Titre: Title:	Méthode de recherche d'information basée sur LDA : étude de cas sur trois revues québécoises en sciences humaines et sociales
Auteur: Author:	Arthur Tobler
Date:	2019
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Tobler, A. (2019). Méthode de recherche d'information basée sur LDA : étude de cas sur trois revues québécoises en sciences humaines et sociales [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/4075/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/4075/
Directeurs de recherche: Advisors:	Catherine Beaudry, & Michel Gagnon
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Méthode de recherche d'information basée sur LDA : étude de cas sur trois revues québécoises en sciences humaines et sociales

ARTHUR TOBLER

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Octobre 2019

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Méthode de recherche d'information basée sur LDA : étude de cas sur trois revues québécoises en sciences humaines et sociales

présenté par **Arthur TOBLER**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Bruno AGARD, président
Catherine BEAUDRY, membre et directrice de recherche
Michel GAGNON, membre et codirecteur de recherche
Michel DESMARAIS, membre

DÉDICACE

« L'éternité de ce qui ne dure pas. »

Fabrice Midal

REMERCIEMENTS

Je tiens à remercier ma directrice de recherche M^me^ Catherine Beaudry de m'avoir accepté comme étudiant à la maîtrise et de m'avoir orienté et donné les moyens de réaliser ce projet durant ces deux années, tout en m'accordant la possibilité de m'essayer à l'enseignement en parallèle de la recherche. Je remercie mon co-directeur de recherche M. Michel Gagnon de m'avoir fait bénéficier de son expertise en analyse de langage, de sa bienveillance et de son humour au cours de nos réunions.

Je souhaite remercier M. Carl St-Pierre pour son aide précieuse dans la validation de ma méthodologie et de mes analyses statistiques et M^me^ Laurence Solar-Pelletier pour sa relecture du sondage, son aide pour faire accepter le questionnaire auprès du comité d'éthique et ses actions pour faciliter le travail des étudiants de la Chaire.

Je remercie l'équipe d'Érudit de m'avoir donné accès à un corpus de données préparées méticuleusement avec lequel ce fut un plaisir de travailler.

Je remercie les chercheurs et professeurs qui ont répondu au questionnaire et ont ainsi grandement valorisé le travail de cette maîtrise.

Je remercie M. Bruno Agard et M. Michel Desmarais d'avoir consacré de leur temps et de leur expertise en composant le jury d'évaluation de ce mémoire.

Je remercie enfin Kim pour sa douceur et toutes ses petites attentions, ma famille pour m'avoir suivi à distance, mes amis de Poly et les étudiants de la Chaire pour tous ces moments vécus à Montréal et ailleurs.

RÉSUMÉ

L'accès rapide à une information pertinente est un enjeu crucial dans un monde contemporain inondé de données. En particulier, extraire efficacement de l'information à partir de données non structurées comme le texte est une tâche difficile. En réponse à ce besoin, l'approche d'apprentissage automatique a montré des résultats prometteurs. Toutefois, la littérature se préoccupe peu de l'opérationnalisation de ces algorithmes à des données différentes de corpus en anglais dans le domaine des sciences naturelles. De fait, les pistes d'application de ces méthodes auprès de la communauté de chercheurs en sciences sociales sont nombreuses mais restent inexploitées. En particulier, une famille de modèles probabilistes, regroupés sous le nom de modèles de thèmes, s'est avérée prometteuse sur certaines tâches telles que la classification et la recherche de documents. Toutefois, du chemin reste à parcourir pour exploiter le potentiel de ces modèles auprès de chercheurs en sciences sociales.

Le but de ce travail est de visualiser, d'évaluer et d'appliquer un modèle de thèmes, le Latent Dirichlet Allocation (LDA), sur un corpus d'articles en sciences humaines et sociales. En particulier, nous proposons une méthodologie d'intégration de ce modèle à une tâche de recherche d'information permettant d'évaluer la pertinence du LDA sur ce type de collection.

L'étude caractérise d'abord quelques paramètres clés dans l'utilisation de ce modèle. Elle montre en particulier que la lemmatisation du vocabulaire n'apporte aucun avantage significatif aux résultats obtenus. Ensuite, elle montre que les méthodes d'évaluation du LDA employées dans la littérature ne sont pas suffisantes pour permettre une application fiable de ces modèles sur les revues étudiées. Pour répondre à ce manque, une méthode de validation externe basée sur une tâche de recherche de documents a donc été développée puis évaluée par des universitaires en SHS. Trois résultats principaux ressortent de cette évaluation opérationnelle. En premier lieu, utiliser directement la représentation vectorielle latente du modèle LDA améliore la pertinence des résultats par rapport à un algorithme utilisant une fréquence de termes liés à la requête. Ensuite, les résultats de fouille sont indépendants du nombre de thèmes du modèle LDA utilisé pour effectuer la fouille dans le cas où on se base sur l'espace latent du LDA. Enfin, l'étude de la spécificité de la requête sur les résultats de fouille n'a pas dégagé d'effet clair sur le corpus étudié.

ABSTRACT

Enabling a quick access to relevant information is a crucial issue in a world saturated with data. Specifically, extracting information from unstructured data like text is a difficult task. The machine learning approach to solve this task has recently shown promising results. However, the literature does not tackle the operationalization of these algorithms to data that differs from corpus published in English in the field of natural sciences. Thus, the application of these methods to research issues in social sciences stays rather unexplored. A specific family of probabilistic models, known as topic models, has shown promising results on some tasks such as document classification. However, further studies are needed to enable a full exploitation of these models for social scientists.

The goal of this work is to visualize and evaluate a specific topic model known as Latent Dirichlet Allocation, on a corpus of articles in social sciences. We develop a methodology to integrate this model into an information retrieval task.

Our study first characterizes some key parameters needed to use this mode. It shows that the lemmatization of vocabulary does not bring any significant benefit to the results. It also indicates that the methods used to evaluate topic models in the literature are not enough to ensure the reliability of these models when applied to a corpus in social sciences. To fill this gap, we developed an external validation method through an information retrieval task evaluated by SHS experts. Three main results were obtained. First, directly using the latent representation of the LDA leads to better relevant results compared to an algorithm using a frequency count of terms related to the query. Second, the relevance of results appears to be independent of the number of topics used in the LDA model. Finally, the specificity of the query does not affect clearly the search results.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

2D : Deux dimensions

ACL : Association for Computational Linguistics

ACP : Analyse en Composantes Principales

AE : Actualité Économique

ANZSRC : Australian and New Zealand Standard Research Classification

EI : Études Internationales

EM : (algorithme) Espérance-Maximisation

K : Nombre de thèmes d'un modèle LDA

IMDB : Internet Movie Database

INEX : INitiative for the Evaluation of XML Retrieval

LDA : Latent Dirichlet Allocation

LEFF : Lexique des Formes Fléchies du Français

LSI : Indexation Sémantique Latente

MCMC : Monte-Carlo par chaînes de Markov

NLTK : Natural Language Toolkit

PDF : Portable Document Formet

PMI : Information Mutuelle Ponctuelle

RI : Relations Industrielles

ROC : Reconnaissance Optique de Caractères

SHS : Sciences Humaines et Sociales

TF-IDF : Term Frequency-Inverse Document Frequency

TREC : Text REtrieval Conference (TREC)

t-SNE : t-distributed Stochastic Neighbor Embedding

UQAM : Université du Québec à Montréal

XML : Extensible Markup Language

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Le texte est le principal média de la communication humaine. Il se retrouve dans de nombreux espaces, à commencer par le Web, la littérature savante et journalistique ou encore au sein des bibliothèques physiques et numériques. L'archivage croissant des contenus écrits et oraux produits par l'être humain engendre une quantité massive de données textuelles accessibles sous forme numérique. Par exemple, le projet Google Livres et le projet Gutenberg ont pour objectif de numériser l'ensemble des livres physiques pour faciliter l'accès à la connaissance (en 2019, Google Livres propose 25 millions de volumes). La valorisation de ces données non structurées est un enjeu majeur de notre siècle. Dans le cadre de l'industrie 4.0, elle répond notamment à des fins prédictives et préventives, par exemple en vue d'optimiser les chaînes de production et d'améliorer la gestion des risques. La littérature scientifique ne fait pas exception : le nombre de publications scientifiques croît exponentiellement et avec elles le nombre d'articles accessibles par le biais des bases de données bibliographiques comme Web of Science, Elsevier ou Google Scholar. Ces données sont aujourd'hui proposées via des interfaces limitées, offrant essentiellement des moteurs de recherche booléens et non paramétrables pour des besoins spécifiques. Un réel besoin existe au sein de la communauté de chercheurs, en particulier en sciences humaines et sociales (SHS), pour disposer d'outils automatiques permettant d'organiser et de rechercher de l'information au sein d'un corpus de documents. Cependant, l'automatisation de ces tâches est ardue : se posent les questions de la préparation du texte pour l'adapter à un traitement automatique, de la modélisation des documents pour en conserver l'information utile et du choix des méthodes d'évaluation des modèles sur certaines tâches spécifiques. Une approche moderne pour extraire efficacement de l'information de ces corpus est d'utiliser des modèles statistiques d'apprentissage automatique. En particulier, la modélisation thématique de documents s'avère une voir prometteuse vers un accès personnalisé à la connaissance. Toutefois, la plupart des travaux de la littérature valident leur modèle sur des articles de science naturelle via des mesures automatiques, décorrélées des besoins réels d'utilisateurs.

L'objectif de ce mémoire est de participer à l'accessibilité et l'adoption d'un modèle de thème populaire, le Latent Dirichlet Allocation (LDA), au sein de la communauté de chercheurs en SHS. Pour cela, nous souhaitons utiliser ce modèle sur une tâche de recherche de documents.

Deux questions de recherche guideront ce travail.

Question 1 : Comment décrire et évaluer thématiquement un corpus d'articles en sciences sociales à l'aide du modèle LDA ?

Question 2 : Comment intégrer le LDA sur une tâche de recherche d'articles en sciences sociales ?

Ce mémoire se situe à l'interface entre la modélisation informatique et statistique du langage d'une part et l'analyse de texte appliquée aux SHS d'autre part. Il consiste à développer et valider des méthodes d'analyse automatique de texte dans un domaine d'application encore peu exploré par la littérature. Il se propose de valoriser des outils techniques auprès de chercheurs non experts de ces outils et exprimant un besoin clair : disposer d'une méthode de gestion, de description et de recherche de documents spécialisés.

Trois objectifs de recherche découlent de ces deux questions.

Premièrement, il s'agit de fournir une représentation thématique d'un corpus d'articles en SHS à l'aide du modèle LDA. Une méthodologie claire et un code détaillé ont été développés pour que les expériences puissent être reproductibles sur d'autres bases de données et par des non experts en sciences de données. Différentes méthodes de visualisation ont été présentées pour faciliter l'appréhension des résultats des modèles de thèmes auprès de non experts.

Ensuite, il s'agit d'évaluer cette représentation afin d'en assurer la validité et la fiabilité. Les mesures automatiques de perplexité et de cohérence seront utilisées pour cette étape d'évaluation.

Enfin, on souhaite utiliser le modèle déterminé à l'aide des deux premières étapes, pour effectuer une tâche de recherche d'information. On veut construire et comparer différents algorithmes de recherche d'information basés sur le LDA selon l'avis d'utilisateurs réels.

Les contributions du mémoire sont les suivantes :

1) Nous proposons une méthode d'évaluation de la validité et de la fiabilité des modèles LDA à l'aide de mesures d'évaluation automatique de la littérature.

2) Nous proposons une méthode de recherche d'information spécialisée, basée sur le LDA et évaluée par des experts sur un corpus d'articles en sciences humaines et sociales.

Ce mémoire comporte six chapitres.

Le chapitre qui suit approche la littérature sur laquelle se base ce travail. Il présente les modèles théoriques actuels qui servent à modéliser des corpus de documents et détaille leurs applications

en sciences sociales. Il souligne en particulier les manques de la littérature que ce travail cherchera à combler.

Les questions de recherche, la méthodologie choisie pour y répondre ainsi que les données utilisées sont exposées dans le troisième chapitre.

Le quatrième chapitre vise à répondre à la première question de recherche. Il présente les visualisations des thèmes obtenus à l'aide de différents modèles LDA et les analyses des résultats de validité et fiabilité de ces modèles.

Le cinquième chapitres s'intéresse à la seconde question de recherche. Il présente une méthodologie pour intégrer les thèmes de modèles LDA sur une tâche de recherche d'information, ainsi que les résultats obtenus suite à une évaluation humaine.

Les conclusions de ce travail, les recommandations et les possibilités d'extension sont formulées dans le dernier chapitre.

CHAPITRE 2 REVUE DE LITTÉRATURE

Ce mémoire a été écrit avec l'intention qu'il puisse servir de support de travail à des chercheurs en SHS souhaitant comprendre et appliquer des modèles de thèmes à leur corpus de textes. À ces fins, la revue de littérature est découpée en trois parties. À la fin de la première partie, le lecteur sera familier avec les enjeux et les méthodes de la modélisation statistique du langage et de l'apprentissage automatique. Une fois ces bases acquises, le cas particulier de l'analyse automatique du texte et de ses applications en SHS lui sera présenté. Enfin, il découvrira le modèle spécifique utilisé dans ce travail : le Latent Dirichlet Allocation (LDA).

2.1 Modélisation statistique du langage et apprentissage automatique

Afin de comprendre la particularité des modèles utilisés dans ce travail, il apparaît nécessaire de rappeler les fondements de l'étude statistique du langage ainsi que les méthodes d'apprentissage automatique qui ont favorisé l'essor de l'analyse massive de textes.

2.1.1 Modélisation statistique du langage

Modéliser le langage

Lorsque l'on présente une collection de données à un ordinateur, il faut fournir à celui-ci des clés de compréhension pour qu'il soit capable de résoudre une tâche précise à l'aide de ces données. Par exemple, supposons que l'on dispose d'un corpus de documents que l'on souhaite classifier selon certains critères prédéfinis. Ces clés de compréhension sont alors la *représentation* des données et la *modélisation* du problème à résoudre. La question de la représentation est traitée dans la section 2.2. La partie de modélisation consiste à déterminer la forme générale du modèle mathématique d'analyse et à définir le but à atteindre par l'algorithme. En pratique, les données du monde réel comportent leur lot d'imperfections qui empêchent une analyse directe : données manquantes, erreurs de mesure, hétérogénéité, variables d'influence inconnues, etc. Pour modéliser ces imperfections, une approche courante est d'utiliser des modèles probabilistes. Le principe est de trouver une distribution de probabilité qui *explique* les données observées. On souhaite donc expliciter les lois de probabilités à partir desquelles les donnes pourraient être reconstruites. Dans le cas des données qui nous intéressent, à savoir du texte, c'est le domaine de la modélisation statistique du langage. Un modèle statistique de langue idéal devrait être capable de recréer un

corpus de documents simplement à partir du vocabulaire utilisé. La distribution de probabilité déterminée par le modèle sera caractérisée par une *forme* que l'on fixe *a priori* et par des *paramètres* que l'on cherche à optimiser. C'est l'approche dite *paramétrique* que nous avons utilisée dans ce travail. Une autre approche, dite *non paramétrique* consiste à laisser le modèle déterminer la famille de distribution de probabilité la plus adaptée au problème. Pour faire le lien entre les données observées et les paramètres inconnus du modèle, on fait appel aux modèles graphiques probabilistes. Ces modèles établissent le lien entre trois composantes :

1) *Les variables aléatoires visibles*, que l'on observe à l'aide de nos données. Par exemple, ce sont les mots qui apparaissent dans un corpus de documents.

2) *Les variables aléatoires latentes*, que l'on n'observe pas directement à l'aide de nos données. Par exemple, ce sont les thèmes abordés dans un corpus de documents.

3) *Une distribution de probabilité* qui établit le lien entre ces deux ensembles, et que l'on cherche à déterminer. Par exemple, une distribution de probabilité jointe de forme normale entre les mots et les thèmes.

Dans les modèles que nous utiliserons, la probabilité jointe sera représentée à l'aide d'un modèle graphique direct (Pearl, 1995). Ce type de modèles permet de représenter les relations de dépendances entre les variables. Ils facilitent la compréhension théorique de l'implémentation et de l'inférence du modèle. Graphiquement, ces modèles sont tracés sous la forme de diagrammes bloc-flèches. Chaque variable aléatoire est représentée par un cercle, transparent pour les variables latentes et grisé pour les variables visibles. Un bloc représente un phénomène de réplication pour plusieurs variables. Les flèches représentent des relations de dépendance entre deux variables. Une fois le modèle probabiliste construit et le graphe dessiné, il reste à inférer la probabilité conditionnelle reliant les variables visibles et latentes. En d'autres mots, on veut recréer l'histoire générative des données à partir de certaines de leurs caractéristiques. La problématique peut se formuler ainsi : quelle est la vraisemblance d'observer les données à partir des variables latentes ? En pratique, on va optimiser la valeur des distributions de probabilités des variables latentes pour retrouver les données observées. C'est la deuxième étape de ces modèles, la plus difficile à effectuer, que l'on qualifie d'inférence. L'équation fondamentale pour comprendre le processus d'inférence est la formule de Bayes :

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Équation 2.1 Théorème de Bayes

Où l'on a :

- θ : paramètre du modèle dont on cherche la valeur (lié aux variables latentes).
- p(θ) : distribution de probabilité *a priori* du paramètre θ.
- D : données observées (les mots des documents par exemple).
- p(θ|D) : distribution de probabilité *a posteriori*, donc du paramètre θ conditionnellement aux données observées. C'est elle qui nous intéresse. Sa valeur exacte nécessite de connaître la valeur de p(D).
- p(D) : distribution de probabilité *marginale* des données observées. Elle correspond à toutes les configurations possibles des données (par exemple, tous les documents que l'on peut former avec un certain nombre de mots). C'est la valeur la plus difficile à estimer dans la formule de Bayes et elle nécessitera des méthodes de calcul d'inférence approximé.
- p(D|θ) : distribution de probabilité de *vraisemblance*. Étant donné le paramètre θ, on souhaite maximiser la vraisemblance d'obtenir les données observées.

Exemple : le modèle unigramme

Nous illustrons ces concepts à l'aide d'un modèle probabiliste de langue simple : le modèle unigramme. Rappelons le problème : on dispose de *D* textes composés à l'aide d'un vocabulaire de *N* mots et l'on cherche à reconstruire l'histoire générative de ces textes. Autrement dit, à partir d'une suite de mots, on souhaite trouver le mot le plus probable qui complète cette suite. Le modèle unigramme suppose que la distribution de probabilité du mot $w_{n,d}$ dépendra uniquement du mot précédent. Chaque mot $w_{n,d}$ est tiré d'une distribution de probabilité multinomiale de paramètre β. L'ensemble des mots du corpus suit alors la probabilité jointe suivante, équivalente à la représentation graphique de la Figure 2.1 :

$$p(w_{11}, \dots, w_{ND}) = p(\beta)\, x \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{n,d}|\beta)$$

Équation 2.2 Probabilité jointe du modèle unigramme

Où $w_{n,d}$ représente le *nième* mot du document *d* et N_d le nombre de mots dans le document *d*.

$p(w_{n,d} \mid \beta)$ est la distribution multinomiale qui paramétrise le mot $w_{n,d}$.

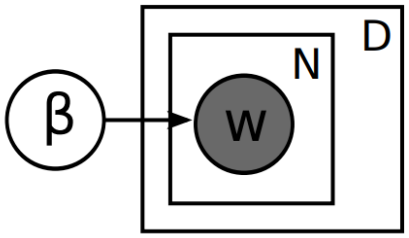

Figure 2.1 Diagramme de plaques du modèle de langue unigramme.

Les documents sont représentés à l'aide d'une collection de mots w_n. β est le paramètre de la distribution de probabilité multinomiale associée.

La probabilité jointe s'interprète comme la probabilité d'observer la séquence de mots w_{11} à w_{ND}. La flèche du graphe de la Figure 2.1 illustre la décomposition de la probabilité jointe en somme de probabilités conditionnelles que le modèle cherchera à estimer. Ces modèles à variables latentes offrent plusieurs avantages pratiques. Ils peuvent décrire et modéliser des phénomènes variés, sont facilement interprétables et ils sont modulaires. De plus, de nombreux outils statistiques existent pour les implémenter, les entraîner et les évaluer. Le succès de l'approche probabiliste (aussi dite bayésienne) dépend de la nature des données, des hypothèses de dépendance entre les variables et des choix de distribution de probabilité qui sont effectués.

Faire face au problème de l'inférence

Dans le cas du modèle unigramme, la distribution de probabilité postérieure $p(\theta|D)$ peut se calculer directement. Toutefois, ce n'est pas le cas de la plupart des modèles statistiques de langue, à cause du dénominateur p(D). Celui-ci s'avère incalculable pour la majorité des modèles, à cause des multiples interactions entre variables observées et variables latentes. Différentes méthodes ont été développées pour approximer cette distribution de probabilité. Nous présentons ici le principe de deux d'entre elles, qui permettront au lecteur d'avoir une meilleure compréhension du mécanisme utilisé par les modèles probabilistes : le maximum *a posteriori* et les chaînes de Markov Monte-Carlo. Nous utiliserons le modèle présenté en 2.3.1 pour illustrer deux méthodes plus avancées qui sont l'échantillonnage de Gibbs et l'inférence variationnelle.

L'idée de l'estimation du *maximum a posteriori* (MAP) est de maximiser la valeur de la distribution de probabilité postérieure indépendante de θ, en oubliant la probabilité jointe incalculable. Le paramètre θ optimal sera déterminé par l'équation :

$$\theta_{MAP} = argmax_{\theta}\ (p(\theta|D)) = argmax_{\theta}\ (p(D|\theta)p(\theta))$$

Pour éviter des dépassements de capacité de calcul, on se ramène la plupart du temps au logarithme de cette fonction. Puis on utilise des méthodes d'optimisation (souvent à base de descente de gradient) pour trouver le θ qui explique le mieux les données. Le choix de la distribution de probabilité *a priori* p(θ) sera d'autant plus important que le nombre de données sera petit. Si p(θ) est fort, l'estimation MAP trouvera une distribution de probabilité dense centrée autour de la meilleure valeur de θ. Si p(θ) est faible, l'estimation MAP trouvera une distribution de probabilité similaire à l'estimation qui maximiserait uniquement la vraisemblance p(D|θ). De façon générale, l'estimation MAP fournit une estimation ponctuelle de θ qui peut servir d'approximation à la distribution de probabilité postérieure. Dans le cas où le maximum de vraisemblance s'avère difficile à calculer directement, on peut utiliser un algorithme *d'espérance-maximisation* (EM) qui scinde le problème en deux étapes. Une étape d'espérance (E) où l'on estime la valeur de la log-vraisemblance en supposant le reste des paramètres connus. Puis une étape de maximisation (M) où l'on estime les paramètres en maximisant la log-vraisemblance estimée dans l'étape (E). On ne détaille pas le fonctionnement de cette méthode car elle ne sera pas utilisée dans notre travail, mais on renvoie le lecteur intéressé à Zeng *et al.* (2016) pour une explication de l'algorithme EM appliqué au modèle LDA.

Les méthodes *Monte-Carlo par chaîne de Markov* (MCMC) sont une autre façon d'évaluer la probabilité postérieure (Bishop, 2006). L'idée des MCMC est d'échantillonner successivement des candidats pour la distribution de probabilité p(θ|D) à partir d'une densité de probabilité jointe non normalisée p(D,θ). La suite d'échantillonnage est construite de façon à converger vers la distribution de probabilité postérieure p(θ|D). Une fois que l'on a collecté suffisamment d'échantillons, on peut les agréger pour former une estimation de la distribution de probabilité postérieure. Cette méthode a longtemps été la plus populaire dans les problèmes d'approximation de distribution de probabilités postérieures (Blei *et al.*, 2017). L'inconvénient des MCMC est le temps de calcul et le nombre d'échantillons nécessaires pour converger suffisamment vers la

distribution de probabilité postérieure, en particulier lorsque l'on travaille avec de grands corpus de données.

Le processus d'inférence correspond en fait à une forme d'apprentissage automatique. On présente maintenant le cadre général de l'apprentissage automatique, qui permet à un ordinateur d'effectuer des tâches spécifiques.

2.1.2 Apprentissage automatique

Un algorithme d'apprentissage automatique est conçu pour « apprendre » à partir de *données*. « Apprendre » peut se comprendre comme un processus où l'algorithme va ajuster certains paramètres d'un modèle mathématique dans le but d'améliorer sa *performance* sur une *tâche* selon une certaine mesure (Mitchell *et al.*, 1990). Le développement de ce type d'algorithmes est motivé par la résolution de tâches trop difficile pour être menée par des humains ou à l'aide de règles explicites codées à la main. Ils se trouvent à l'interface entre trois domaines : le design du modèle repose sur des outils d'algèbre linéaire, l'évaluation et l'interprétation se basent sur des méthodes statistiques d'analyse de données et l'implémentation utilise des outils de programmation.

Données

Pour apprendre, un algorithme a besoin de données, et plus précisément de caractéristiques quantitatives extraites de ces données. Dans le cas d'un article par exemple, on pourra compter la fréquence de chaque mot le composant et fournir à l'algorithme un vecteur de fréquences représentant le document. La section 2.2 détaille les méthodes de représentations du texte. De façon générale, un vecteur de nombres est fourni à l'algorithme pour chaque point de données que l'on analyse. En pratique, on forme une matrice où chaque ligne correspond à un point de données et chaque colonne une de ses caractéristiques sous forme de nombre. Cette représentation nécessite d'avoir le même nombre de caractéristiques pour chaque point de données. Comme pour toute analyse statistique, la qualité des données en entrée sera cruciale à la performance de l'algorithme sur la tâche d'intérêt. Plus le modèle disposera de paramètres à estimer, plus on devra lui fournir un grand nombre de données. Parfois, les données viennent avec des informations supplémentaires comme une étiquette. Par exemple, on peut disposer d'articles scientifiques avec le nom du journal de publication et souhaiter classifier chaque article selon le journal. C'est un cas d'apprentissage *supervisé*, où l'on dispose d'un lien explicite entre les caractéristiques des données et leur étiquette.

Si pour la même tâche, nous n'avions pas accès au nom du journal, ce serait un cas d'apprentissage *non supervisé*. La différence entre les deux types d'apprentissage peut se comprendre à l'aide des probabilités. On considère chaque vecteur de donnée comme une valeur prise par un vecteur aléatoire $\vec{x}$. Dans le cas non supervisé, on cherche à déterminer la distribution de probabilité $p(\vec{x})$. Dans le cas supervisé, on dispose des étiquettes de chaque donnée considérée comme des exemples d'un vecteur aléatoire $\vec{y}$ et l'on cherche à déterminer la distribution de probabilité $p(\vec{y}|\vec{x})$.

Tâches

Différentes tâches peuvent être abordées par des algorithmes d'apprentissage. En classification, on souhaite étiqueter les données selon certains critères. Cet étiquetage peut être déterministe (on attribue une étiquette pour chaque point de donnée) ou probabiliste (on attribue une distribution de probabilité sur les étiquettes pour chaque point de donnée). En régression, on souhaite prédire une valeur de sortie numérique à partir d'un point de donnée d'entrée. En traduction automatique, on souhaite établir une correspondance entre deux langues. En détection d'anomalies, on cherche à identifier des points atypiques parmi l'ensemble des données. En estimation de densité, on cherchera à déterminer une distribution de probabilité qui permet de générer vraisemblablement les données observées, etc.

Performance

Pour comparer différents algorithmes, il faut construire des mesures adaptées à chaque tâche. Par exemple, pour la classification discrète, on s'intéressera souvent au taux d'erreur commis par l'algorithme tandis que pour l'estimation de distribution de probabilité, on utilisera la log-vraisemblance. La mesure de performance utilisée doit être valide (mesurer ce qu'on veut effectivement mesurer) et fiable (donner des résultats reproductibles). Dans certains cas, il n'est pas évident de déterminer ce qui doit être mesuré et ce qui doit être pénalisé. Par exemple, dans le cas d'une régression, on peut sanctionner les erreurs graves et rares ou bien les erreurs moyennes plus fréquentes. Il arrive aussi qu'on sache ce que l'on souhaite optimiser mais que l'on ne puisse y accéder directement : c'est le cas de la plupart des modèles probabilistes, où la distribution de probabilité que l'on cherche est incalculable.

Généralement, on s'intéresse à la performance de l'algorithme sur des données nouvelles. Par exemple, est-il capable de classifier correctement de nouveaux articles scientifiques à partir du corpus sur lequel il a été entraîné ? Si l'algorithme a une bonne performance sur ces données

nouvelles (qui forment l'ensemble de *test*), on dira qu'il *généralise* bien. La capacité de généralisation d'un algorithme est cruciale et a motivé les développements récents en apprentissage, dont les réseaux de neurones profonds (Goodfellow *et al.,* 2016). Un des critères de réussite de l'évaluation sur des données nouvelles est que celles-ci soient indépendantes et identiquement distribuées par rapport aux données d'entraînement. En pratique, on dispose souvent d'un ensemble de données fixe. La méthode pour entraîner l'algorithme est de décomposer cet ensemble en deux sous-ensembles :

i. L'ensemble d'*entraînement*. L'algorithme est optimisé sur cet ensemble à l'aide d'une certaine mesure de performance.

ii. L'ensemble de *test.* La performance de l'algorithme est évaluée sur ces données.

Une heuristique de séparation fréquemment utilisée est de 80/20, mais cela peut changer en fonction des données et des tâches. Un bon algorithme sera capable d'avoir une erreur faible sur l'ensemble d'entraînement et un écart de performance faible entre les deux ensembles. Deux familles d'erreurs sont possibles :

i. *Sous-apprentissage*. L'algorithme est incapable d'avoir une performance suffisante sur l'ensemble d'entraînement.

ii. *Surapprentissage.* L'algorithme est incapable d'avoir un écart de performance faible entre les deux ensembles.

Pour augmenter la performance sur l'ensemble d'entraînement, les possibilités théoriques sont d'augmenter le nombre de données et de diminuer la *capacité* du modèle. La capacité se définit par le nombre de paramètres, la forme du modèle et l'espace de solutions accessibles au modèle. Cependant, diminuer la capacité du modèle va naturellement diminuer sa capacité de généralisation. Il faut donc trouver un compromis, c'est ce qu'on appelle le compromis *biais-variance*. Le biais mesure l'écart entre les paramètres trouvés par l'algorithme et les paramètres optimaux qu'il aurait pu trouver pour le modèle qu'il optimise : un biais élevé correspond au scénario de sous-apprentissage sur les données d'entraînement, à savoir une faible capacité du modèle. La variance mesure la sensibilité du modèle aux données d'entraînement : un modèle avec une forte variance aura un résultat qui fluctue fortement selon les données utilisées pour l'entraîner. Plus la tâche sera complexe, plus le modèle devra avoir une grande capacité. Mais une grande

capacité implique aussi de disposer d'une grande puissance de calcul pour l'entraînement du modèle et plus sa variance sera élevée car il va capturer du bruit inutile.

Illustrons cela avec un problème de régression. Supposons que l'on veuille ajuster un modèle de régression sur un ensemble de points générés à l'aide d'une fonction de degré 2. Une fonction linéaire sera incapable de capturer la courbure des données : c'est un modèle avec un biais élevé et une variance faible. À l'inverse, un polynôme de degré 9 va capturer le bruit qui accompagne les points de données : c'est un modèle avec un biais faible et une variance élevée. Finalement, un polynôme de degré 2 aura la bonne complexité pour s'ajuster aux données : il a suffisamment de paramètres pour assurer le compromis biais-variance. La Figure 2.2 illustre ce comportement.

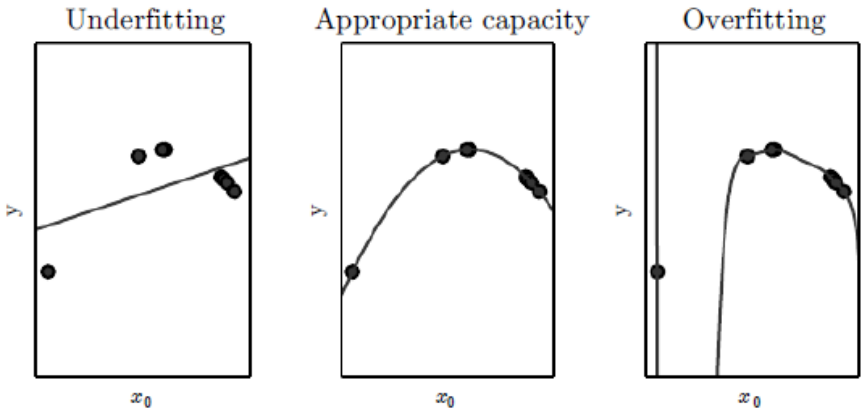

Figure 2.2 Illustration de la capacité d'un modèle sur un problème de régression d'après Goodfellow *et al.* (2016)

Il faut noter que la plupart des tâches sont trop complexes pour que l'algorithme soit capable de trouver un optimum global de performance. La solution proposée correspondra à un minimum local qui dépendra de la méthode d'optimisation utilisée (voir Sra *et al.*,2012 pour une revue des méthodes d'optimisation utilisées en apprentissage automatique). Par ailleurs, le meilleur des modèles n'arrivera pas à une erreur nulle sur l'ensemble de test à cause de la nature bruitée des données : cette erreur minimale est appelée *erreur de Bayes*. Pour aider l'algorithme à converger vers un optimum satisfaisant, on dispose de certains leviers d'action que l'on appelle *hyperparamètres*. Dans le cas de notre problème de régression, c'était le degré du polynôme. Contrairement aux paramètres du modèle, ils ne sont pas appris directement par l'algorithme pour éviter un comportement de surapprentissage. Un troisième sous-ensemble de données peut être construit pour adapter ces hyperparamètres, qu'on appelle ensemble de *validation*.

On détaille maintenant quelques mesures d'évaluation généralement utilisées dans les modèles d'apprentissage probabiliste. L'objectif de l'évaluation est de tester la performance du système sur une tâche et de pouvoir mener une analyse critique de la robustesse, la fiabilité et les limitations du modèle. Pour évaluer des modèles d'apprentissage probabilistes, la mesure traditionnelle consiste à considérer la *log-vraisemblance*. La vraisemblance est définie comme la probabilité d'observer un ensemble de données. Autrement dit, on évalue la capacité du modèle à générer nouveau document similaire à ceux utilisés lors de l'entraînement. Il est d'usage d'en calculer le logarithme afin d'éviter un dépassement de capacité de calcul. On peut la mesurer sur l'ensemble d'entraînement ou de test. Dans le premier cas, on évalue alors la capacité du modèle à s'ajuster à un ensemble de données. La log-vraisemblance du modèle est la somme des log-vraisemblance pour chaque point de donnée. L'inconvénient majeur d'évaluer la log-vraisemblance sur l'ensemble d'entraînement est de ne pas pouvoir évaluer le surapprentissage. On utilise cette méthode davantage comme critère d'arrêt de l'entraînement dans le cas d'une estimation de maximum *a* priori ou d'un algorithme de maximum de vraisemblance. Dans le second cas, une fois le modèle entraîné sur un ensemble d'entraînement, on cherche à mesurer sa capacité de généralisation, autrement dit s'il est capable de modéliser un autre ensemble de données similaires aux données d'entraînement. La log-vraisemblance est alors définie à partir de la probabilité d'observer des points de données de test étant donnés les points de données d'entraînement. La log-vraisemblance du modèle est égale à la somme des log vraisemblance de chaque point de test *étant donné* l'ensemble d'entraînement.

À partir de la log-vraisemblance, on peut construire une nouvelle mesure adaptée aux modèles de thèmes utilisés dans ce mémoire : *la perplexité* (Blei *et al.*, 2003). On détaille les modèles de thèmes à la section 2.3. Pour l'instant il suffit de savoir que ce sont des modèles qui visent à apprendre une représentation globale d'un corpus de documents. L'idée est ensuite d'évaluer la capacité du modèle à « générer » de nouveaux documents similaires à ceux du corpus. Une fois entraîné sur un corpus d'entraînement, on évalue donc sa capacité à recréer les documents de l'ensemble de test : c'est le but de la mesure de perplexité. Pour l'ensemble de test D_{test} composé de M documents, elle se définit par l'équation suivante :

$$perplexité(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d}\right\} = exp\left\{-\frac{log - vraisemblance}{nombre\ de\ jetons}\right\}$$

Équation 2.3 Définition de la perplexité

Au numérateur on retrouve la log-vraisemblance évoquée plus haut, qui mesure la capacité du modèle à prédire les mots d'un nouveau document. Au dénominateur, c'est un facteur de normalisation qui prend en compte le nombre total de mots présents dans un documents pour pouvoir comparer les documents de taille différente entre eux. La perplexité correspond en fait à l'inverse de la moyenne géométrique de la log-vraisemblance, divisée par le nombre total de jetons dans le corpus. C'est une fonction décroissante de la log-vraisemblance de l'ensemble de test : un bon modèle génératif aura une faible perplexité. La log-vraisemblance des modèles de thème est incalculable donc on procède à des méthodes d'approximations (Wallach *et al.*,2009). Par ailleurs, tous les modèles visent *in fine* à être utilisés dans des tâches appliquées. Il est donc nécessaire de valider la qualité et l'utilité de la représentation des modèles sur des applications potentielles réelles. Ce type d'évaluation indirecte à l'aide d'une tâche utilise des mesures de performance déjà établies. Dans le cas de modèles supervisés, on peut évaluer la performance du modèle par rapport aux humains ou à d'autres modèles. Par exemple, lors d'une tâche de classification binaire de documents, il est d'usage de comparer trois mesures : l'exactitude (*accuracy*), la précision (*precision)* et le rappel (*recall)* (Kent *et al.*, 1955). L'exactitude correspond au ratio du nombre de documents bien classés sur le nombre total de documents classés. La précision mesure la « pureté » des classes de documents, autrement dit le ratio de vrais positifs d'une classe sur l'ensemble des documents attribués à cette classe. Le rappel mesure la proportion de documents pertinents que l'algorithme sélectionne pour une tâche parmi tous les documents qui devraient être attribués à cette classe.

Tableau 2.1 Matrice de confusion d'une tâche de classification binaire

	Pertinents	Non pertinents
Documents retournés	Vrais Positifs (VP)	Faux Positifs (FP)
Documents non retournés	Faux Négatifs (FN)	Vrais Négatifs (VN)

On peut exprimer mathématiquement ces trois mesures à partir des notations données dans le Tableau 2.1.

$$exactitude = \frac{VP + VN}{VP + FP + VN + FN} ; précision = \frac{VP}{VP + FP} ; rappel = \frac{VP}{VP + FN}$$

Équation 2.4 Mesures de performance d'une classification binaire

Dans le cas de modèles non supervisés, on ne dispose pas *a priori* de vérité absolue contre laquelle comparer les résultats de l'algorithme. Une des possibilités est alors de construire une tâche *ad hoc* qui permet de comparer les résultats de différents modèles. Dans le cadre de notre travail, nous avons par exemple construit un moteur de recherche utilisant la représentation de documents calculée par le modèle. Puis nous avons évalué la pertinence des résultats proposés par l'algorithme à l'aide d'un questionnaire proposé à des humains. La section 5.2 détaille cette évaluation externe. Pour une présentation plus détaillée des concepts mentionnés dans cette section, on renvoie le lecteur à Goodfellow *et al.* (2016).

2.2 Analyse automatique de textes en sciences humaines et sociales

Une fois rappelé le cadre général de l'analyse automatique de données, on s'intéresse plus en détail à un type de données particulier : le texte et, plus spécifiquement, les textes de sciences sociales. La plupart des tâches d'analyse automatique de textes peuvent se décomposer en quatre étapes (Kowsari *et al.,* 2019), qui sont résumées par le schéma de la Figure 2.3.

1) **Extraction de caractéristiques du corpus**. Il s'agit d'extraire des propriétés structurées à partir du texte afin de pouvoir l'inclure dans un modèle d'analyse. Par exemple, on pourra appliquer certains filtres au texte puis le convertir en une représentation sac de mots.

2) **Réduction de dimensionnalité.** Il s'agit de ne garder qu'une partie des propriétés extraites. Cette étape optionnelle permet de réduire le temps de calcul de l'algorithme et le stockage mémoire nécessaire pour les données.

3) **Entraînement d'un algorithme (de classification, d'extraction d'information etc.).** Le choix des algorithmes dépendra des données, de la tâche à résoudre et des puissances de calcul et de mémoire dont le chercheur dispose.

4) **Évaluation.** Il s'agit de comprendre et de mesurer les forces et faiblesses du système d'analyse développé et de permettre une rétroaction sur le choix et le design de l'algorithme utilisé.

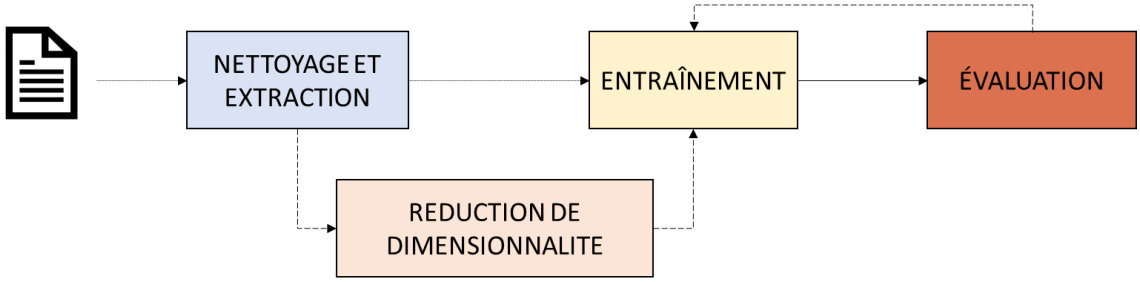

Figure 2.3 Pipeline d'analyse de texte

La section 2.2.1 détaille les deux premières étapes tandis que la section 2.2.2 détaille les deux dernières étapes en s'appuyant sur les deux types de tâches que l'on retrouve en SHS : la classification de documents et la recherche d'information.

2.2.1 Représentation informatique du texte

Quand on souhaite analyser automatiquement du texte, une première question se pose : comment transformer le texte d'une forme lisible par un humain à une forme interprétable par une machine ?

Le texte respecte les règles que lui impose sa langue d'écriture. Ces règles dépendent de la langue utilisée (vocabulaire, grammaire, conjugaison…) mais aussi du style du rédacteur (formel ou non, soutenu ou familier…). Un même mot peut avoir un sens différent (exemple : l'homonyme « compte », dans « compte en banque » ou « il compte son argent ») et deux phrases peuvent avoir un même sens sans pour autant partager de mots (exemple : « Matthieu se lève de bonne heure. » et « Il n'aime pas se réveiller tard. »). Le contexte d'utilisation est ce qui permet de lever l'ambiguïté des situations. C'est cette intuition qui a servi de développement à des algorithmes modernes de représentation comme on verra ci-après. Qui plus est, la langue n'est pas un système statique, les mots et leur usage dépendent du temps et des civilisations. Ces difficultés inhérentes au texte et à l'écriture restent aujourd'hui encore des sujets de recherche actifs. Malgré tout, des méthodes de traitement ont été développées ces cinquante dernières années pour aider les machines à analyser du texte.

Nettoyer le texte

Comme pour toute donnée, une première étape nécessaire aux analyses est de *nettoyer* le texte. En effet, la plupart des documents contiennent des caractères ou mots nuisibles aux analyses. Le

nettoyage consiste à retirer ce bruit qui affecte les performances des algorithmes et augmente inutilement le vocabulaire stocké en mémoire. Diminuer la taille du vocabulaire permet de limiter la problématique de sous-apprentissage (trop peu de documents par rapport au nombre de mots) et de malédiction de dimension (trop de mots pour chaque document). Le processus de nettoyage varie en fonction du modèle utilisé mais on retrouve les étapes suivantes dans la plupart des travaux.

Tokenisation. La tokenisation consiste à séparer un texte en jetons (*tokens*). Un jeton peut être un mot, un bigramme (ex : New_York) voir un n-grammes ou bien un symbole de ponctuation (Gupta et Malhotra, 2015). Le but de cette première étape est d'identifier les entités de chaque phrase (Aggarwal, 2018).

Mots arrêts, ponctuation et chiffres. Les symboles de ponctuation et certains mots fréquemment utilisés dans une langue (par exemple en français : « *le* », *« et », « car », « alors »*…) apportent peu d'information pour la classification ou l'extraction d'information. Il est d'usage de les retirer du vocabulaire. De même, les chiffres s'avèrent peu porteurs de sens pour ces tâches donc on les enlève.

Mise en minuscule. On ramène la plupart du temps chaque mot en lettres minuscules afin d'agréger les différentes versions d'un même mot. Cela permet de limiter la multiplicité des formes d'un même mot mais peut induire certaines erreurs d'identification comme « EU » (États-Unis) qui devient « eu » (participe passé de « avoir »). Ce type de problème est plutôt rare, mais il est possible de le corriger en appliquant en amont un algorithme de reconnaissance d'entités nommées (REN).

Correcteur orthographique. Parfois, il peut être utile de vérifier la typographie de certains mots afin d'éviter les redondances dans la constitution du vocabulaire. Par exemple, dans le cas de fichiers XML obtenus à partir de fichiers PDF à l'aide d'un système de reconnaissance optique de caractères (ROC), des erreurs de conversion apparaissent fréquemment (Lopresti, 2009). Différents algorithmes existent pour corriger les erreurs orthographiques. On renvoie le lecteur à Christanti *et al.* (2018) pour un tour des méthodes actuelles. La plupart utilisent des mesures de distance d'édition (ou distance Levenshtein) sur des n-grammes (Angell *et al.*, 1983) ou bien des tables de hachage (Hantler *et al.*, 2017).

Racinisation et lemmatisation. Un mot peut apparaître sous différentes formes (conjuguées, pluriel…) mais garder un sens identique. La racinisation permet d'identifier les différentes formes d'un mot et de les considérer comme une seule entité. Singh et Gupta (2017) proposent une revue

de littérature récente sur la théorie, les méthodes et applications de la racinisation. L'algorithme de Porter en anglais et l'algorithme de Curry en français en sont des exemples typiques. La racinisation est une étape difficile à automatiser compte tenu de la quantité de variations offertes par une langue comme le français. En pratique, on peut utiliser un processus plus simple qu'on appelle la lemmatisation. Il consiste à identifier et retirer le suffixe d'un mot pour le ramener à sa forme canonique (le lemme). Enfin pour certaines applications, il peut être nécessaire d'ajouter un étiquetage morphosyntaxique sur chaque entité : on détermine si l'entité est un nom propre ou commun par exemple.

Choisir une représentation du texte

Une fois le vocabulaire du corpus nettoyé, il faut procéder au choix d'une représentation du corpus. Le texte est une donnée de nature non structurée. Cela rend difficile la définition de caractéristiques propres au texte. Nous ne disposons pas encore de machines intelligentes capables de « comprendre » un texte comme le ferait un humain à l'aide d'opérations d'association mentale, de logique et de connaissance contextuelle. L'approche par défaut est plutôt d'extraire certaines caractéristiques du texte qui pourront être utilisées comme entrée de modèles d'analyse. L'extraction de caractéristiques utiles est donc une étape cruciale et un domaine de recherche majeur en analyse automatique de langue. Avant de présenter quelques méthodes modernes, rappelons qu'il n'existe pas de méthode meilleure qu'une autre dans l'absolu. Seule la performance sur une tâche et l'utilité dans une application réelle donnée permet de déterminer si un algorithme est meilleur qu'un autre. En bref, « *all models are wrong but some are useful.* » (George E. P. Box, 1976).

Un premier choix méthodologique important est celui de l*'échelle* de représentation. Choisit-on de représenter le texte à l'échelle d'un document, d'un paragraphe, d'une phrase ou d'un mot ? On choisira ici l'échelle du document. C'est la méthode généralement adoptée dans le cas des applications en SHS car la plus intuitive pour interpréter les résultats (Grimmer et Stewart, 2013) et la plus adaptée aux tâches d'intérêt (classification, recherche d'information). Un texte peut naturellement être interprété comme une chaîne de caractères ou mots, qui composent un document. Le mot forme l'unité sémantique naturelle du texte. Un document est alors défini par l'ensemble des mots qui le composent et un ensemble de documents forment un corpus. Mais cette représentation ne peut être utilisée telle quelle par une machine. Cela s'explique par la difficulté à

maintenir de façon efficace l'ordre entre les mots pour la plupart des corpus de documents. De plus, maintenir cet ordre s'avère non significatif pour la plupart des tâches d'intérêt (Aggarwal, 2015). Plutôt que de représenter un document comme une séquence de caractères ou de mots, on utilise donc une représentation du texte plus adaptée aux analyses automatisées : celle d'un vecteur en haute dimension. Un document est alors représenté par un vecteur dont chaque composante correspond à une caractéristique d'un mot du document. On présente différentes méthodes pour calculer les composantes de ce vecteur.

La méthode de calcul la plus simple est celle du sac de mots (Martin et Jurafsky, 2009). Chaque composante sera égale au nombre d'occurrences du mot au sein de ce document (pouvant être normalisé par la longueur du document). L'indice du mot dans le vocabulaire correspondra à la dimension du vecteur représentant chaque document. Le principe est illustré par la Figure 2.4. Une variante binaire est aussi possible où la composante sera égale à 1 si le mot est présent dans le document et 0 sinon. Certaines variantes proposent de normaliser le compte à l'aide d'une fonction logarithmique. On peut de plus intégrer des bigrammes, trigrammes voire des n-grammes dans le vocabulaire d'un document. Cela permet par exemple de considérer l'entité « New York » comme un unique jeton. En pratique, pour les modèles d'analyse de sentiment, de modélisation de thèmes et de fouille de documents, inclure des n-grammes améliore peu les performances des algorithmes (Hopkins et King, 2007 ; Manning *et al.*, 2009).

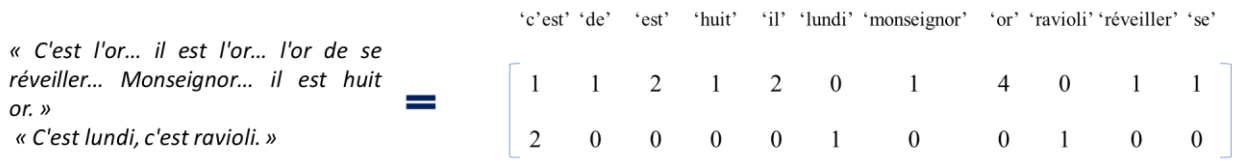

Figure 2.4 Exemple de représentation sac de mots

Avec ce choix de représentation, le texte devient une donnée multidimensionnelle, dont les dimensions sont les mots du corpus. Le corpus de textes est représenté sous la forme d'une matrice de taille (nombre de documents x taille du vocabulaire). Pour un corpus de textes de taille modéré, le vocabulaire obtenu comporte entre 10 000 et 50 000 termes (Grimmer et Stewart, 2013). Chaque document contient peu de mots relativement à l'ensemble du vocabulaire utilisé dans le corpus. La matrice représentant le corpus est donc *creuse* : autrement dit, la plupart de ses coefficients sont nuls. Cette propriété sert de motivation dans le développement de modèles spécifiques pour l'analyse de textes.

Une variante du sac de mots consiste à normaliser cette fréquence par la fréquence du mot dans l'ensemble du corpus : c'est la représentation tf-idf (Manning *et al.*, 2009). L'intérêt de cette variante est d'ajuster le poids donné aux mots rares et aux mots fréquents et de limiter le vocabulaire final. On considère notamment que les mots apparaissant dans beaucoup de documents auront peu de valeur discriminante et donc peuvent être retirés du vocabulaire. La transformation tf-idf pondère pour chaque mot *i* entre sa fréquence au sein du document *j* ($tf_{,ij}$) et sa fréquence au sein du corpus (idf_i, calculée comme l'inverse de la proportion de documents du corpus qui contiennent le mot *i*).

$$tfidf_{i,j} = tf_{i,j} \times idf_i \; = \; tf_{i,j} \; \times \; log \; \frac{|D|}{\left|d_j : t_i \in d_j\right|}$$

Équation 2.5 Formule de la mesure tf-idf

En pratique, la perte du contexte local dans la représentation sac de mots ou tf-idf n'empêche pas d'obtenir des résultats utiles en analyse exploratoire qualitative, en particulier dans le cas d'un petit nombre de données et de puissances de calcul modérées (Hopkins et King, 2007). Pour permettre la mesure de similarité entre les jetons, de nouvelles méthodes ont été développées lors de la dernière décennie. En particulier, depuis l'introduction de l'algorithme Word2Vec en 2013 (Mikolov *et al.*, 2013), certains modèles utilisent une représentation distribuée des mots (Pennington *et al.*, 2014 ; Peters *et al.*, 2018). L'hypothèse justifiant ces modèles est de considérer que des mots de sens proche seront utilisés dans un contexte similaire. Le vecteur de représentation sera alors calculé par un algorithme qui varie en fonction des méthodes. Chaque mot est représenté par un vecteur d'une taille *D* fixée par le chercheur. L'avantage de ces algorithmes dits de « plongement lexical » est de limiter la taille du vecteur représentant chaque mot et d'inclure une notion de contexte dans la représentation en lieu et place du compte fréquentiel : ils intègrent une information sémantique et syntaxique. D'autres modèles incorporent une information à l'échelle des caractères plutôt que des mots (Bojanowski *et al.*, 2017). Enfin, des réseaux de neurones profonds comme des réseaux récurrents ont été utilisés pour améliorer les représentations vectorielles des mots (Devlin *et al.*, 2018 ; Melamud *et al.*, 2016).

Réduction de dimension. Les représentations informatiques du texte se trouvent dans des espaces de haute dimension : le vocabulaire comporte au moins 10 000 jetons. Chaque document nécessitera donc un vecteur avec plus de 10 000 composantes. Dans le cas de représentations

vectorielles de mots, chaque mot aura aussi besoin d'un vecteur pour le représenter. Ce nombre important de caractéristiques engendre un traitement long et un stockage mémoire coûteux. De plus, il est impossible pour un humain de visualiser et donc interpréter les résultats de ces modèles de façon intuitive. Pour faire face à ces deux problèmes ont été développés des algorithmes de réduction de dimension. *L'Analyse en Composantes Principales* (ACP ; Jolliffe, 2011) est la technique la plus utilisée pour procéder à une réduction de dimensions. À partir des caractéristiques existantes, elle vise à former de nouvelles variables non corrélées et maximisant la variance de l'ensemble. C'est une méthode algébrique qui construit de nouvelles caractéristiques d'information maximale à l'aide de projections orthogonales des données. L'ACP peut être utilisée en amont d'un modèle supervisé comme moyen d'extraire un nombre de caractéristiques limité, ou comme un moyen de réduire le bruit obtenu dans les résultats et de limiter le surapprentissage. Sa principale limite en est la complexité computationnelle (Sharma et Paliwal, 2007). L'algorithme *t-SNE* (Maaten et Hinton, 2008) est une méthode probabiliste non linéaire qui maintient la structure locale des données entre l'espace initial de haute dimension et l'espace de projection plus petit. Il cherche à conserver la distance entre les points de données entre les deux espaces. La méthode t-SNE est essentiellement utilisée pour des fins de visualisation.

Le choix d'une représentation dépend finalement de la tâche que l'on souhaite effectuer. Nous détaillons maintenant les deux tâches traditionnelles concernant les documents de sciences sociales.

2.2.2 Applications en sciences humaines et sociales

Le texte sert de média pour toute communication en SHS, que ce soit par la littérature savante (articles, livres, rapports…) ou grise (non répertoriés et informels). La quantité de littérature disponible rend le coût d'analyse manuelle exorbitant. Les algorithmes d'analyse automatique visent d'une part à résoudre ce problème de *volume* de données à analyser en filtrant l'information utile selon certains critères. D'autre part, ils peuvent être construits de sorte à dégager des *tendances* nouvelles, impossibles à identifier *a priori* pour des humains. Les outils d'analyse de données automatique répondent donc la plupart du temps à des objectifs qualitatifs (étude exploratoire, visualisation, tendances générales d'un corpus) atteints à l'aide méthodes quantitatives (optimisation et mesures d'évaluation statistique).

Les applications des algorithmes d'analyse de texte peuvent se scinder en cinq catégories (Clark *et al.*, 2010) :

1) **Classification de contenu.** On souhaite organiser un corpus selon certains critères connus *a priori* ou à déterminer.
2) **Extraction d'information.** Le but est de parcourir et retourner les documents d'un corpus les plus pertinents en fonction d'un critère de recherche.
3) **Traduction et correction automatique.** L'objectif est d'établir une correspondance sémantique, syntaxique et orthographique entre plusieurs langues.
4) **Génération automatique de texte.** On souhaite générer du texte dans une langue aussi proche que possible de la langue humaine.
5) **Réponse aux questions.** On veut construire un système informatique capable d'analyser et répondre à des questions en langage naturel.

On se centre ici sur les deux premières catégories qui contiennent l'essentiel des applications en SHS actuelles et au sein desquelles se situe le travail de ce mémoire. Après avoir présenté comment organiser un corpus de documents, on exposera comment y rechercher de l'information. Naturellement, ces deux tâches sont dépendantes : vous trouverez toujours plus facilement ce que vous cherchez si votre base de recherche est bien structurée.

2.2.2.1 Classification automatique de documents

Pour faciliter la recherche d'un document dans une bibliothèque, les livres sont souvent rangés selon leur thématique (économie, mathématique, etc.) puis selon une cote spécifique (nom d'auteur, année de publication, etc.). Au XXe siècle, cette classification thématique faisait partie du travail de bibliothécaire et était faite manuellement. Avec la numérisation massive des contenus textuels, il devient nécessaire de disposer d'un système d'organisation de documents automatique capable de gérer des millions de documents. Passer de la classification manuelle à la classification automatique implique de procéder à des choix méthodologiques et à définir des hypothèses de modélisation. En fonction des données étudiées, on peut procéder à de la classification *supervisée* ou *non supervisée*. La classification supervisée de documents correspond au cas où les catégories d'organisation sont connues *a priori*. Par exemple, dans le cas d'articles, on sait dans quel journal a été publié chaque article. La classification non supervisée est plus exploratoire : l'algorithme va

lui-même regrouper les documents selon certains critères de similarité. On parle aussi de regroupement (*clustering*). Les méthodes de regroupement sont abordées dans la section 2.3, où l'on détaille en particulier l'une d'entre elles qui a été utilisée dans ce travail : les modèles de thème. Deux configurations de problèmes sont ensuite possibles : classification dure (*hard*) où une unique catégorie est attribuée à chaque document ou classification élastique (*soft*) où l'on attribue à chaque document une probabilité d'appartenance à chaque catégorie. La première est plus simple d'analyse mais la seconde est plus riche en information. On présente ici le cas de la classification supervisée et ses applications en sciences sociales. Le problème est le suivant : des documents sont étiquetés par des humains selon un ensemble de catégories prédéfinies (date, journal, thématique, etc.). Ces documents forment ce qu'on a appelé dans la section 2.1.2 l*'ensemble d'entraînement*. Le but est de construire un algorithme de classification qui va pouvoir discriminer chaque document selon son étiquette. À partir des caractéristiques (des mots par exemple) des documents d'une classe, l'algorithme déduit des propriétés spécifiques de la classe qui la différencie des autres. La représentation calculée par l'algorithme est ensuite évaluée sur l'ensemble de test, composé d'un autre ensemble de documents munis des mêmes étiquettes. Si l'algorithme est pertinent, il saura retrouver les catégories des documents de l'ensemble de test. On utilisera classiquement les métriques d'exactitude, de précision et de rappel pour évaluer sa performance.

Aggarwal et Zhai (2018) distinguent 5 grandes familles d'algorithmes de classification :

1) **Classifieurs à base de règles.** Ce classifieur détermine des patrons de mots qui apparaissent souvent dans des documents ayant les mêmes étiquettes. Ces patrons servent à construire manuellement un ensemble de règles permettant de déterminer la classe de nouveaux documents.

2) **Arbres de décision.** Ce classifieur utilise une représentation hiérarchique des données en utilisant certaines des caractéristiques comme règles successives de classification. Les différentes catégories possibles sont situées sur les feuilles de l'arbres et sont atteintes en fonction des décisions prises à chaque nœud.

3) **Machine à vecteurs de support (*SVM*).** Ce classifieur très populaire sépare l'espace de représentation des documents selon chacune des classes à l'aide de frontières. L'idée de ces classifieurs est de trouver la frontière de séparation qui maximise l'écart entre des documents de classes différentes.

4) **Réseaux de neurones artificiels.** Les réseaux de neurones adaptés au texte servent à discriminer les classes en utilisant les caractéristiques des mots. Contrairement aux deux premiers algorithmes, aucune règle de classification n'est déterminée manuellement car celles-ci sont calculées par l'algorithme. De plus, à l'aide de couches de calcul successives les réseaux de neurones peuvent effectuer des opérations non linéaires pour séparer efficacement les données.

5) **Classifieurs Bayésiens.** Ce classifieur cherche à inférer le processus de création des documents observés. Il utilise ensuite ce processus (qui correspond à une probabilité postérieure) comme élément discriminant chaque classe.

Les trois premières catégories sont des classifieurs dits « discriminants », dans le sens où ils cherchent à maximiser la performance de classification sur un ensemble de test. Les classifieurs bayésiens visent plutôt à approximer la distribution de probabilité des données en maximisant la vraisemblance de les observer. Les réseaux de neurones sont initialement de nature discriminante mais peuvent être convertis en réseaux bayésiens (Neal, 2012).

Un critère de choix d'algorithme de classification est le compromis entre interprétabilité et puissance du modèle. Les classifieurs à base de règle sont facilement interprétables car leur processus de décision est explicite. En revanche, les réseaux de neurones bien que plus performants sont des boîtes noires de calcul et posent des problèmes de transparence dans l'apprentissage. Pour un détail des avantages et inconvénients de chaque méthode, on renvoie le lecteur à la revue de littérature proposée par Kowsari *et al.* (2019). Il existe de nombreux autres algorithmes de classification (k-voisins, forêts aléatoires, algorithmes génétiques, etc.). Plutôt que de détailler le fonctionnement technique de chacun des algorithmes, nous préférons exposer les principales applications de ces algorithmes dans le domaine des sciences sociales.

Une première application d'intérêt est la *classification par thématique.* Hillard *et al.* (2007) ont identifié quatre caractéristiques que doit comporter un système de classification par thématique. Celui doit avoir un pouvoir *discriminant* suffisant, autrement dit les thèmes doivent être mutuellement exclusifs et chaque document doit être clairement identifié comme traitant d'un thème en particulier. Ensuite, il doit avoir une bonne *exactitude* : les thèmes attribués à chaque document doivent représenter le contenu global du document. Une mesure de performance adaptée au système doit être accessible. De plus, le système doit être *fiable* : des documents partageant une

thématique similaire écrits à différentes périodes devraient recevoir la même étiquette thématique. Cela implique que le modèle soit capable d'identifier des changements de vocabulaire pour un même thème au cours du temps. Enfin, le système doit être *efficace*, peu coûteux à implémenter et à stocker. Hopkins et King (2010) proposent des méthodes adaptées aux SHS pour construire un tel système. Wang et Manning (2012) étudient la variabilité des classifieurs bayésien et SVM sur différentes banques de documents et différentes configurations méthodologiques.

Une autre application fréquente des algorithmes de classification est *l'analyse de sentiment.* Prenons l'exemple d'une analyse de sentiment à deux classes : pour chaque document du corpus, on cherche à le classer comme positif ou négatif. Une des méthodes initialement utilisées pour cette tâche a été la méthode à base de dictionnaire (Schrodt *et al.*, 1994). Le principe de ces méthodes à base de dictionnaire est d'utiliser les mots composant les documents comme indicateur de classification. Un premier algorithme permet d'attribuer une pondération positive ou négative à chaque mot du vocabulaire (discrète ou continue). Puis, à l'aide des fréquences des mots dans chaque document, on attribue une classe au document (Eshbaugh-Soha, 2010). Un certain nombre de dictionnaires ont été établis, attribuant différentes mesures à chaque mot d'un vocabulaire, et peuvent être réutilisés pour de nouvelles données ou questions de recherche. Toutefois, un dictionnaire construit sur un ensemble de données peut s'avérer inutile sur un autre ensemble de documents. Loughran et McDonald (2011) montrent que des dictionnaires standards ne prennent pas en compte le contexte spécifique de certains corpus : les mots *coût* ou *taxe* peuvent par exemple avoir une connotation positive dans des rapports financiers. Ces méthodes sont limitées par le coût de construction manuelle du dictionnaire, la difficile validation des résultats et la faible capacité de généralisation sur d'autres données. En pratique, elles ne sont utilisées que pour de la classification binaire. Plus récemment, Nobles *et al.* (2018) utilisent un classifieur basé sur un réseau de neurones profond pour identifier un risque de suicide dans des messages textes. On citera aussi Ofoghi et Verspoor (2017) qui évaluent un classifieur d'émotions à partir de messages Twitter ; Tang *et al.* (2015) qui mènent une analyse de sentiments de commentaires IMDB à l'aide d'un réseau de neurones récurrent et Schrodt (2000) qui détermine si des déclarations étatiques sont de nature pacifique ou belliqueuse. Enfin, comme toute tâche d'apprentissage automatique, la classification doit être évaluée pour permettre la comparaison des modèles et leur utilité opérationnelle. Sebastiani *et al*. (2002) proposent un tour des méthodes d'évaluation des algorithmes de classification.

Une fois les documents classifiés, on peut efficacement fouiller la collection selon certains critères spécifiques : c'est la tâche de recherche d'information.

2.2.2.2 Recherche d'information

La recherche d'information est la tâche de « *trouver du contenu (souvent des documents) de nature non structurée (souvent du texte) qui satisfont un besoin d'information parmi une large collection.* » (Manning *et al.*, 2009). Prenons l'exemple suivant : sur une base de données d'articles dans le domaine de l'économie, un chercheur s'intéresse aux articles traitant de politique monétaire en Afrique. Le chercheur entre sa *requête* « politique monétaire en Afrique » et l'algorithme se propose de retourner un ensemble de documents *pertinents* en rapport à cette requête.

On présente succinctement les méthodes existantes pour construire un système de recherche de documents. On se concentre sur la recherche d'information parmi une collection de documents appartenant à un domaine spécifique et d'une taille modérée. C'est le cas de données stockées sur un serveur unique centralisé auquel on accède avec quelques machines. Pour des détails sur la recherche d'information sur le Web et comment gérer de l'information à cette échelle, on renvoie le lecteur à Manning *et al.* (2009). Les systèmes de recherche d'information comportent deux volets : la construction d'un index inversé puis la récupération de documents en fonction d'une requête (parcours de l'index et mesure de pertinence).

Par exemple, si l'on s'intéresse à la thématique « économie » au sein d'une collection d'articles, une idée intuitive est de parcourir la collection et de retourner tous les documents qui contiennent le mot « économie ». Rappelons que la seule représentation du corpus dont l'ordinateur dispose pour l'instant est une matrice documents-mots, où chaque composante est la fréquence du mot dans le document. Pour que la fouille soit efficace, il nous faut disposer d'une matrice mots-documents en amont de la requête. Cette matrice sera essentiellement remplie de 0, ce qui permet des approches de construction efficace (Aggarwal, 2015). La méthode est similaire à la construction de la matrice documents-mots : après nettoyage du texte, on crée un dictionnaire qui associe à chaque mot la liste de documents où il apparaît. Plusieurs structures sont possibles pour stocker la liste de documents de chaque mot (voir Manning *et al.*, 2009). Une fois cet index inversé construit, on souhaite lui adresser une requête. Plusieurs algorithmes de parcours de l'index inversé sont possibles. L'idée générale est de retourner l'intersection des listes de documents qui contiennent les mots « politiques monétaires en Afrique ». Ce système de recherche est dit « booléen ». L'index

inversé nous indique si un mot apparaît ou non dans un document. La requête précédente devient par exemple : trouver les documents qui contiennent « politiques ET monétaires ET Afrique ». De nombreuses améliorations sont possibles à ce système trivial. Les systèmes actuels permettent notamment quatre possibilités additionnelles (Manning *et al.*, 2009) :

1) Ils considèrent la fréquence d'apparition d'un mot en plus de sa présence ou absence dans un document.

2) Ils intègrent des n-grammes dans le vocabulaire pour pouvoir rechercher des concepts comme « économie internationale » (Evans *et al.* (1991) ; Mitra *et al.* (1997)). Ils peuvent aussi intégrer des connaissances externes au contenu même des documents comme des liens relationnels entre les concepts (avec la base *Wordnet*® pour l'anglais) ou des métadonnées des documents (comme la langue).

3) Ils intègrent une tolérance orthographique et une approximation sémantique dans la requête de l'utilisateur.

4) Ils proposent une mesure de hiérarchisation des résultats, qui se met à jour en fonction des choix passés de l'utilisateur.

Le point clé permettant ces améliorations est de passer d'une modélisation ensembliste (où l'on accède aux documents avec des opérations logiques) à des modèles vectoriels ou probabilistes. L'approche vectorielle s'intéresse à construire des mesures de similarité entre les vecteurs représentant chaque document dans l'espace du vocabulaire (Wong *et al.*, 1985). Un document sera pertinent s'il est « proche » de la requête dans cet espace. Un exemple de mesure de pertinence est la mesure de similarité cosinus. Elle consiste à calculer le produit scalaire normalisé entre le vecteur de la requête et le vecteur de chaque document. Plus ce produit scalaire sera élevé, plus le document sera pertinent par rapport à la requête. Un exemple d'approche vectorielle est le modèle Analyse sémantique latente (LSA) que l'on présente en section 2.3.

L'approche probabiliste cherche à déterminer la vraisemblance qu'un document du corpus soit pertinent conditionnellement à la requête de l'utilisateur (Callan *et al.*, 2017 ; Turtle et Croft, 1991). L'intérêt de cette approche est de modéliser l'incertitude dans la compréhension de la requête par la machine. Le modèle original et encore utilisé aujourd'hui est le Binary Independence Model (Yu et Salton, 1975). Il approxime chaque document par un vecteur binaire comptant l'absence ou présence des mots. Puis il retourne les documents pertinents après avoir calculé le produit scalaire

entre les représentations binaires de la requête et des documents. Un modèle plus récent est le modèle Latent Dirichlet Allocation (LDA), sur lequel se fonde ce travail et que l'on développe en section 2.3. Wei et Croft (2006) et Wang *et al.* (2007) montrent par exemple l'efficacité d'un modèle LDA dans l'extraction d'information sur des corpus TREC. Une autre approche probabiliste moderne qui s'est avérée très efficace en recherche d'information est les modèles de langue. L'idée est la suivante : un document sera pertinent s'il est susceptible de générer la requête. Autrement dit, à partir d'un modèle probabiliste du document, on calcule la vraisemblance d'observer la requête de l'utilisateur (Ponte et Croft, 1998). La limite de ces modèles est de ne pas inclure de mesure de pertinence explicite et d'utiliser des modèles de langue simples (unigramme).

Quelle que soit l'approche et tout comme la tâche de classification, des mesures d'évaluation sont nécessaires. La différence avec la tâche de classification est l'interaction fondamentale avec l'utilisateur. Un système de recherche d'information performant permet à l'utilisateur de trouver ce qu'il cherche rapidement. Pour mener l'évaluation d'un système de recherche d'information ad hoc, il faut trois composantes : une collection de documents à interroger, un besoin d'information (transmis sous la forme de requête) et un retour de pertinence de la part des utilisateurs sur les paires requête-documents à l'aide d'une échelle de pertinence. Les bases de données d'évaluation de système de recherche les plus connues sont les collections *Cranfield, Text Retrieval Conference (TREC), NTCIR, Reuters-RCV1 et 20 Newsgroups*. La mesure d'exactitude n'est pas appropriée à l'évaluation d'une tâche de recherche d'information car pour la majorité des requêtes, 99,9 % des documents sont non pertinents (Manning *et al.*, 2009). Maximiser l'exactitude a de grandes chances de mener à une grande proportion de faux négatifs voire à inciter le système à ne plus proposer aucun document. En fonction de l'application, on souhaitera plutôt maximiser la précision (pertinence de tous les résultats proposés) ou le rappel (récupérer tous les résultats pertinents possibles). Un bon système de recherche d'information trouve un compromis satisfaisant pour l'utilisateur entre ces deux mesures, qui varient inversement avec le nombre de documents proposés. C'est pourquoi on utilise fréquemment la moyenne harmonique des deux, appelée le score F1 (van Rijsbergen *et al.*, 2012).

$$F_1 = \frac{2 \times précision \times rappel}{précision + rappel}$$

Équation 2.6 Score F1

Pour une liste hiérarchique de documents, on peut utiliser des courbes précision = f(rappel) pour déterminer la précision moyenne de l'algorithme ou des courbes ROC qui trace le rappel en fonction du taux de faux positifs (Buckley et Voorhees, 2017). Pour une liste hiérarchique de documents, on peut utiliser des courbes précision = f(rappel) pour déterminer la précision moyenne de l'algorithme ou des courbes ROC qui trace le rappel en fonction du taux de faux positifs (Buckley et Voorhees, 2017). Une fois ces mesures établies, il s'agit d'établir la pertinence des résultats de la recherche. Pour des requêtes spécifiques, il est nécessaire de s'adresser à des experts du domaine pour assurer la validité de la réponse. Dans le cas d'un moteur de recherche dynamique, on peut ajuster les poids du modèle latent en fonction des résultats de pertinence des requêtes. L'évaluation humaine étant un processus long et coûteux, on choisit un sous-ensemble de requêtes à évaluer.

Ensuite, il s'agit d'évaluer la variabilité des réponses données, autrement dit la fidélité interjuges. En effet, les êtres humains sont loin d'être homogènes concernant leur opinion (Swanson, 1988). Pour assurer la validité de la mesure, l'indicateur utilisé doit incorporer le fait que deux évaluateurs peuvent être d'accord par chance. Pour mesurer l'accord entre plusieurs juges évaluant plusieurs items à l'aide d'une échelle cardinale, le test non paramétrique le plus stable, robuste et précis est l'alpha de Krippendorff (Hayes et Krippendorff, 2007). L'alpha de Krippendorff (α) généralise en fait les autres tests non paramétriques d'accord interjuges comme le π de Scott, le κ de Fleiss, le ρ de Spearman et le r de Pearson pour une évaluation avec plus de deux juges. Il est basé sur le ratio entre le désaccord observé et le désaccord qui serait obtenu par chance.

$$\alpha = 1 - \frac{D_o}{D_e}$$

Équation 2.7 Définition du α de Krippendorff

D_o représente le désaccord observé et D_e le désaccord théorique dû uniquement à la chance. La méthode de calcul de ces deux désaccords dépend du type d'échelle utilisée. L'hypothèse nulle est que les désaccords observés sont dus uniquement à la chance. Un α de 0 indique un désaccord total

tandis qu'un α de 1 indique un accord total. Le seuil d'acceptation en recherche en sciences sociales se situe à 0.67 (Krippendorff, 2004).

Une critique émise envers ces mesures de pertinence est qu'elles n'évaluent qu'un seul document à la fois. Or, dans un système de recherche, l'utilisateur ne veut pas récupérer deux fois le même résultat. Pour y remédier, on peut par exemple inclure une mesure de diversité et de nouveauté des résultats (Carbonell et Goldstein, 1998). Par ailleurs, un système de recherche ne peut être évalué uniquement par la pertinence des résultats qu'il propose. D'autres critères sont importants et peuvent être mesurés (Manning *et al.*, 2009) :

1) La latence de l'indexation (nombre de documents indexés par heure par exemple)
2) La latence de la recherche (notamment en fonction de la taille de l'index)
3) Les possibilités de requête pouvant lui être adressées (complexité, langue naturelle, vocabulaire restreint etc.) et l'interface utilisateur
4) La diversité thématique de la base de données

Fondamentalement, l'objectif est de satisfaire les utilisateurs du moteur de recherche. On peut mesurer cette satisfaction à l'aide d'études qualitatives et effectuer une rétroaction sur le système en fonction des résultats.

Que l'on veuille mener une tâche de classification ou une tâche de recherche d'information, il faut d'abord choisir un modèle qui va traiter la collection de documents, l'organiser selon certains critères et être capable d'en extraire l'information souhaitée. L'approche de modélisation probabiliste présentée dans la section 2.1 est couramment utilisée pour effectuer ces tâches (Zhang, 2004). La section qui suit présente le modèle probabiliste retenu pour notre étude : le Latent Dirichlet Allocation.

2.3 Modélisation thématique : le Latent Dirichlet Allocation (LDA)

Dans le cas où l'on ne dispose pas d'étiquettes *a priori* ou que l'étiquetage par des humains est trop coûteux en temps et en argent, on utilise des méthodes d'analyse non supervisées, comme le regroupement. Ces méthodes limitent le besoin d'intervention humaine pour la formation du corpus d'entraînement et la sélection de règles de classification. Les algorithmes de regroupement consistent à identifier des groupes de documents similaires selon une certaine métrique. Différents

niveaux de granularité sont possibles, depuis l'échelle du document jusqu'à l'échelle du mot. Ce regroupement est une méthode efficace pour organiser une collection de documents : par exemple, Cutting *et al.* (1993) construisent ainsi une table de contenu d'un corpus d'articles tandis que Anick et Vaithyanathan (1997) érigent un système de recherche d'information. Pour une présentation détaillée des méthodes de regroupement, on renvoie le lecteur à Kaufman et Rousseeuw (2009). On présente ici les trois familles d'algorithmes de regroupement (Allahyari *et al.*, 2017) distinguent trois types de regroupement :

1) **Regroupement hiérarchique**. Le but est de construire des groupes de documents liés par des relations hiérarchiques. Ils peuvent être de nature descendante ou agglomérative (Murtagh, 1983). Dans le premier cas, on commence avec un groupe contenant tout le corpus et l'on divise progressivement en sous-groupes. Dans le second cas, chaque document correspond initialement à un groupe et l'on fusionne les documents en des plus clusters plus gros à chaque étape. La fusion se fait à l'aide de mesures de similarité entre les documents ou les groupes de documents (voir Aggarwal, 2018 pour un détail de ces mesures). Par exemple, Bolton et Hand (2001) utilisent des regroupements hiérarchiques pour détecter des fraudes sur des cartes de crédit et Quinn *et al.* (2006) l'utilisent pour regrouper des arguments rhétoriques parmi des discours donnés au Congrès américain.

2) **Regroupement à partition.** Le but est de répartir les documents au sein de k clusters sans relation hiérarchique. L'algorithme le plus connu est l'algorithme des k-moyennes. On procède par itérations en ajustant l'appartenance de chaque document à un groupe ainsi que la position de chaque groupe en fonction des documents qui le composent.

3) **Regroupement probabiliste.** Les regroupements probabilistes sont des regroupements élastiques qui utilisent des probabilités d'appartenance à chaque groupe plutôt qu'une attribution déterministe. L'un des modèles les plus fréquents de ce type de regroupement est les modèles de thèmes. On détaille maintenant ces modèles.

Modèles de thèmes

Les modèles de thèmes sont un cas d'algorithme de regroupement probabiliste. Ils sont construits à partir d'une analyse bayésienne des données, c'est-à-dire qu'ils tentent d'inférer le processus génératif d'un ensemble de documents. Une fois ce processus d'inférence complété, on peut valider

et utiliser la représentation calculée par ces modèles pour des tâches de classification et de fouille de documents. Ces modèles ont été pensés pour explorer et appréhender de larges ensembles de données représentés sous forme discrète, par exemple à l'aide d'une représentation sac de mots du texte. Nous présentons d'abord les modèles statiques de thème qui construisent une représentation thématique globale d'un corpus. Ces modèles sont utilisés notamment pour organiser et fouiller des corpus de documents.

Latent Dirichlet Allocation (LDA)

La forme d'un modèle probabiliste dépend du type de données qui nous intéresse, des hypothèses de modélisation de ces données et du but que l'on souhaite atteindre à l'aide du modèle. Dans le cas d'une collection de documents que l'on souhaite classifier, chaque document peut contenir différentes thématiques. Or, le modèle unigramme présenté dans la section 2.1 ne permet d'attribuer qu'un thème pour chaque document. Le modèle pLSI (Indexation Sémantique Latente probabiliste) développé en 2001 par Hofmann (Hofmann, 2001) répond à cette problématique en permettant à un document d'appartenir à un mélange de thèmes. Peu avant ce modèle probabiliste, une version purement algébrique et non probabiliste a été pensée (LSI). De façon similaire à l'ACP, le LSI utilise une décomposition en valeurs singulières de la matrice documents/mots pour déterminer les meilleures caractéristiques discriminantes entre catégories. L'idée latente est que des mots de forte cooccurrence auront un sens similaire. Dans le cas d'une recherche d'information, on calculera le vecteur de projection de la requête dans ce nouvel espace et le comparera aux vecteurs des documents du corpus. Le modèle LSI ne propose toutefois pas de processus génératif à l'échelle du document. Hoffmann a donc développé une version probabiliste du LSI, qui définit une fonction objectif explicite que l'on cherche à maximiser. En l'occurrence, chaque mot est attribué à un thème tandis qu'à chaque thème sont affectés plusieurs mots. Un document est alors représenté comme un mélange de thèmes. On maximise la vraisemblance d'observer les données générées par rapport aux données observées. Le processus d'inférence peut se faire avec l'une des méthodes présentées en 2.1 comme l'algorithme Espérance-Maximisation ou bien une chaîne de Monte-Carlo. En bref, pour chaque mot d'un document d, on tire un thème z selon une multinomiale conditionnée selon le document d. Puis on tire un mot w_d selon une multinomiale paramétrée par z. La Figure 2.5 résume le processus génératif de ce modèle.

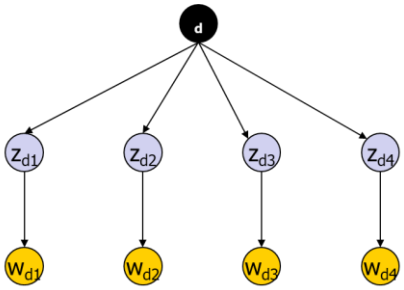

Figure 2.5 Modèle génératif du modèle pLSI

Un problème majeur de ce modèle est qu'il étend mal la représentation à l'échelle du document. De plus, il s'avère propice au surapprentissage : il est en particulier difficilement capable de générer un nouveau document à partir de la représentation entraînée. Pour pallier ces problèmes, le modèle LDA (Blei *et al.*, 2003) étend le modèle pLSI et reste aujourd'hui encore l'un des modèles les plus pertinents pour explorer une collection de documents thématiquement (Griffiths et Steyvers, 2004). Le modèle LDA montre plusieurs avantages théoriques et opérationnels : il permet une flexibilité dans l'attribution des thèmes comparé au modèle pLSI ; le processus d'inférence est relativement simple ; le design est modulaire et facilite la construction de variantes adaptées à de nouvelles tâches.

Le succès du LDA part d'une hypothèse simple : un document traite de peu de thématiques et une thématique peut se caractériser à l'aide de peu de mots. Comme le pLSI, l'objectif du modèle LDA est de proposer un scénario qui est capable *a posteriori* de reconstruire le corpus de documents observé. La stratégie générative du modèle LDA est la suivante : postulant l'existence de K thèmes propres à une collection de documents, le modèle va chercher à trouver le mélange de thèmes qui compose chaque document. Pour cela, il va itérer en attribuant successivement un poids à chaque mot pour chaque thème, puis un poids pour chaque thème à chaque document. À la différence du pLSI et pour permettre une meilleure représentation à l'échelle des documents, on souhaite que le mélange de thèmes pour chaque document soit tiré d'une distribution de probabilité *a priori*. Cette distribution de probabilité *a priori* doit engendrer des multinomiales, elle peut donc être vue comme une distribution de probabilités. Le choix du LDA est d'utiliser pour cela la distribution de probabilité conjuguée à la multinomiale qui est la distribution de probabilité de Dirichlet, ce qui permet à la distribution de probabilité postérieure d'avoir la même forme que la distribution de probabilité *a priori*. Les étapes d'entraînement du modèle sont détaillées à la Figure 2.6.

1) On tire des proportions β_1, …, β_K à partir d'une loi de Dirichlet de paramètre η. À la fin de cette étape, chaque thème a été initialisé avec une proportion pour chaque mot du vocabulaire, proportion contenue dans les vecteurs β_1 à β_K.
2) Ensuite on itère pour chaque document *d* de la collection :
    a. On initialise un mélange de thèmes θ_d à partir d'une loi de Dirichlet de paramètre α. À la fin de cette étape, chaque document a été initialisé avec une proportion de chaque thème, contenue dans le vecteur θ_d.
    b. Pour chaque mot *n* du document *d* :
        i. On tire un thème $z_{d, n}$ à l'aide d'une multinomiale paramétrée par θ_d.
        ii. On tire un mot $w_{d, n}$ à partir d'une multinomiale paramétrée par $\beta_{z, d, n}$.

Figure 2.6 Processus génératif du LDA

À la fin de ces deux sous-étapes, on a ajusté la proportion des N mots du vocabulaire au sein des K thèmes et la proportion des K thèmes au sein des D documents.

Les vecteurs d'hyperparamètres η et α sont définis à l'échelle du corpus et servent de paramètres à la loi de Dirichlet que l'on choisit en amont de l'entraînement. η correspond à l'*a priori* que l'on donne au modèle concernant la distribution de probabilité de mots au sein de chaque thème. Un η élevé engendre une distribution de probabilité initiale de mots uniforme dans chaque thème, autrement dit chaque mot est attribué indifféremment à chaque thème. Un η faible engendre une distribution de probabilité de mots initiale fortement discriminante selon les thèmes. α correspond à l'*a priori* que l'on donne au modèle concernant les mélanges de thèmes dans chaque document. Un α élevé engendre une distribution de probabilité initiale de thèmes similaire entre les documents et proche d'une distribution de probabilité uniforme. Un α faible engendre une distribution de probabilité initiale des thèmes présentant une forte variance entre les documents.

Les vecteurs de paramètres β_i contiennent les proportions initiales de mots pour chaque thème tandis que les vecteurs de paramètres θ_d contiennent les proportions initiales des mélanges de thèmes pour chaque document *d*. Ces différents paramètres sont liés par la probabilité jointe suivante :

$$p(\beta, \theta, Z, W) = \left(\prod_{i=1}^{K} p(\beta_i|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})\right)$$

Équation 2.8 Probabilité jointe du modèle LDA

Cette probabilité jointe contient le processus génératif décomposé selon les différentes étapes d'entraînement. On peut le visualiser à l'aide des Figure 2.7 et Figure 2.8.

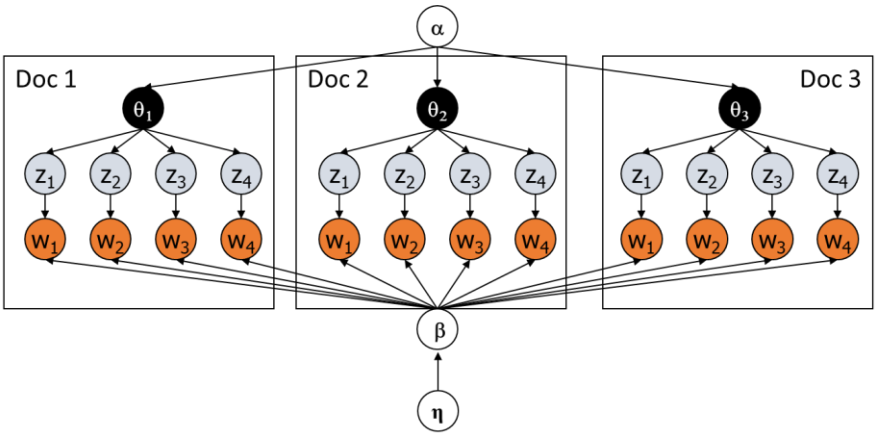

Figure 2.7 Modèle génératif du modèle LDA

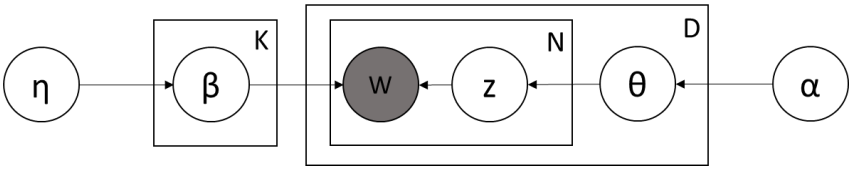

Figure 2.8 Diagramme de plaques du modèle LDA

En pratique, on n'observe que les mots du corpus, via une représentation sac de mots. À de chaque document. À partir de ces mots, on veut donc déterminer la distribution de probabilité des mots pour chaque thème (matrice de paramètres **β)**, les mélanges de thèmes pour chaque document (paramètre $\boldsymbol{\theta_d}$) et constituer le processus génératif des documents (paramètre **z**). Cela revient à estimer la valeur de la distribution de probabilité postérieure suivante (où l'on a retiré les hyperparamètres α et η pour une meilleure lisibilité) :

$$p(\beta, \theta, Z|W) = \frac{p(W \mid \beta, \theta, Z) + p(\beta, \theta, Z)}{p(W)}$$

Équation 2.9 Probabilité postérieure du modèle LDA

On fait alors face à un problème de calcul. En effet, cette distribution de probabilité ne peut être déterminée exactement car la constante de normalisation p(W) est insoluble (Dickey, 1983). Des méthodes simples pour approximer cette distribution de probabilité postérieure ont été présentées dans la section 2.1. À cause du nombre de paramètres à estimer dans le cas du modèle LDA, d'autres méthodes sont utilisées : l'échantillonnage de Gibbs (Blei *et al.,* 2003) et les méthodes variationnelles (Hoffman, 2010 ; Kingma et Welling, 2013 ; Blei *et al.,* 2017 ; Srivastava et Sutton, 2017) sont les plus populaires.

Échantillonnage de Gibbs. C'est un cas particulier de MCMC (Carlo, 2004), exposées rapidement en section 2.1.1. L'échantillonnage de Gibbs s'applique aux modèles probabilistes disposant de plusieurs paramètres comme dans le modèle LDA. Le principe est d'approximer la distribution de probabilité postérieure par une suite d'échantillonnages de la distribution de probabilité conditionnelle. Dans le cas du LDA, la distribution de probabilité conditionnelle attribue les mots au sein de chaque thème. Pour chaque mot $w_{d,n}$ du document *d* et du thème *k*, l'échantillonnage est paramétré par quatre facteurs : le nombre de mots du document *d* déjà attribués au thème *k*, le nombre de fois où le mot $w_{d,n}$ a déjà été attribué au thème *k* et les hyperparamètre de Dirichlet α et η. En bref, l'échantillonnage pondère deux objectifs : l'affinité entre le document *d* et le thème *k* et l'affinité entre le thème *k* et le mot $w_{d,n}$. Ces échantillons sont ensuite intégrés pour récupérer des valeurs pour la distribution de probabilité postérieure qui nous intéresse. Une limite majeure de cette méthode est que l'on ne dispose d'aucun critère stable de convergence.

Inférence variationnelle. Les méthodes d'inférence variationnelle ont été développées afin de converger vers une solution satisfaisante plus rapidement que les méthodes d'échantillonnage (Jordan *et al.,* 1999). Elles consistent à formuler le problème d'inférence comme un problème d'optimisation. L'idée des méthodes variationnelles est d'approximer la distribution de probabilité postérieure par une famille de distribution de probabilité plus simple, caractérisée par des paramètres variationnels ν. L'objectif est alors de trouver la distribution de probabilité $q_\nu(x)$ la plus proche de la probabilité postérieure p(x|y) selon une divergence KL (Kullback et Leibler, 1951) :

$$argmin_\nu \, KL(q_\nu||p) = \; argmin_\nu \int q_\nu(x) log \frac{q_\nu(x)}{p(x|y)} dx$$

Mais par principe, on ne connaît pas la valeur de la distribution de probabilité postérieure p(x|y) ! Au lieu de minimiser cet objectif, on va donc optimiser une borne inférieure L_ν (méthode ELBO),

indépendante de la distribution de probabilité postérieure. Le choix de la famille q_v dépendra de l'algorithme d'inférence utilisé. Une hypothèse usuelle est de supposer que la famille q_v peut se factoriser entièrement selon des distributions de probabilités marginales : c'est l'hypothèse de l'inférence variationnelle des champs moyens. Les familles trouvées par cette hypothèse s'avèrent utilisables en pratique (Gerrish et Blei, 2011 ; Jordan *et al.,* 1999). Une fois la famille q_v déterminée, on développe le calcul de L_v et de son gradient pour l'exprimer en fonction de v. Il faut parfois borner cette borne elle-même pour aboutir à un résultat exploitable. Un inconvénient des méthodes variationnelles est son manque de flexibilité pour un nouveau modèle. En effet, pour chaque nouveau design il faut : 1) Choisir une famille de distribution de probabilité q_v ; 2) Former un objectif à optimiser (à l'aide de bornes) ; 3) Calculer les gradients de cet objectif ; 4) Entraîner l'algorithme. Pour des améliorations récentes dans les méthodes variationnelles utilisant en particulier des réseaux neuronaux, on renvoie le lecteur à Blei *et al.* (2017).

Évaluation interne et externe

Une fois l'inférence accomplie, il est nécessaire d'évaluer le modèle. Nous avons présenté dans la section 2.1.2 la mesure de perplexité, souvent utilisée pour évaluer les modèles probabilistes. La perplexité mesure une certaine capacité de généralisation globale du modèle face à un ensemble de données non observées pendant l'entraînement. Elle a une finalité prédictive plus qu'explicative. Un modèle LDA avec une valeur de perplexité faible est validé comme un bon modèle génératif. Mais cela n'assure en rien l'interprétabilité des thèmes proposés par le modèle et ne permet pas de filtrer les thèmes incohérents ou redondants. En fait, une corrélation négative a même été trouvée entre le jugement humain et la mesure de perplexité par Chang *et al.* (2010). Notons que Mimno et Blei (2011) ont introduit une méthode d'analyse bayésienne pour évaluer la qualité des thèmes extraits mais celle-ci reste une mesure globale de qualité.

Pour pallier aux limites de la mesure de perplexité et valider l'interprétabilité de chaque thème issu du LDA par des humains, des mesures de *cohérence* ont été développées. Elles reposent essentiellement sur l'hypothèse que des mots avec des sens similaires apparaîtront dans des contextes similaires. Une mesure de cohérence « valide » doit montrer une corrélation satisfaisante avec le jugement humain. Newman *et al.* (2010) proposent les premiers une mesure de cohérence qui repose sur l'information mutuelle ponctuelle (*PMI)* et des sources extérieures de connaissance

comme Wikipédia et Google : c'est la cohérence UCI, qui est par nature extrinsèque. Elle est définie pour le thème *k* et l'ensemble W_k des mots représentatifs du thème :

$$C_{UCI} = c(k, W_k) = \sum_{w_i \neq w_j} log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}$$

Où les probabilités de cooccurrence $p(w_i, w_j)$ sont calculées à l'aide des bases de connaissances externes et ε sert de constante de lissage. Intuitivement, la mesure UCI compte parmi les articles de Wikipédia les cooccurrences des mots-clés représentatifs de chaque thème et les normalise par leur fréquence dans l'ensemble de la base de données de Wikipédia. Par exemple, si les termes « sport » et « équipe » sont souvent utilisés dans le même article Wikipédia, alors leur PMI sera forte et si un thème qui attribue un poids important aux deux alors il sera considéré comme cohérent. Newman *et al.* (2010) évaluent la corrélation entre les scores de cohérence pour chaque thème et les résultats d'évaluation humaine où l'on demande d'évaluer chaque thème selon une échelle de pertinence à trois points : bonne, neutre, mauvaise. Ils obtiennent des scores de corrélation supérieurs au seuil de concordance interjuges (0.6).

Mimno *et al.* (2011) proposent une autre mesure pour la cohérence d'un thème *k* contenant l'ensemble des mots dominants W_k :

$$C_{UMass} = c(k, W_k) = \sum_{w_i \neq w_j} log \frac{d(w_i, w_j) + \epsilon}{d(w_i)}$$

Où $d(w_i)$ est le nombre de documents qui contiennent le mot w_i, $d(w_i, w_j)$ est le nombre de documents qui contiennent à la fois les mots w_i et w_j et ε est de nouveau une constante de lissage. Avec cette définition, un thème est cohérent si les mots qui le composent avec le plus de poids ont une forte cooccurrence parmi l'ensemble des documents du corpus. C'est une mesure intrinsèque au corpus contrairement à la mesure UCI, qui s'appuie sur Wikipédia. Cette mesure est capable d'identifier des thèmes peu interprétables mais n'est pas bonne pour identifier les thèmes facilement interprétables. Cela s'explique essentiellement par le fait que de nombreux thèmes sont composés en majorité de mots arrêts qui ont une forte cooccurrence et que la mesure C_{UMass} est incapable de distinguer les mots fréquents des mots réellement informatifs (Nikolenko *et al.*, 2017). Notons que pour ces deux mesures, Stevens *et al.* (2012) montrent qu'une petite constante de lissage ε montre une meilleure performance que la valeur ε = 1 choisie initialement.

Parmi les nombreuses autres mesures de cohérences développées, on citera Aletras et Stevenson (2013) qui utilisent des vecteurs de contexte pour représenter les mots les plus représentatifs de chaque thème et construisent une variation de la cohérence C_{UCI} nommée C_{NPMI}.

$$C_{NPMI} = c(k, W_k) = \sum_{w_i \neq w_j} \frac{log \frac{p(w_i, w_j) + \varepsilon}{p(w_i)p(w_j)}}{-log(p(w_i, w_j) + \varepsilon)}$$

Enfin, Röder *et al.* (2015) ont proposé un cadre unificateur de l'ensemble de mesures de cohérences et l'ont proposé une nouvelle mesure agrégée appelée cohérence C_V. Les auteurs ont paramétré la cohérence C_V après avoir exploré l'ensemble des configurations possibles de leur cadre unificateur. Elle combine une mesure cosinus avec la cohérence C_{NPMI} et une fenêtre de glissement pour évaluer les probabilités de co-occurrence.

Évaluation sur une tâche spécifique. Enfin, certaines approches favorisent l'évaluation externe des représentations thématiques, à l'aide de tâches spécifiques. Cela ne permet pas de qualifier directement l'espace latent calculé par le LDA mais plutôt de valider l'utilité du modèle sur un besoin opérationnel. Ainsi, Titov et McDonald (2008) effectuent une analyse de sentiment non supervisée sur des avis utilisateurs à partir d'un modèle LDA capable d'extraire des thèmes corrélés à des sentiments positifs ou négatifs. Wei et Croft (2006) construisent un système d'extraction d'information qui repose sur les thèmes issus d'un modèle LDA. Chang *et al*. (2010) proposent deux nouvelles tâches d'évaluation qui seront fréquemment reprises par la suite par des chercheurs en modélisation thématique. La première consiste à proposer une liste de mots-clés représentatifs d'un thème, d'y insérer un *mot intrus* issu du vocabulaire et demander à des évaluateurs humains d'identifier le mot intrus pour différents thèmes. La seconde tâche consiste à introduire un *thème intrus* parmi le mélange de thèmes qui composent un document et demander à des évaluateurs humains d'identifier ce thème intrus pour différents documents. Le travail de Lau *et al*. (2014) établit un lien entre les mesures de cohérence et l'évaluation humaine en comparant le classement de cohérence établi par les mesures automatiques à celui établi par des humains.

Extensions du LDA

Depuis sa création, le modèle LDA s'est révélé utile en phase d'exploration de données et facile à adapter à de nouvelles données discrètes. Blei (2012) identifie cinq hypothèses du modèle comme pistes d'améliorations explorées par les chercheurs. *L'hypothèse de sac de mots* : l'ordre de mots

n'est pas pris en compte. Des travaux visent notamment à intégrer des n-grammes dans le processus d'inférence (Wallach, 2006 ; Wang *et al.*, 2007). *L'hypothèse de l'interchangeabilité des documents*. La date de publication n'est pas prise en compte dans le modèle LDA. Pour modéliser l'évolution temporelle des thèmes, des approches dynamiques ont été développées (Blei et Lafferty, 2006 ; Wang *et al.*, 2012). *L'hypothèse d'un nombre fixe de thèmes*. Ce choix *a priori* est difficile à optimiser et peut changer au cours du temps. Teh *et al.* (2006) propose un modèle non paramétrique qui détermine le nombre de thèmes pendant le processus d'inférence et permet l'apparition de nouveaux thèmes lors de l'intégration de nouveaux documents. Les variations non paramétriques permettent de façon générale d'introduire une hiérarchie entre les thèmes, du plus général au plus spécifique par exemple (Blei *et al.*, 2007). *L'hypothèse de non corrélations entre les thèmes*. Dans les collections réelles, des thèmes peuvent montrer différents niveaux de corrélation (il se peut qu'un document traitant d'une problématique de *politique internationale* traite aussi de *guerre* mais peu probable qu'il traite de *chimie*). Deux modèles récents prennent en compte cette corrélation, au prix de modèles complexes en calculs (Blei et Lafferty, 2007 ; He *et al.*, 2017). *Aller au-delà du contenu textuel*. La plupart des documents textuels contiennent des métadonnées. Dans le cas d'articles scientifiques, on a souvent accès aux caractéristiques de l'auteur (nom, prénom, université…), de la revue de publication (nom, facteur d'impact…) ou encore de l'article lui-même (langue, longueur, nombre de citations…). Pour cela, Rosen-Zvi *et al.* (2004) ont développé le modèle auteur-thèmes. Ils ont attribué des proportions de thèmes à des auteurs d'articles scientifiques et l'ont mené l'inférence sur les auteurs en plus des documents. Ils construisent une mesure de similarité entre auteurs à partir des thèmes d'intérêt qu'ils partagent. Autre exemple, les documents peuvent être munis de liens entre eux comme les hyperliens des pages Web ou les citations des articles scientifiques. Chang et Blei (2010) ont construit un modèle de thèmes relationnel qui applique un modèle LDA à chaque document et établit une mesure de distance entre les documents à partir de leurs thèmes. Kim *et al.* (2017) incorporent les citations des auteurs d'un papier dans leur calcul d'inférence. Des travaux incorporent d'autres métadonnées : la structure linguistique ou une structure syntaxique (Boyd-Graber et Blei, 2009 ; Thomas L Griffiths *et al.*, 2005), la reconnaissance d'entités (Newman *et al.*, 2006) ou encore l'influence universitaire (Dietz *et al.*, 2007 ; Kataria *et al.*, 2011). Enfin, on peut appliquer le modèle LDA pour d'autres données de nature discrète que le texte : en génétiques par exemple pour trouver des ancêtres communs (Pritchard *et al.*, 2000) ou en reconnaissance d'images où l'on

considère chaque image comme un mélange de patrons (Fei-Fei et Perona, 2005 ; Li *et al.*, 2010). D'autres métadonnées ont été utilisées : la structure linguistique ou une structure syntaxique externe comme Wordnet (Boyd-Graber et Blei, 2009 ; Griffiths *et al.*, 2005), la reconnaissance d'entités (Newman *et al.*, 2006) ou encore l'influence académique (Dietz *et al.*, 2007 ; Kataria *et al.*, 2011). Enfin, on peut appliquer le modèle LDA pour d'autres données que le texte : en génétique pour trouver des ancêtres communs (Pritchard *et al.*, 2000) ou en reconnaissance d'images où l'on considère chaque image comme un mélange de patrons (Fei-Fei et Perona, 2005 ; Li *et al.*, 2010). Daud *et al.* (2010) proposent une revue exhaustive des variantes du LDA antérieures à 2010. Finalement, le modèle LDA a été utilisé en sciences sociales comme moyen de dégager de l'information de la langue elle-même : en sciences politiques (Grimmer, 2010), en psychologie (Socher *et al.*, 2009) ou encore en bibliométrie (Gerrish et Blei, 2010).

2.4 Synthèse

La revue de littérature nous a permis de cerner le potentiel d'application des modèles de thèmes en sciences humaines et sociales. La numérisation récente de nombreuses ressources textuelles a facilité l'accès à des connaissances étendues. Se pose maintenant la problématique d'organiser et de fouiller de façon efficace parmi une collection titanesque de documents spécialisés. L'approche d'apprentissage automatique et probabiliste offre des outils adéquats pour répondre à cette problématique et assister les utilisateurs disposant de telles collections.

Cependant, ces outils doivent encore être affinés et mieux compris pour les rendre plus accessibles et utilisables par des personnes non expertes en science des données. En outre, la plupart des travaux existant dans la littérature ne s'intéressent pas directement à l'application opérationnelle de ces algorithmes auprès de la communauté de chercheurs en SHS. Pourtant, il existe de multiples facteurs qui bloquent la mise en application des modèles probabilistes hors des sciences naturelles, dont les principaux sont la compréhension et l'interprétation des résultats et la validité et fiabilité des modèles (Ramage *et al.*, 2009).

La motivation essentielle de ce mémoire est donc d'évaluer le modèle LDA sur un corpus de textes en sciences sociales pour en faciliter l'adoption auprès de chercheurs en SHS. Pour cela, on souhaite *décrire* intuitivement les résultats obtenus à l'aide de visualisations claires et *évaluer* la

fiabilité et la capacité des modèles de thèmes à être opérationnalisés sur une tâche de fouille de documents. Le chapitre suivant expose la méthodologie développée pour répondre à cette motivation.

CHAPITRE 3 MÉTHODOLOGIE ET DONNÉES

Ce chapitre consiste à définir les objectifs de cette recherche, décrire les données sélectionnées pour y répondre et les traitements préliminaires qui leur ont été appliqués en vue de leur modélisation thématique.

3.1 Questions de recherche

L'objectif de ce mémoire est de participer à l'accessibilité et l'adoption de modèles de thèmes au sein de la communauté de chercheurs en SHS, afin de les aider à valoriser leurs collections numériques de textes. Pour cela, nous souhaitons employer ce modèle sur une tâche de recherche de document.

Il est d'abord nécessaire de mener une première analyse descriptive des résultats produits par un modèle LDA appliqué sur un corpus de sciences humaines.

Question 1 : Comment *décrire* et *évaluer* thématiquement un corpus d'articles en sciences sociales à l'aide du modèle LDA ?

Cette première question vise à déterminer la meilleure configuration méthodologique pour implémenter et entraîner des modèles LDA sur la collection étudiée. Il s'agit d'abord de visualiser les thèmes obtenus, puis de de mesurer la performance du modèle selon les métriques d'évaluation usuelles dans la littérature des modèles de thèmes : la perplexité et la cohérence. C'est le sujet du chapitre 4.

Une fois ces premières analyses effectuées, on s'intéresse à l'intégration du modèle LDA sur une tâche de recherche de documents. Le but est d'évaluer l'opérationnalisation de modèle LDA sur un besoin réel.

Question 2 : Comment *intégrer le LDA* sur une tâche de recherche d'articles en sciences sociales ?

Cette seconde question concerne l'apport méthodologique le plus important du mémoire. On s'intéresse à mesurer la performance de différents algorithmes basés sur le LDA pour rechercher des documents au sein d'un corpus en sciences sociales. Le chapitre 5 traitera de cette question.

Nous décrivons maintenant la méthodologie adoptée pour répondre à ces deux questions de recherche.

3.2 Méthodologie

Le travail vise à tester trois hypothèses de recherche.

Hypothèse 1 : « Le modèle LDA permet de décrire de façon valide et fiable des articles en sciences humaines et sociales. »

Pour tester cette première hypothèse, il s'agira d'appliquer des modèles LDA sur un corpus d'articles complets, dans le domaine des SHS et publiés en français. Une première étape de visualisation sera menée pour appréhender les résultats obtenus. Puis, nous caractériserons l'effet du nombre de thèmes du modèle LDA et celui de la lemmatisation du vocabulaire à l'aide des mesures de perplexité et de cohérence. Ces mesures seront enfin utilisées pour caractériser la fiabilité des modèles LDA.

Hypothèse 2 : « Les thèmes extraits par un modèle LDA appliqué sur un corpus de textes en français en sciences humaines et sociales peuvent être utilisés pour faire une expansion de requête sur une tâche de recherche de documents. »

Hypothèse 3 : «Augmenter la requête à l'aide de la représentation thématique des documents donne de meilleurs résultats qu'utiliser des mots ayant une forte cooccurrence avec ceux de la requête. »

Ces deux dernière hypothèses seront testées via une tâche de recherche de documents, dont on proposera la méthodologie de construction et d'évaluation. Il s'agira en particulier de comparer la performance de deux algorithmes d'expansion de requête basés sur le modèle LDA, ainsi que l'influence de différents paramètres sur la pertinence des résultats (nombre de thèmes *a priori*, domaine de la revue et spécificité de la requête).

La Figure 3.1 résume les contributions de ce mémoire (marquées en orange) par l'intermédiaire d'un schéma rappelant les différentes étapes de l'analyse de données.

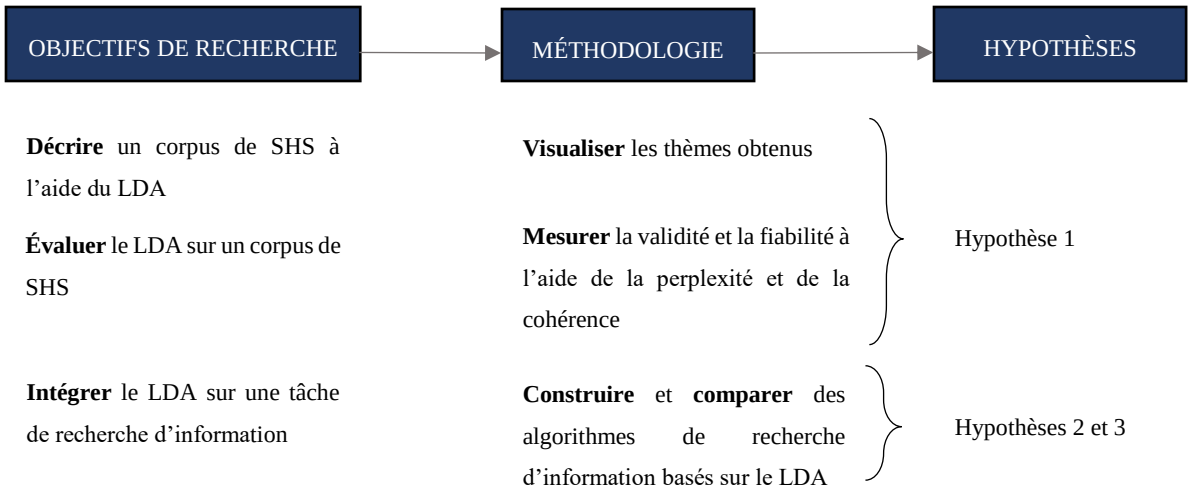

Figure 3.1 Vue générale de la recherche

3.3 Choix des données : étude de cas

Pour répondre aux questions de recherche, il nous faut disposer d'un corpus de textes en sciences sociales sur lequel on pourra entraîner et évaluer des modèles LDA. Pour cette étude, le choix des données s'est porté sur la collection d'articles proposée par la plateforme Érudit[1]. Érudit est une plateforme québécoise qui fournit un accès centralisé à des documents de recherche dans plus de trente disciplines de SHS : des articles en texte complet de revues savantes et culturelles, des livres, des mémoires et thèses et des données de recherche. Érudit est le plus important diffuseur de ressources francophones en SHS d'Amérique du Nord avec plus de 150 revues et 200 000 documents accessibles. Nous avons sélectionné trois revues québécoises accessibles sur Érudit pour former notre ensemble de données : Actualité Économique (AE), Études Internationales (EI) et Relations Industrielles (RI). Ce choix est motivé par deux raisons :

1) Ces revues traitent de thématiques de recherche différentes. On pourra entraîner trois modèles LDA et comparer la performance de chacun en fonction du domaine de recherche.
2) Ces revues ont une histoire de publication suffisamment longue pour disposer de plus de 1000 articles à analyser chacune. La longueur des articles (mesurée à l'aide du nombre de

[1] https://www.erudit.org/fr/

jetons total dans un article après filtre du vocabulaire) suit une distribution de probabilité normale.

Les revues Études Internationales et Relations Industrielles ont été créées à l'Université de Laval tandis qu'Actualité Économique a été fondée au HEC Montréal. Chaque revue publie 4 numéros par an et une vingtaine d'articles. Remarquons dès maintenant que notre base de données est petite (4000 articles). Elle présente un risque de surapprentissage si on lui applique des modèles trop complexes. Le Tableau 3.1 résume les caractéristiques importantes des trois revues.

Tableau 3.1 Caractéristiques des trois revues sélectionnées

	Actualité Économique	**Études Internationales**	**Relations Industrielles**
Nombre d'articles en français	1534	1094	1372
Nombre moyen de jetons par article	2872	3911	2916
Période de publication Total (années)	1955 – 2017 62 ans	1970 – 2017 47 ans	1945 – 2018 73 ans
Thématiques	Science économique et finance	Relations internationales Science politique et économique	Etude du travail et de l'emploi

Le travail effectué dans la suite de ce mémoire pourrait naturellement être appliqué sur d'autres documents en sciences sociales pour en évaluer la capacité de généralisation et effectuer une comparaison entre corpora.

3.4 Description des données et traitements préliminaires

Une partie des articles a été numérisée par les équipes d'Érudit tandis que les articles les plus récents ont été publiés sous une forme numérique d'origine. Les équipes d'Érudit proposent aussi une version XML annotée de chacun des articles en plus de leur version PDF. La version XML des articles a été obtenue en appliquant un système de reconnaissance optique de caractère (Mori *et al.*, 1999). Certaines erreurs de transcription ont eu lieu durant cette conversion, sur lesquelles nous reviendrons dans la section 3.4. Le format XML donne un accès direct au contenu textuel des

articles et est muni d'une organisation structurelle précisée par Érudit[2]. On retiendra en particulier l'attribut « typeart » qui qualifie le genre de chaque publication : *article*, *compte-rendu*, *note* ou *autre*. Nous avons retenu seulement les publications de type *article*, qui sont suffisamment longues et sont construites autour de quelques thématiques. Selon les articles, le traitement XML peut également varier : le traitement est dit « minimal » quand un humain a identifié le corps de l'article et l'a séparé des métadonnées et il est dit « complet » quand un humain a procédé à une identification sémantique approfondie de la publication (par exemple, séparation par section au sein des paragraphes de texte et identification des citations). La Figure 3.2 présente la répartition du type de traitement XML pour les trois revues. La plupart des articles des trois revues concernées ont subi un traitement « minimal », donc les étiquettes supplémentaires attribuées dans le cas d'un traitement « complet » n'ont pas été utilisées. Elles pourraient être utiles pour des études complémentaires, par exemple sur une tâche d'extraction de citations.

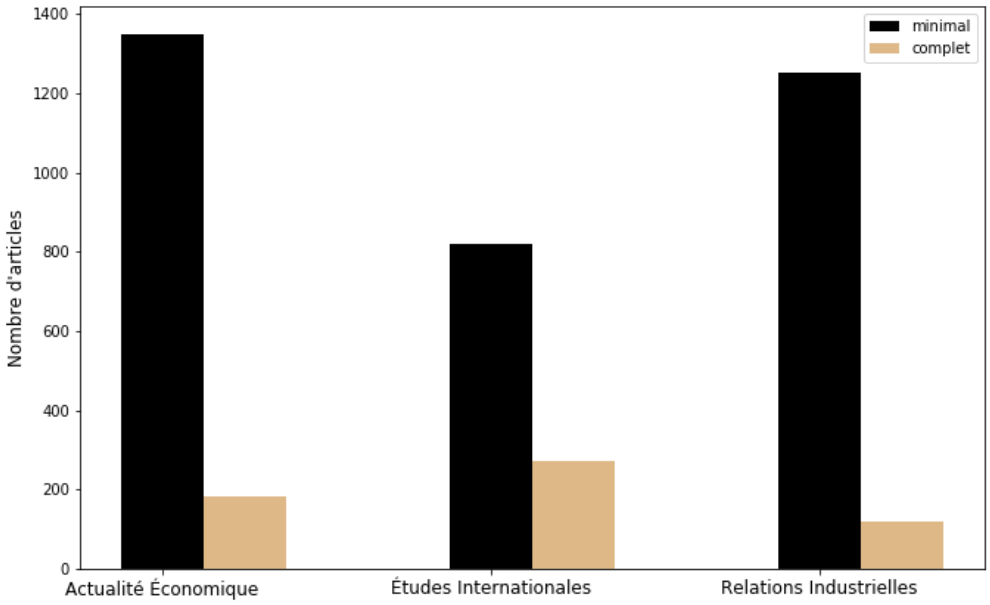

Figure 3.2 Répartition du type de traitement XML des articles

[2] http://retro.erudit.org/xsd/article/3.0.0/doc/

Après avoir caractérisé qualitativement les données utilisées dans notre travail, nous présentons maintenant les traitements préliminaires que nous leur avons appliqués en amont des modèles de thèmes.

Nous commençons par souligner les erreurs inhérentes aux données manipulées. Une première source d'erreur est la conversion automatique effectuée pour passer d'un format PDF au format XML. Le processus de reconnaissance optique de caractères (ROC) engendre un certain nombre d'erreurs connues qui nuisent aux analyses qui succèdent à cette première étape (Lopresti, 2009). En premier lieu, l'algorithme ROC n'est pas capable d'identifier sur une page PDF donnée la fin d'un article et le début du suivant. À l'aide d'un algorithme de comparaison de séquences, nous avons quantifié ce problème au sein de chaque revue. Seule la revue Relations Industrielles est concernée avec 2,8 % des articles dont une partie est convertie deux fois en XML par le ROC. Il a donc été décidé d'ignorer ce problème. Par ailleurs, les mots découpés en fin de ligne par un tiret sont considérés comme deux mots distincts dans la conversion XML : par exemple, le mot « auto-mobile » devient « auto » et « mobile ». De plus, la séquence « li » est lue comme un « U » par l'algorithme OCR. Enfin, des textes en français peuvent contenir des passages en anglais comme une version anglaise du résumé (*abstract*) ou des termes techniques non traduits. Ces problèmes ont été abordés à l'aide d'un algorithme de correcteur orthographique. Le mécanisme de c*et al*gorithme sera détaillé plus avant, lors de la constitution du vocabulaire pour chaque revue.

Le passage des articles en version PDF à la version XML est effectué par l'équipe d'Érudit. Nous avons travaillé sur la transformation du texte issu des fichiers XML en données utilisables pour les modèles de thème. La procédure de transformation suit les recommandations de la littérature exposées en section 2.2.1. Le langage de programmation Python a été utilisé pour l'ensemble du travail, car il offre un ensemble de bibliothèques d'analyse de textes efficaces et faciles d'accès (Van Rossum et Drake, 2011).

1) *Tokeniser*. Nous avons utilisé le tokeniseur de la bibliothèque NLTK[3]. (Loper et Bird, 2002). Nous avons ajouté les bigrammes les plus fréquents (apparaissant dans plus de 20 documents) : ils forment environ 4 % du vocabulaire final pour chaque revue (voir Tableau

[3] https://www.nltk.org/

3.2). De plus, nous avons retiré les articles trop courts (moins de 300 jetons) car ils sont susceptibles de bruiter les résultats du modèle de thèmes.

2) *Enlever les mots-arrêts*. 672 mots-arrêts pour la langue française ont été retirés. Ces mots-arrêts ont été extraits d'une liste de mots proposée par les bibliothèques NLTK et Spacy[4].

3) *Enlever les chiffres, la ponctuation et mettre en minuscule.* Nous avons utilisé des fonctions intégrées au langage de programmation Python.

4) *Appliquer un correcteur orthographique.* Nous avons utilisé une distance de Levenshtein qui calcule pour chaque mot du vocabulaire l'ensemble des mots à distance 1 (suppression, insertion ou substitution d'une lettre du mot). On compare ensuite cet ensemble de mots à un dictionnaire connu de la langue française, qui fait le lien entre chaque mot et sa fréquence générale dans la langue. Dans le cas d'une erreur d'orthographe, on retourne le mot de distance 1 avec la plus grande fréquence dans la langue française. Si aucun mot ne correspond, on l'élimine. Entre 45 et 70 % des jetons (unigrammes) ont été corrigés (voir Tableau 3.2).

5) *Lemmatiser les jetons.* Nous avons utilisé un lemmatiseur développé pour le français par Claude Coulombe[5], basé sur le Lexique des Formes Fléchies du Français (LEFF). Il retourne la forme infinitive des verbes et la forme masculine singulière des autres mots.

Enfin, la littérature recommande de limiter la taille du vocabulaire pour accélérer l'entraînement des modèles et améliorer l'interprétabilité des résultats (Blei et Lafferty, 2009 ; Lu *et al.*, 2017). Pour limiter le nombre de jetons conservés pour chaque revue, nous avons appliqué différents filtres sur la fréquence des mots. Nous avons mesuré l'influence de la force du filtre sur la taille du vocabulaire (i.e. la diversité lexicale) pour chaque revue. La Figure 3.3 montre l'effet des filtres de mots rares et de mots fréquents sur le vocabulaire non lemmatisé des trois revues. La ligne rouge indique la taille de vocabulaire retenue. La tendance est similaire sur le vocabulaire lemmatisé. La colonne de droite concerne le nettoyage des mots fréquents : le filtre enlève les mots qui apparaissent dans plus d'un certain pourcentage de documents. Ce filtre peut être vu comme une

[4] https://spacy.io/

[5] https://github.com/ClaudeCoulombe/FrenchLefffLemmatizer

généralisation des mots-arrêts. On choisit une valeur de filtre égale à 30 % des documents, ce qui permet de limiter la taille du vocabulaire de 5 % : tous les mots qui apparaissent dans plus de 30 % des documents sont retirés. Pour les mots rares, le filtre enlève les mots qui apparaissent dans moins d'un certain nombre de documents. On choisit une valeur de filtre égale à 3 documents, ce qui permet de limiter la taille du vocabulaire entre 55 % et 60 % : tous les mots qui apparaissent dans moins de 3 documents sont retirés du vocabulaire.

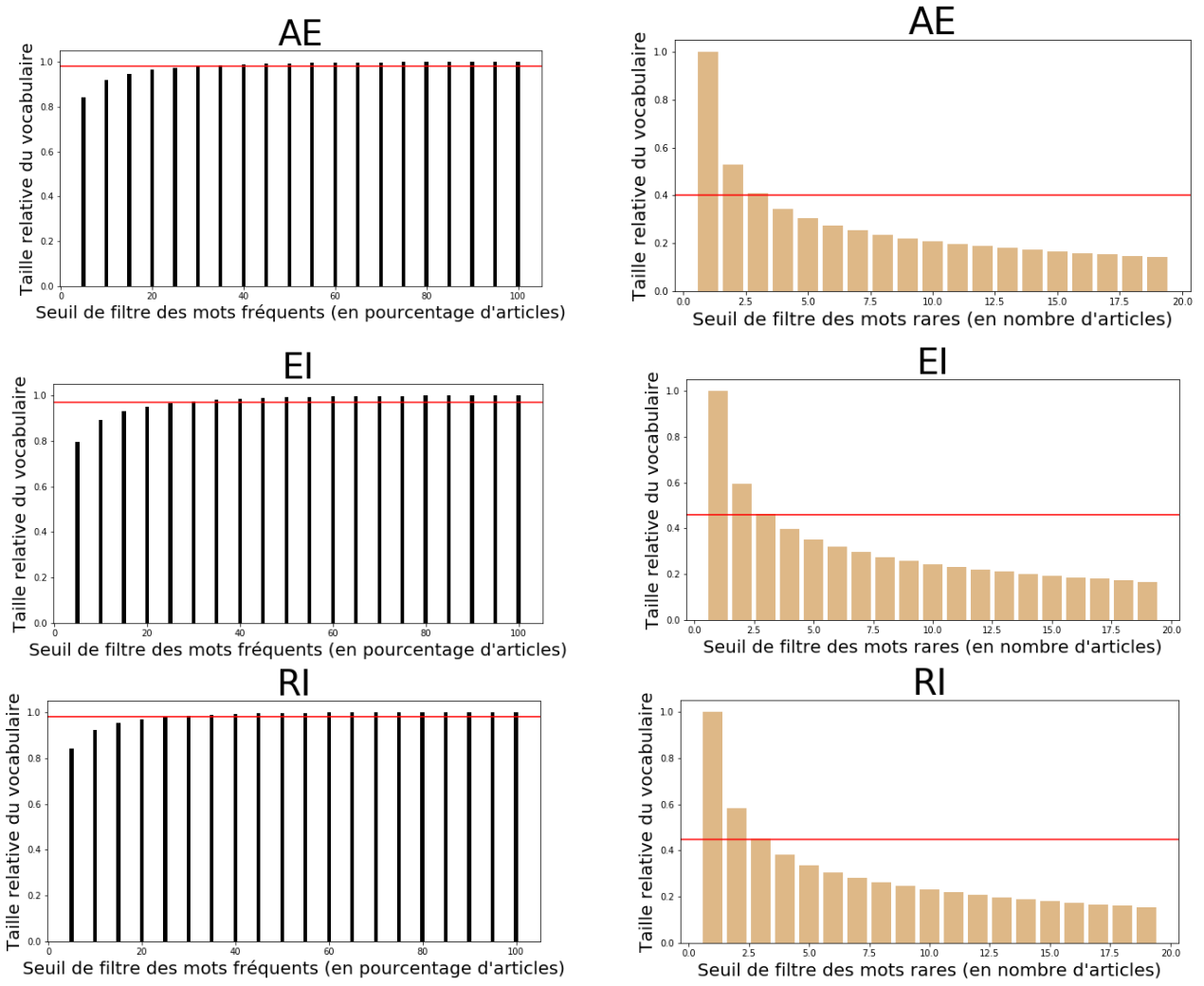

Figure 3.3 Effet des filtres de mots rares et fréquents sur la taille du vocabulaire

On résume les opérations effectuées sur le vocabulaire dans le Tableau 3.2. La ligne « % de jetons corrigés » correspond au pourcentage de jetons corrigés par le correcteur orthographique. Les vocabulaires avec lemmatisation sont logiquement plus réduits. Le pourcentage de bigrammes est stable, autour de 4 %.

Tableau 3.2 Propriétés du vocabulaire pour chaque revue

		Actualité Économique	Études Internationales	Relations Industrielles
SANS Lemmatisation	Taille du vocabulaire avant filtres	112 028	113 672	104 139
	% de jetons corrigés	44	65	65.9
	% de bigrammes dans le vocabulaire	4.1	4.2	4.4
	Taille du vocabulaire final	36 100	41 585	39 205
AVEC Lemmatisation	Taille du vocabulaire avant filtres	104 341	105 048	97 333
	% de jetons corrigés	70.7	67.7	68.1
	% de bigrammes dans le vocabulaire	3.6	3.8	4.4
	Taille du vocabulaire final	32 374	37 343	36 169

Finalement, la Figure 3.4 présente le pipeline d'analyse. Il est important de noter qu'on analyse chaque revue *séparément*. Les articles traités sont convertis d'un format PDF en un format XML par l'équipe d'Érudit. Puis, nous extrayons le contenu XML et préparons le texte pour le transformer en représentation sac de mots. La matrice sac de mots est utilisée comme entrée des modèles de thèmes, dont les résultats sont visualisés et évalués selon différentes métriques.

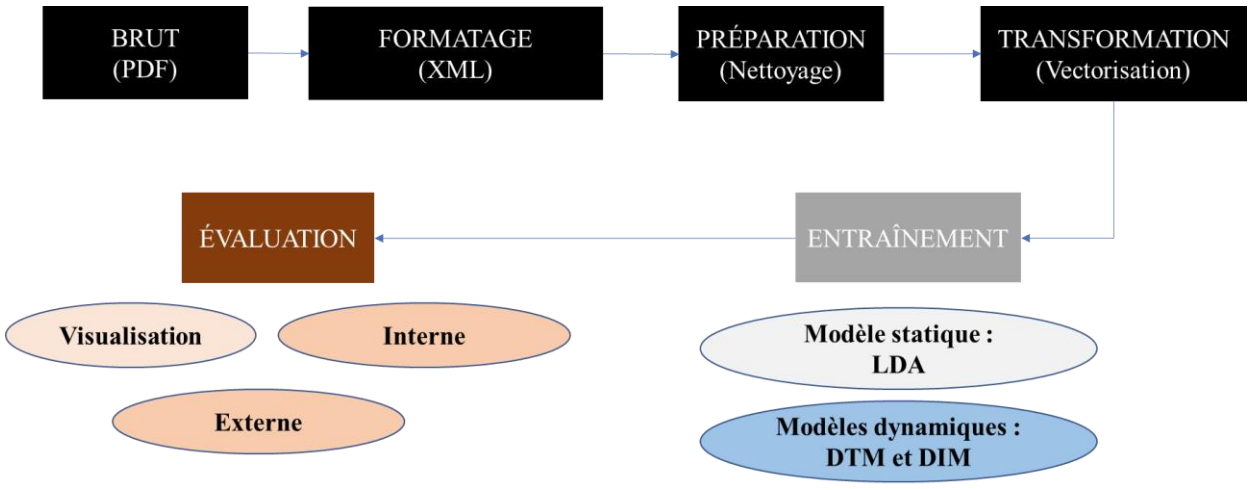

Figure 3.4 Pipeline d'analyse des données

3.5 Synthèse

Ce chapitre a détaillé les manques de la littérature que ce mémoire souhaite combler. Les objectifs, questions de recherche et hypothèses de travail ont été exposés. Puis les données utilisées pour répondre à la problématique ont été décrites ainsi que les traitements préliminaires aux analyses.

Le prochain chapitre détaille l'application du modèle LDA aux trois revues de SHS et les méthodes de visualisation et d'évaluation des résultats pour répondre à la première question de recherche. Le cinquième chapitre s'intéressera à l'intégration du LDA dans une tâche de fouille de documents pour répondre à la seconde question de recherche.

CHAPITRE 4 VISUALISATION ET ÉVALUATION AUTOMATIQUE DE MODÈLES LDA

Ce chapitre traite des moyens de visualisation des résultats produits par le modèle LDA (Section 4.2.1) ainsi que des mesures d'évaluation automatique permettant de comparer la validité et fiabilité des modèles sur le corpus d'étude (Section 4.2.2).

4.1 Introduction

La bibliothèque Python Gensim[6] développée depuis 2010 et dédiée aux outils de modélisation de thèmes a été utilisée pour implémenter le modèle LDA (Rehurek et Sojka, 2010). La méthode d'inférence variationnelle utilisée (*Variational Bayes*) est décrite dans Hoffman et Blei (2010). L'intérêt de cette méthode est qu'elle exige un espace mémoire constant contrairement aux méthodes d'échantillonnage (telle que celui de Gibbs) qui requièrent un espace mémoire linéairement croissant avec le nombre de documents.

Pour le choix des deux *a priori* du modèle que sont α et η, on se base sur les recommandations de Wallach *et al.* (2009). Ainsi, un *a priori* asymétrique a été choisi pour l'hyperparamètre α et un *a priori* symétrique a été choisi pour η. Rappelons que α contrôle la distribution de probabilité des thèmes au sein de chaque document et que η contrôle les proportions initiales des mots dans chaque thème.

Pour chaque revue, différents modèles LDA ont été entraînés, où l'on a fait varier le nombre de thèmes et où l'on applique ou non une étape de lemmatisation. L'entraînement d'un modèle LDA prend typiquement 30 min sur un cœur Intel® i5. Le Tableau 4.1 résume les hyperparamètres choisis pour l'entraînement des modèles LDA.

[6] https://radimrehurek.com/gensim/

Tableau 4.1 Hyperparamètres des modèles LDA implémentés

Hyperparamètre	Plage de valeur
Nombre de thèmes	[1–10,20,30,40,50,60,70,80,90,100]
Nombre d'itérations de l'algorithme d'inférence	400
Type de représentation du corpus	Sac de mots
Lemmatisation du vocabulaire	Oui / Non

Chaque revue a été séparée en deux ensembles : un ensemble d'entraînement regroupant 80 % des articles et un ensemble de test contenant 20 % des articles. Cette séparation a pour but d'entraîner les modèles LDA sur l'ensemble d'entraînement et d'évaluer leur perplexité sur l'ensemble de test.

4.2 Résultats

La section 4.2.1 présente des outils de visualisation des résultats pour *décrire* thématiquement chacune des revues. La section 4.2.2 consiste à *évaluer* quantitativement les résultats obtenus.

4.2.1 Visualisation

On présente d'abord quelques méthodes de visualisation adaptées aux besoins de chercheurs non experts en modèles de thèmes. Disposer de méthodes de visualisation claires et exploitables est essentiel pour favoriser l'adoption des modèles de thèmes dans la communauté des chercheurs en sciences sociales et humaines. Ramage *et al.* (2009) ont ainsi identifié la réticence qu'ont les chercheurs en sciences sociales à faire *confiance* aux résultats des modèles de thèmes à cause de la difficulté rencontrée pour caractériser les résultats obtenus. En effet, les modèles LDA produisent des distributions de probabilités de mots au sein de thèmes, et de thèmes au sein de documents. Visualiser cette information est difficile à cause de la haute dimensionnalité des données manipulées : des dizaines ou centaines de thèmes, des milliers de documents et des dizaines de milliers de mots. L'approche récente de visualisation propose des interfaces compactes et interactives. Deux échelles de visualisation sont proposées : l'échelle du corpus et l'échelle du document. Ces deux échelles de visualisation sont illustrées sur les modèles LDA à 10 thèmes, sans lemmatisation et avec une représentation sac de mots du corpus. Les résultats des modèles LDA avec 10 et 20 thèmes pour chaque revue sont présentés à l'annexe A. Pour les résultats des modèles

avec un plus grand nombre de thèmes, on renvoie le lecteur au fichier Excel « *Résultats LDA — jetons.xlsx* » stocké sur le dépôt GitHub[7] associé au projet.

4.2.1.1 Visualisation à l'échelle du corpus

Une première approche pour visualiser les résultats produits par le modèle est de représenter chaque thème par un nuage de mots, où chaque mot aura une police proportionnelle à son poids dans le thème. La Figure 4.1 présente la visualisation nuages de mots pour les modèles LDA à 10 thèmes. Elle est utile pour des modèles avec peu de thèmes, permet d'identifier d'éventuels thèmes incohérents et donne une image générale des résultats (Viégas et Wattenberg, 2008).

La figure montre la présence de thématiques visiblement bien définies pour les trois revues, avec certains thèmes qui apparaissent toutefois moins cohérents, présentant un mélange de thématiques. C'est le cas du thème 6 de la revue AE avec des termes liés à des mesures économiques comme « fonction de coût » ou « coût marginal » et des termes génériques comme « firme ». Le thème 0 de AE apparaît comme le moins cohérent et il contient en fait des termes qui étaient présents en bas de page des articles et non filtrés lors du traitement préliminaire. On retrouve aussi un mélange dans le thème 8 de la revue EI qui contient des termes liés à la pêche et d'autres moins définis comme « japonais » ou « soviétiques ». Par ailleurs, les thèmes 7 des revues AE et EI et le thème 8 de la revue RI contiennent des termes liés à la méthodologie générale de ces recherches (avec les termes « variables », « résultats », « théorie », « approche », « performance »). Les thèmes 0 et 9 de la revue RI montrent que le modèle est capable d'extraire des thèmes en langue anglaise. Ces termes viennent essentiellement des *abstracts* présents au début des articles et de termes techniques non traduits. Un détecteur de langue n'a en effet été appliqué que sur l'ensemble du contenu des articles et non sur chaque phrase pour une raison de précision. En outre, cette représentation permet d'identifier les mots récurrents dans plusieurs thèmes qui « écrasent » les autres par leur poids. C'est le cas des bigrammes « taux_chômage », « marché_travail » et « millions_dollars » pour la revue AE. Pour les revues EI et RI, aucun terme ne ressort visiblement dans différents thèmes. Ces mots dominants indiquent des concepts ou méthodes fréquemment discutés dans la revue et servant de thèmes généraux communs à différentes problématiques d'étude. Enfin, on remarque un effet

[7] https://github.com/arthurlemon/topic-models-SHS

de redondance thématique entre les thèmes 1 et 2 (traitant de travail) de la revue AE d'une part et les thèmes 1 et 2 (traitant de chômage et d'emploi) et les thèmes 5 et 6 (traitant de justice) de la revue RI. Par ailleurs, il est à noter que le modèle LDA ne fournit pas d'étiquette pour nommer chaque thème extrait. Cette étape est traditionnellement réalisée par des humains, experts ou non sur les thématiques du corpus. Sur l'exemple, certaines thématiques extraites par le modèle peuvent être étiquetées en première approximation.

L'analyse de la représentation en nuages de mots a été menée sur des modèles à 10 thèmes pour en illustrer le principe. Elle permet de retrouver les thématiques de recherche générale de chaque revue : économie pour AE, relations internationales pour EI, et étude du travail pour RI. Elle pourrait être effectuée par des chercheurs en SHS suivant la même logique sur des modèles avec un plus grand nombre de thèmes afin d'obtenir une représentation thématique plus fine. Toutefois, l'inconvénient de cette représentation en nuage de mots est qu'elle réduit l'information extraite par les modèles à des groupes de mots, sans donner de précision sur le contexte d'utilisation des termes ni de moyen de comparaison entre les thèmes.

Pour approfondir l'analyse, nous avons effectué une première analyse qualitative de l'influence des trois hyperparamètres du modèle LDA à partir des tableaux de résultats présentés à l'annexe A : le nombre de thèmes du modèle et la lemmatisation (ou non) du vocabulaire.

L'effet du nombre de thèmes est qualitativement difficile à estimer. En effet, comment évaluer la cohérence sémantique d'un ensemble de mots sans expertise du domaine et à partir de l'observation de groupes de termes ? Bien que certains thèmes apparaissent *visiblement* incohérents (Tableau A.5 Thèmes extraits pour la revue RI avec un modèle LDA à 10 thèmes : par exemple, le thème 7 du modèle LDA à 10 thèmes avec lemmatisation appliqué sur la revue RI qui rassemble les termes « allait — propriété — individualisées — rapport salarial — engendré »), la plupart sont difficilement comparables. Par exemple, pour les modèles à 10 et 20 thèmes appliqués à la revue EI sans lemmatisation (Tableau A et Tableau A.4), quel thème a la meilleure cohérence visuelle entre « identité — violence — terrorisme — otan — irak » (modèle à 10 thèmes) et « communiste — chinois — partis — communistes — socialiste » (modèle à 20 thèmes) ? C'est pourquoi des méthodes d'évaluation quantitatives sont explorées à la section 4.2.2.

Nous nous sommes également interrogés sur les poids que le modèle attribue aux mots prépondérants de chaque thème afin de vérifier . En effet, en cas l'incohérence visuelle de certains

thèmes pourrait s'expliquer par une non convergence du modèle LDA. Dans ce cas, l'algorithme attribue un poids faible à tous les termes du vocabulaire et ne dégage pas de termes réellement « représentatifs » pour chaque thème.

Figure 4.1 Représentation des thèmes de 3 modèles LDA à 10 thèmes, sous forme de nuage de mots

Pour vérifier le comportement de l'algorithme, nous avons tracé en Figure 4.2 l'évolution du poids moyen attribué aux mots les plus représentatifs de tous les thèmes en fonction du nombre de thèmes du modèle LDA. Le poids moyen est défini mathématiquement à l'équation :

$$p_m(K) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{m} \sum_{w \in top_m(k)} p(w|k)$$

Équation 4.1 Poids moyen attribué au mots les plus représentatifs des K thèmes d'un modèle LDA

Où

- $top_m(k) =$ Ensemble des m mots les plus représentatifs du thème k.
- K = nombre de thèmes *a priori* du modèle LDA

Pour donner un exemple de lecture de la figure, pour le modèle LDA à 30 thèmes avec une représentation sac de mots du corpus et sans lemmatisation du vocabulaire, un poids moyen de 0,01 a été calculé. Cette valeur s'interprète comme le poids que le modèle attribue en moyenne aux 10 termes les plus représentatifs de chacun des 30 thèmes qu'il a extraits. La valeur de ce poids moyen est un indicateur de la représentativité des termes composant les thèmes extraits par un modèle LDA. Les graphes montrent une tendance similaire pour les 3 revues. Pour les modèles utilisant une représentation sac de mots du corpus, le poids moyen augmente avec le nombre de thèmes de façon logarithmique. Ces modèles montrent donc une bonne capacité à sélectionner des termes représentatifs pour chaque thème. Par ailleurs, aucun effet significatif de la lemmatisation du vocabulaire n'est constaté, les tendances et valeurs étant similaires à celles obtenus sans l'étape de lemmatisation. On propose d'autre part de s'intéresser à la variabilité des poids attribués par le modèle aux différents thèmes extraits. L'Annexe B présente les résultats des boîtes à moustache obtenues pour chaque revue et chaque configuration de modèle. La représentation en « boîte à moustache », aussi appelée diagramme en boite ou boîte de Tukey, permet de comparer un même caractère entre des populations de tailles différentes. Le caractère comparé ici est le poids moyen attribué par chaque modèle ayant un nombre de thèmes *a priori* différent. La variabilité est faible pour les modèles avec un petit nombre de thèmes (K ≤ 10) où la taille de l'échantillon est petite. Elle devient plus grande pour les modèles avec un nombre de thèmes plus important, mais très peu de valeurs aberrantes sont observées. Cependant, il y a peu de points extrêmes (moins de 5 même

pour les modèles à 100 thèmes), ce qui montre que tous les thèmes extraits par le modèle LDA comportent des termes représentatifs.

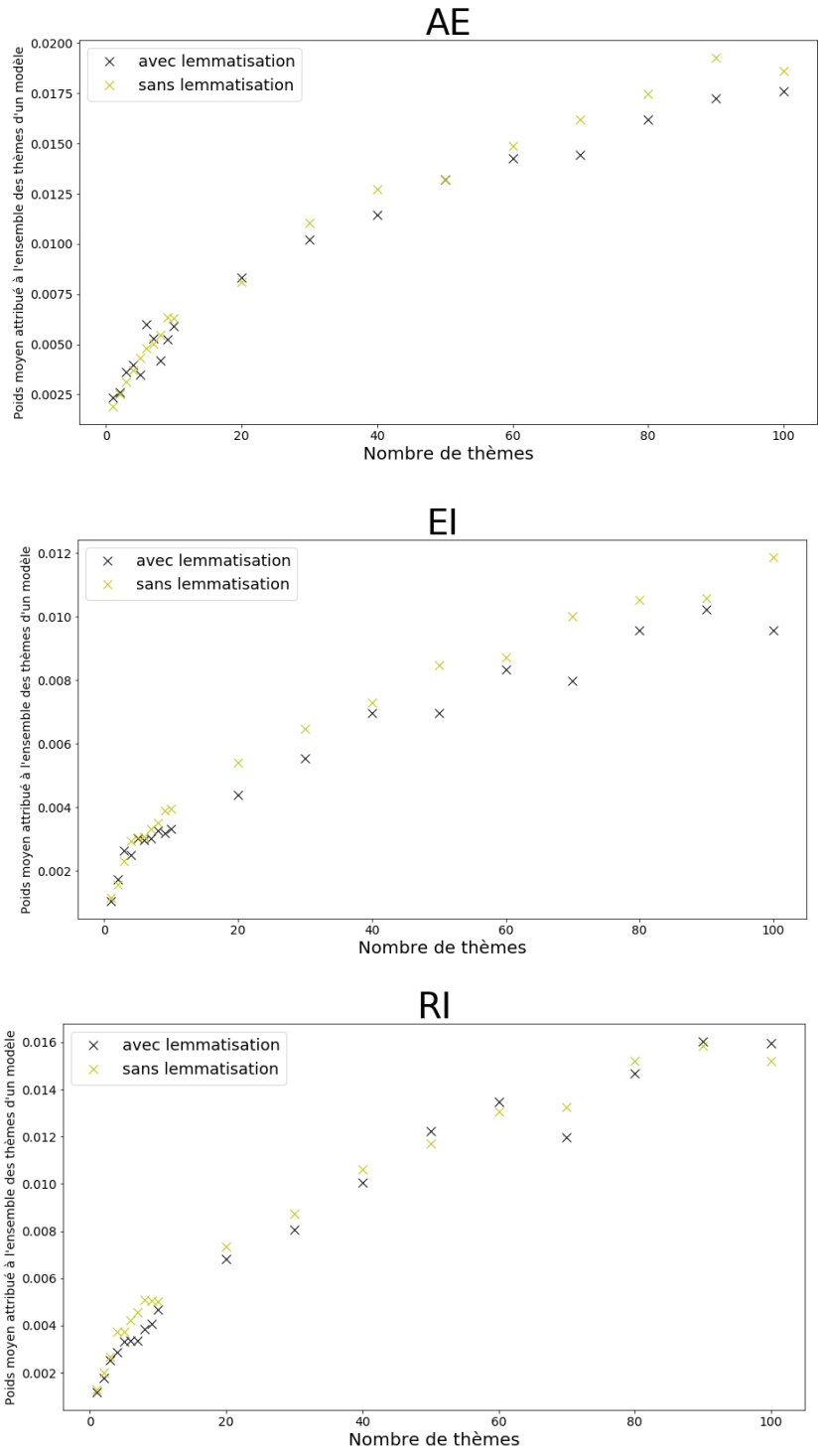

Figure 4.2 Poids moyens attribués aux mots prépondérants des thèmes par différents modèles LDA

En conclusion, le modèle LDA répartit de façon relativement homogène les poids des mots-clés sur les différents thèmes : les termes représentatifs de chaque thème ont un poids similaire et aucun thème ne domine les autres en probabilité. Aucun effet de lemmatisation n'est observé sur la distribution de probabilité des poids.

Comment aller au-delà de la simple distribution de probabilité des mots par thème et des thèmes par document ? La bibliothèque PyLDAvis (Sievert et Shirley, 2014) a été développée comme moyen de représenter le sens de chaque thème, leur importance relative dans le corpus et les similarités entre les thèmes. Elle implémente deux mesures pouvant être ajustées de façon interactive : la saillance et la pertinence.

La mesure de saillance a été introduite par Chuang *et al.* (2012) pour distinguer les termes qui discriminent le mieux un thème *par rapport* aux autres thèmes. Pour un mot *w*, les auteurs définissent sa « spécificité » comme la divergence Kullback-Leibler entre *P(T|w)*, la probabilité de tirer le thème *T* étant donné le terme *w* et *P(T)*, la probabilité marginale du thème T. Autrement dit, si *w* apparaît dans la plupart des thèmes, sa spécificité sera faible : il apporte peu d'information sur la génération du thème *T*. La saillance du mot *w* est alors définie comme le produit entre sa spécificité et sa distribution de probabilité marginale *P(w)*, qui représente le poids du mot dans l'ensemble du corpus.

$$saillance(w) = P(w) \times spécificité = P(w) \times \left(\sum_{T} P(T|w) log \frac{P(T|w)}{P(T)} \right)$$

Équation 4.2 Définition de la saillance du terme *w*

La mesure de pertinence est propre à la bibliothèque PyLDAvis. Elle sert de critère de hiérarchisation des termes *au sein* d'un thème. On note *P(w|T)* la probabilité de tirer le terme *w* étant donné le thème *T* et *P(w)* la probabilité marginale de *w*. La pertinence du mot *w* relativement au thème *T* est définie par l'équation suivante :

$$pertinence(w, T|\lambda) = \lambda log\ P(\mathrm{w}|\mathrm{T}) + (1 - \lambda) log \left(\frac{P(\mathrm{w}|\mathrm{T})}{P(w)} \right)$$

Équation 4.3 Définition de la pertinence du terme *w*

Le facteur λ est un hyperparamètre pouvant varier entre 0 et 1 et que l'utilisateur peut moduler pour ajuster la liste de termes les plus représentatifs d'un thème. Le choix λ = 1 ordonne les mots selon

leur distribution de probabilité au sein du thème, ce qui revient à la méthode classique de hiérarchisation. Le choix $\lambda = 0$ ordonne les mots selon leur *lift*, c'est-à-dire le ratio entre la distribution de probabilité du mot au sein du thème et la distribution de probabilité du mot au sein du corpus. Plus le mot *w* est rare dans le corpus, plus son *lift* sera important. Les auteurs évaluent différentes valeurs de λ à partir d'un modèle de 50 thèmes entraîné sur 13 700 documents du corpus Newsgroups. Il a été demandé à 29 sujets de trouver le mot intrus dans une liste de 5 termes issus d'un thème du modèle LDA. Une valeur optimale de λ égale à 0.6 a été trouvée, que nous avons utilisée dans nos visualisations. Il faut toutefois noter que cette valeur est dépendante du corpus d'étude et du nombre de thèmes du modèle. Elle devrait idéalement être réévaluée pour chaque nouvelle expérience.

L'interface PyLDAvis se décompose en deux parties permettant à l'utilisateur d'appréhender ces deux mesures. La partie gauche de l'interface représente la projection 2D de chaque thème sous forme de cercle. Les coordonnées 2D du centre du cercle sont calculées en fonction de la similarité inter thèmes, mesurée via une mesure cosinus entre les distributions de mots composant chacun des thèmes. L'aire indique l'importance relative de ce thème dans le corpus mesurée avec les poids attribués à chaque mot du vocabulaire par le modèle. La partie droite de l'interface représente la mesure de pertinence des termes composant le thème sélectionné sur la partie gauche, pouvant être ajustée par l'utilisateur à l'aide d'un curseur gérant le facteur λ. La Figure 4.3 compare deux méthodes de projections 2D (ACP et t-SNE) via l'interface PyLDAvis sur la revue Actualité Économique, avec un modèle LDA à 30 thèmes. Naturellement, les mots les plus représentatifs extraits par le modèle sont les mêmes entre les deux représentations. La différence repose sur la visualisation des thèmes composant le corpus. La méthode t-SNE permet ici de mieux séparer les thèmes sur une projection 2D que la méthode ACP.

Deux caractéristiques de la projection proposée par PyLDAvis peuvent servir à évaluer la qualité d'un modèle LDA :

- On s'attend à ce que les thèmes extraits par le modèle puissent être discriminés géométriquement deux à deux. Les projections effectuées ont montré que les modèles utilisant la représentation sac de mots séparent nettement la plupart des thèmes dans le plan.
- On s'attend à ce qu'il y ait peu de termes pertinents pour représenter chaque thème. Cette capacité de sélectivité de termes représentatifs du modèle se vérifie à l'aide de la partie

droite de l'interface. Pour les modèles avec peu de thèmes ($K \leq 50$), peu de mots-clés ont une pertinence importante sont identifiés par le modèle pour chaque thème. Pour des modèles avec davantage de thèmes, on constate une multiplicité des termes avec un poids important dans chaque thème : cette multiplicité s'interprète comme un surapprentissage du modèle qui crée des thèmes regroupant des termes peu porteurs d'information. Ce constat peut se faire à l'aide de l'interface interactive mais se représente difficilement sur une page statique comme celle-ci. Le lecteur est donc invité à utiliser le « notebook » « Visualisations_LDA.ipynb » accessible dans le dépôt GitHub pour observer les résultats des visualisations PyLDAvis sur les trois revues.

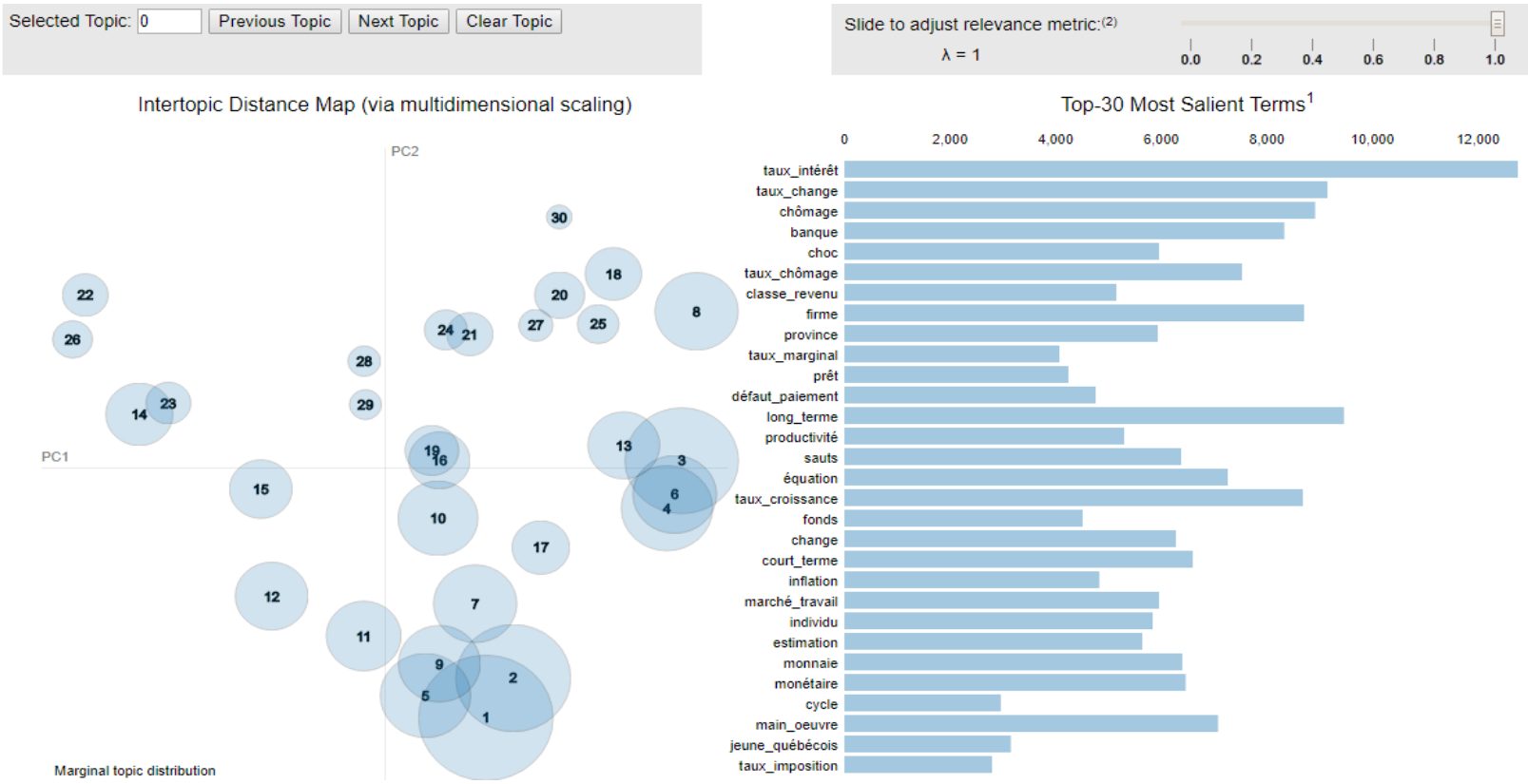

Figure 4.3 Comparaison de deux méthodes de projection d'un modèle LDA à 30 thèmes entraîné sur Actualité Économique. Sur cette page, projection ACP.

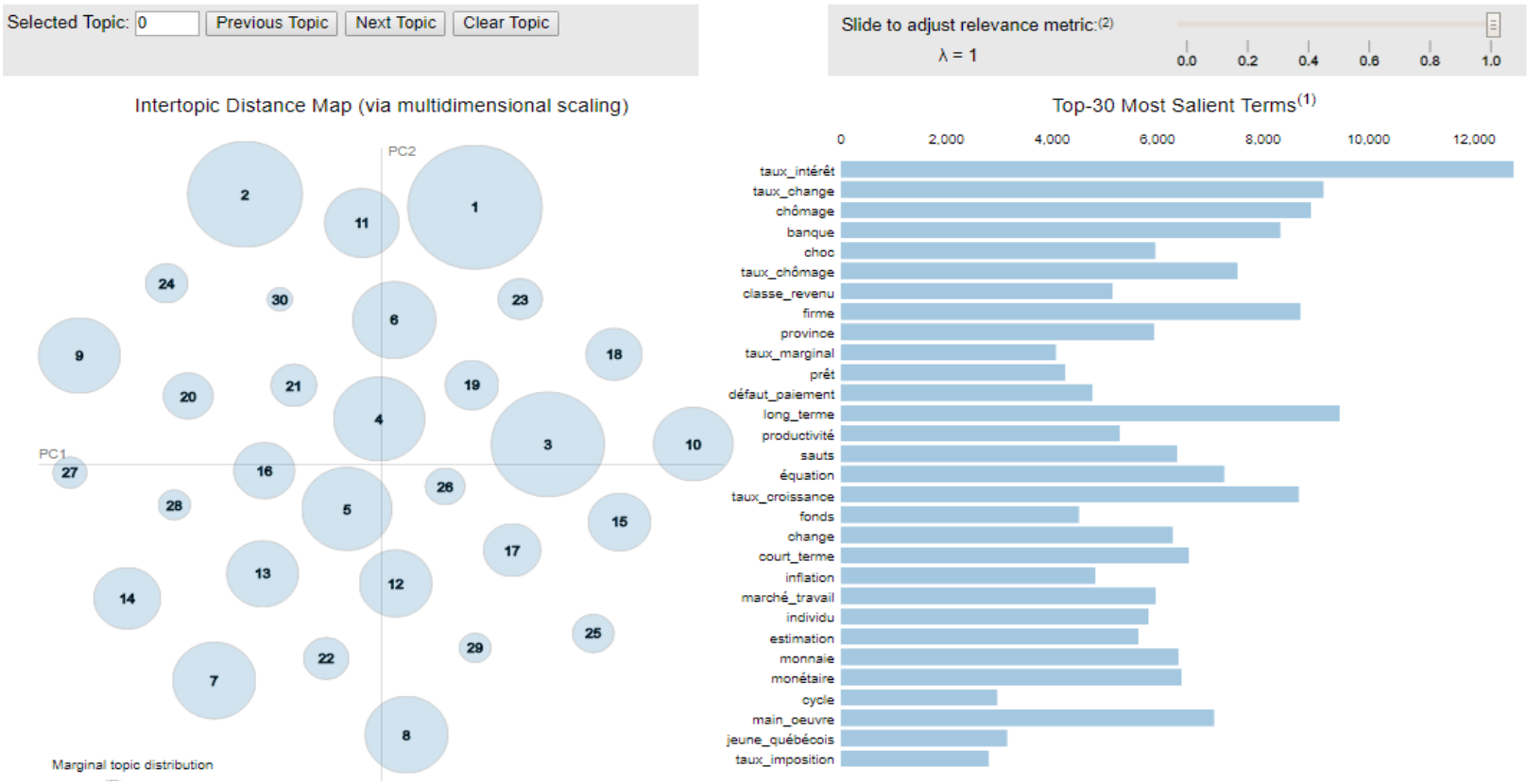

Figure 4.3 Comparaison de deux méthodes de projection d'un modèle LDA à 30 thèmes entraîné sur Actualité Économique. Sur cette page, projection t-SNE. (suite et fin)

4.2.1.2 Visualisation à l'échelle des documents

Outre la visualisation des résultats à l'échelle du corpus, on peut aussi exploiter l'information extraite par le modèle à l'échelle des documents. La Figure 4.4 montre les résultats d'un algorithme de regroupement sur le corpus qui projette le vecteur de mélange de thèmes représentant chaque document. Le nombre de groupes correspond au nombre de thèmes du modèle LDA, c'est-à-dire 10 dans le cas représenté. L'algorithme t-SNE a été utilisé pour construire un point à deux coordonnées pour chaque document, que l'on représente dans un plan. La couleur associée à chaque document correspond au thème majoritaire qui le compose. Le but de cette projection est donc de comparer la valeur discriminante des thèmes à partir de la projection thématique des documents où ils sont majoritaires. Les titres des articles sont utilisés pour évaluer cette discrimination thématique. La bibliothèque Python de visualisation Bokeh[8] a été utilisée pour comparer les titres d'un article avec le thème que l'algorithme de projection lui a attribué. Par exemple, pour Actualité Économique, les articles « Persistance du chômage et insertion » et « Comment lutter contre le chômage lorsque les travailleurs sont hétérogènes ? » sont étiquetés avec le thème 2, dont deux des mots-clés les plus représentatifs sont « taux_chômage » et « marché_travail » et sont projetés dans la même région du plan. Pour les trois revues, on observe une bonne capacité de discrimination des documents selon le thème majoritaire qui les compose. Cependant, il faut noter l'algorithme t-SNE ne permet pas de comparer une distance entre les groupes (qui pourrait servir de mesure de similarité) car il est construit sur une mesure de distance probabiliste (voir section 2.2.1). Finalement, des visualisations interactives fournissant les titres des articles sur le graphe peuvent être lancées à partir du notebook « Visualisations_LDA.ipynb ».

Ces différentes méthodes de visualisation offrent aux chercheurs en sciences sociales utilisant des modèles de thèmes une première possibilité d'interprétation des résultats. Ceux-ci peuvent ainsi identifier des thèmes visiblement incohérents, reconnaître des termes majoritaires dans plusieurs thèmes, interagir et appréhender le corpus d'étude à l'aide d'une projection 2D des thèmes d'une part et des représentations thématiques des documents d'autre part. Ils peuvent aussi dégager les tendances thématiques générales traitées dans un corpus en vue d'une analyse plus spécifique. La

[8] https://bokeh.pydata.org/en/latest/

visualisation des modèles de thèmes n'est pas une tâche facile à cause de la complexité des résultats produits par ces modèles. Afin de pouvoir comparer rigoureusement les résultats des modèles de thèmes, il reste nécessaire d'introduire des mesures d'évaluation interne et externe.

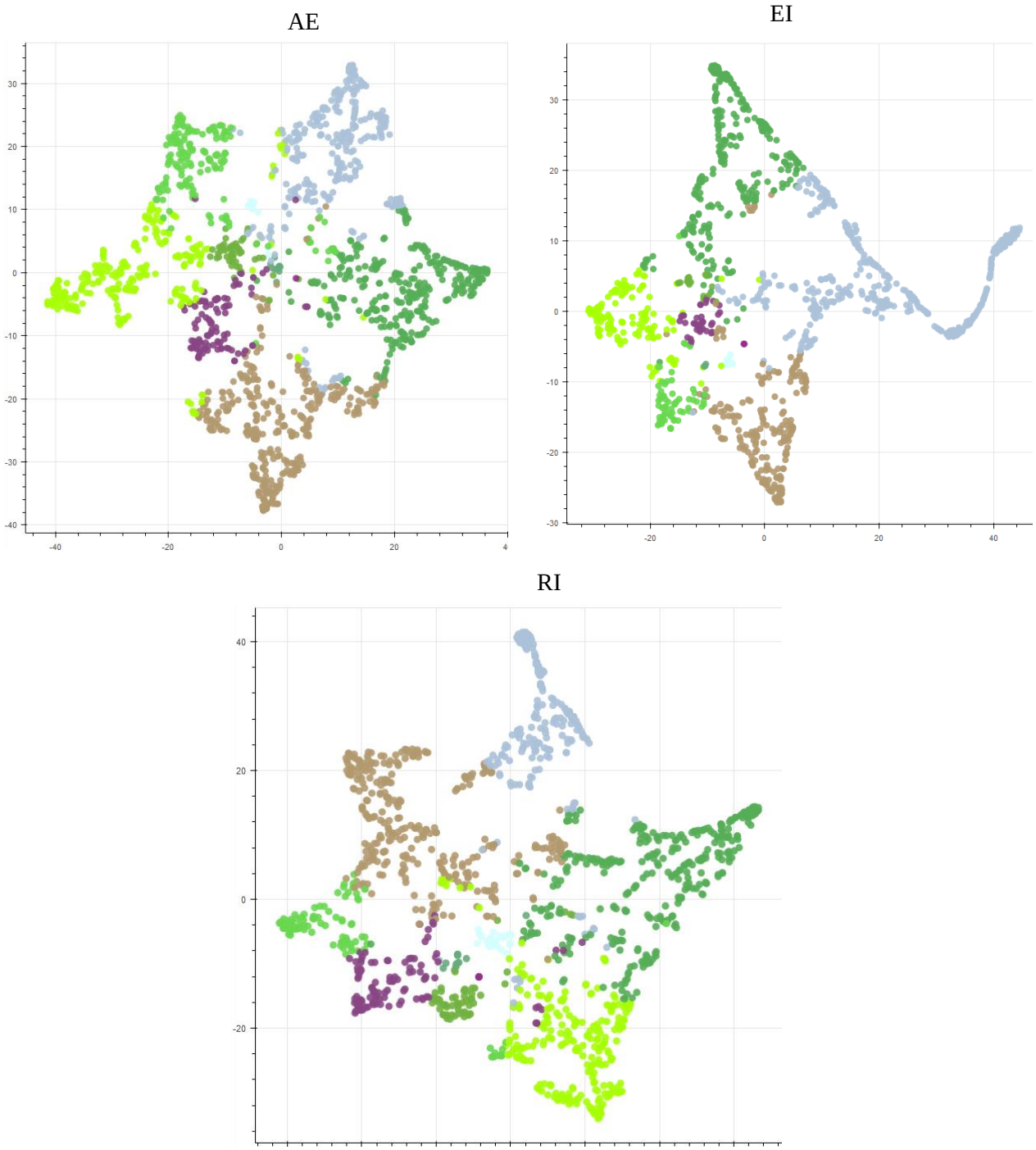

Figure 4.4 Regroupement t-SNE des 3 revues à partir d'un modèle LDA à 10 thèmes. Un point correspond à un document, et sa couleur au thème LDA majoritaire qui le compose.

4.2.2 Évaluation automatique : mesures de perplexité et de cohérence

4.2.2.1 Étude de la perplexité

La mesure traditionnelle pour comparer les modèles LDA est la perplexité (Blei *et al.*, 2003). La perplexité mesure la capacité du modèle à prédire la composition thématique de nouveaux documents. Pour chaque revue ont été comparées deux modalités du modèle LDA : le nombre de thèmes du modèle et la lemmatisation ou non du vocabulaire. La Figure 4.5 présente les courbes de perplexité obtenues avec une échelle logarithmique. L'ordonnée correspond à la perplexité mesurée sur un modèle LDA et l'abscisse au nombre de thèmes *a priori* de ce modèle LDA. Les croix en noire correspondent aux perplexité obtenues pour les modèles sans lemmatisation tandis que celles en jaune correspondent aux perplexités mesurées sur les modèles avec lemmatisation. L'ordre de grandeur et la tendance suivie par les valeurs de perplexité sont similaires pour chaque revue.

Influence de la lemmatisation du vocabulaire. Aucun effet significatif de la lemmatisation du vocabulaire n'est observé (différence entre croix noire et jaune sur une verticale constante). Ce résultat est en accord avec le peu de littérature existante ayant étudié l'effet de la lemmatisation sur les modèles de thèmes (Schofield et Mimno, 2016).

Influence du nombre de thèmes. D'autre part, Les modèles utilisant une représentation sac de mots du corpus montrent une valeur de perplexité stable autour de -10 pour un nombre de thèmes inférieur à 70. Un décrochement apparaît entre les modèles à 70 et 80 thèmes, léger pour les revues AE et EI et remarquable pour la revue RI (pour laquelle la log-perplexité chute de -10 à -14). La cause de ce décrochement n'est pas comprise. La lemmatisation n'ayant aucun effet sur la visualisation ni sur la perplexité des résultats, il a été décidé de ne pas appliquer cette étape dans les analyses suivantes afin de diminuer la complexité de l'étape de préparation des données.

Pour résumer, après avoir effectué ces premières analyses, la configuration « sac de mots/sans lemmatisation » a été retenue pour les études suivantes. De plus, mesurer la perplexité des modèles apparaît comme inutile dans l'évaluation des modèles LDA au vu de l'absence de corrélation entre sa valeur et l'interprétabilité visuelle des thèmes.

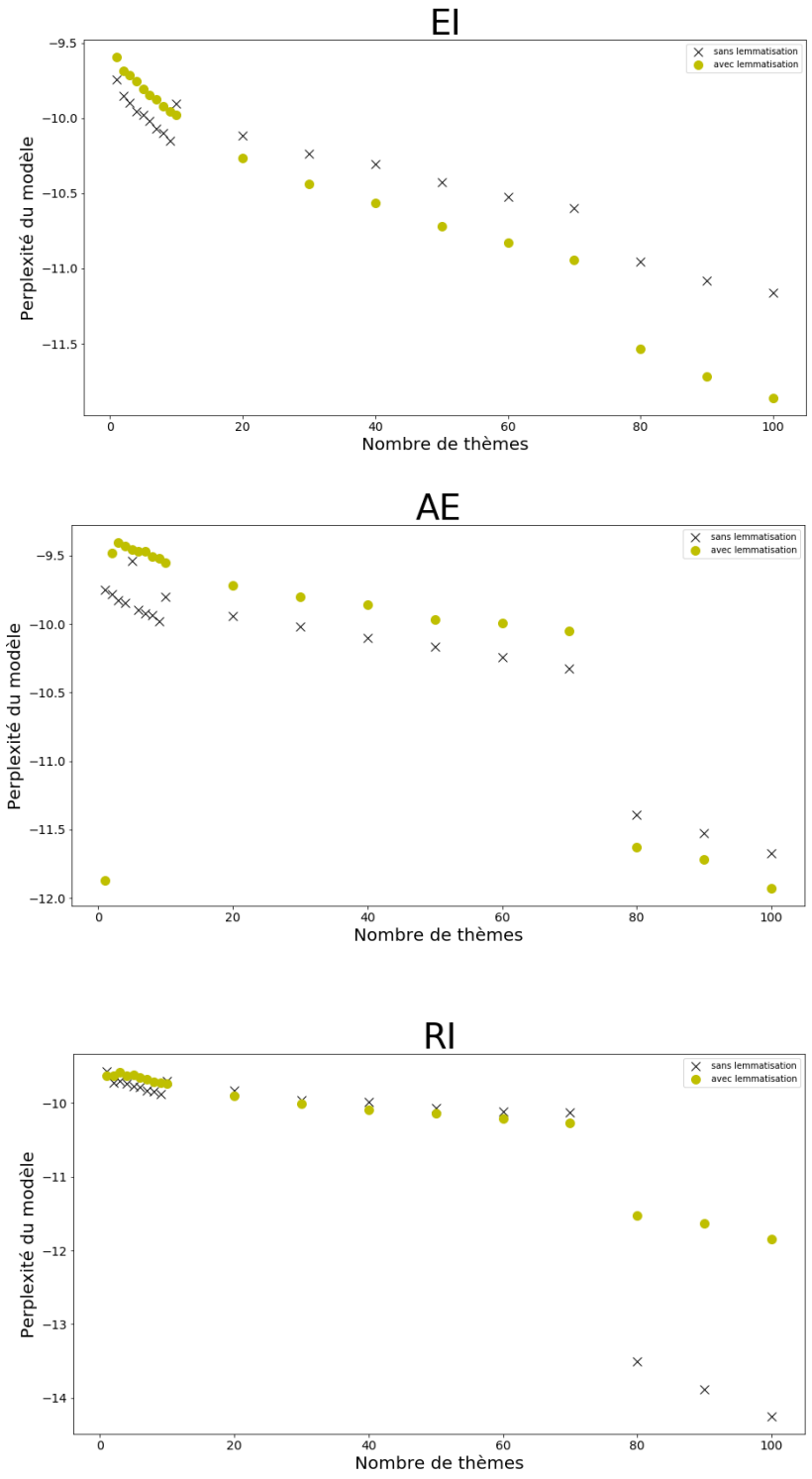

Figure 4.5 Influence du nombre de thèmes *a priori* sur la perplexité des modèles LDA

L'ordonnée est en échelle logarithmique.

4.2.2.2 Étude de la cohérence

La perplexité évalue la capacité prédictive du modèle de thèmes, autrement dit elle mesure sa vraisemblance à générer un document non observé pendant l'entraînement. Mais elle ne donne aucune information explicative sur la cohérence des thèmes produits. À la suite des premiers résultats des visualisations, il a été décidé d'évaluer la cohérence des modèles utilisant une matrice sac de mots du corpus et sans lemmatisation du vocabulaire. Les quatre mesures de cohérence présentées dans la section 0 ont été évaluées pour des modèles LDA appliqués sur les trois revues. Les résultats sont présentés à la Figure 4.6. Une couleur et une forme de points sont caractéristiques de chaque revue : croix verte pour AE, cercle jaune pour EI et carré rouge pour RI. Chaque graphe correspond représente l'évolution d'une des quatre mesures de cohérence en fonction du nombre de thèmes *a priori* du modèle LDA.

Notons que les valeurs de cohérence ne peuvent se comparer de façon absolue entre les différentes mesures. L'ordre de grandeur des valeurs obtenues pour les quatre mesures est le même que celui indiqué dans la littérature (Stevens *et al.*, 2012). Pour les trois revues et les quatre mesures de cohérence, on observe des valeurs de cohérences maximales pour un nombre de thèmes K compris entre 5 et 10, puis une convergence décroissante vers une valeur cohérence qui se stabilise à partir de K = 50 (excepté pour la mesure C_V où la stabilisation apparaît dès K = 10). Cela indique qu'un nombre de thèmes supérieur à 10 n'apporte pas d'information utile supplémentaire sur la collection analysée selon les métriques de cohérence. Les 4 courbes des revues EI et RI montrent des tendances similaires, avec des valeurs de cohérence en moyenne supérieures à celles obtenues pour la revue AE. Les courbes des cohérences C_{UCI}, C_{NPMI} et C_{UMass} ont une évolution semblable avec le nombre de thèmes. Or, elles se basent sur des statistiques de cooccurrence, C_{UMass} à partir du corpus d'étude et C_{UCI} et C_{NPMI} à partir de Wikipédia. On en déduit que pour les revues étudiées, l'utilisation de Wikipédia comme corpus externe de statistiques textuelles n'apporte pas d'information supplémentaire dans le calcul de la cohérence thématique.

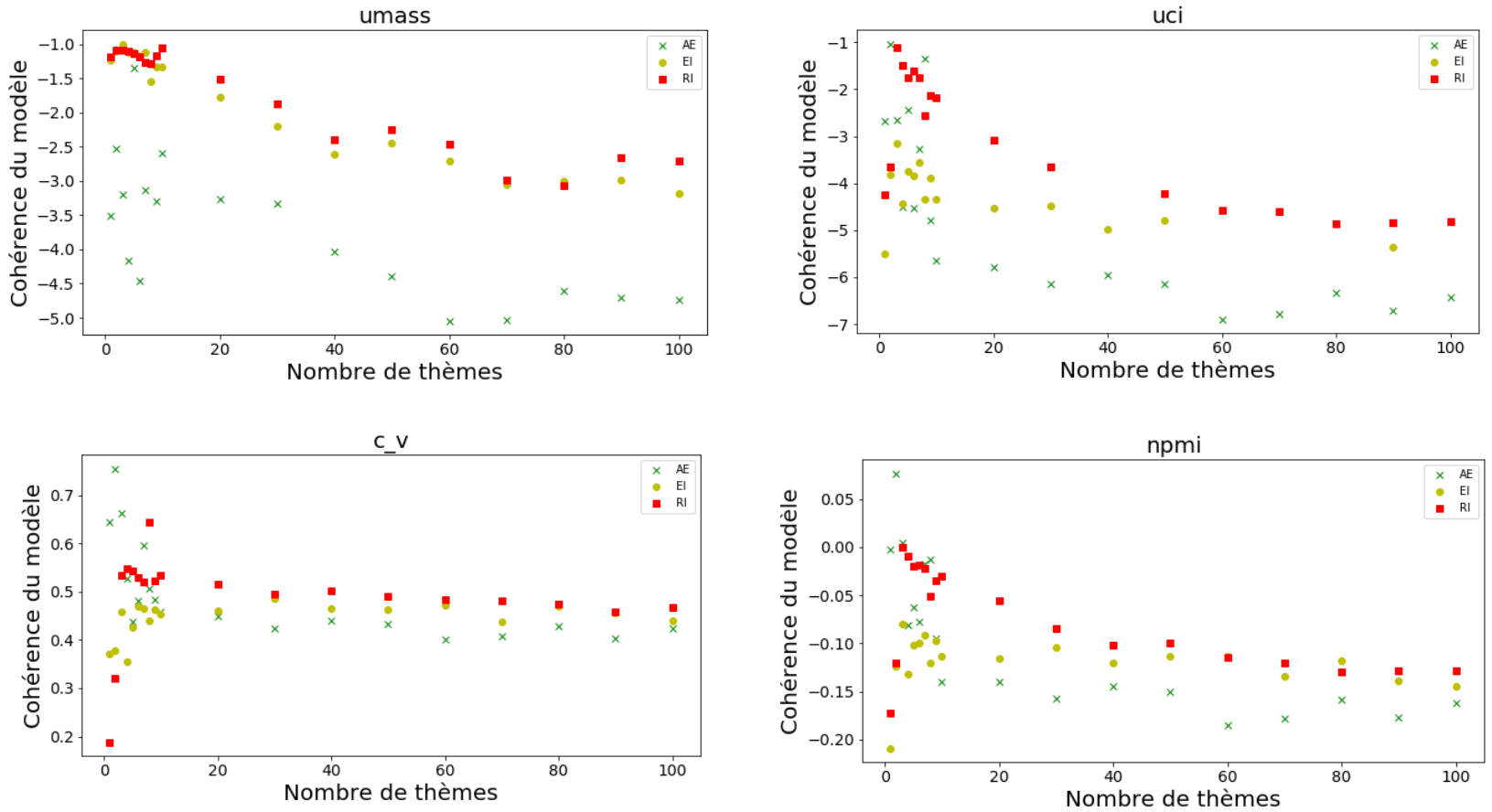

Figure 4.6 Influence du nombre de thèmes *a priori* sur la cohérence des modèles LDA.

Une couleur et une forme de points sont caractéristiques de chaque revue : croix verte pour AE, cercle jaune pour EI et carré rouge pour RI.

4.2.2.3 Étude de la fiabilité

L'étude de la fiabilité des modèles est rarement entreprise dans la littérature des modèles de thèmes. Pourtant elle est cruciale si l'on souhaite tester la reproductibilité des résultats obtenus, en particulier lorsque l'on travaille avec des modèles probabilistes. Ce manque de fiabilité est en outre un des principaux freins à l'adoption des modèles de thèmes par la communauté de chercheurs en SHS (Ramage *et al.*, 2009).

On se propose d'étudier la fiabilité des modèles LDA à l'aide des mesures de perplexité et de cohérence. Cinq modèles LDA à 10 thèmes et à 40 thèmes ont été entraînés indépendamment sur chaque revue, avec une représentation sac de mots et sans lemmatisation du vocabulaire. La configuration d'entraînement des modèles est la même pour les cinq versions des modèles. Le but est d'évaluer l'incertitude causée par l'initialisation probabiliste du tirage des mots pour chaque thème et des mélanges de thèmes pour chaque document (voir le processus génératif du LDA à la section 0).

La Figure 4.7 montre la variabilité des résultats de perplexité obtenus pour les 2 types de modèles LDA et les 3 revues. Par exemple, les valeurs que l'on observe pour « LDA AE 10 » correspondent à la variabilité des mesures de perplexité obtenues sur les 5 modèles LDA à 10 thèmes entraînés indépendamment sur la revue Actualité Économique.

Toutes les mesures de perplexité tombent entre le premier et le troisième quartile sauf une pour le modèle à 40 thèmes pour la revue Études Internationales. Les écarts-types mesurés, indiqués dans le Tableau 4.2 sont faibles. Il n'y a pas d'effet mesurable du nombre de thèmes sur la variabilité des résultats de perplexité. Par conséquent, les résultats des modèles LDA se montrent fiables selon la métrique de perplexité.

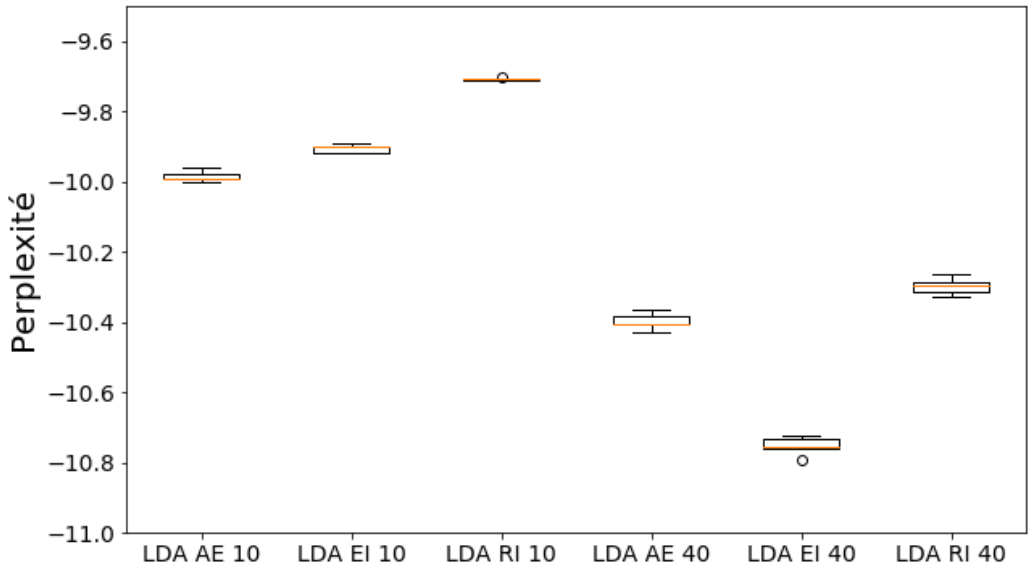

Figure 4.7 Boîte à moustaches des résultats de perplexité pour l'étude de fiabilité

La Figure 4.8 présente la variabilité des valeurs de cohérence obtenues, selon les quatre mesures populaires. La grille de lecture est la même que pour la Figure 4.7. De même que pour les résultats de perplexité, on constate un très faible écart de valeur de cohérence entre les différents modèles, indépendamment de la revue ou du nombre de thèmes. La mesure C_{UCI} montre le plus de variabilité. Pour les mesures C_{UCI}, C_V et C_{NPMI} les modèles à 10 thèmes présentent plus de variabilité que les modèles à 40 thèmes.

Tableau 4.2 Mesures de perplexité pour l'étude de fiabilité

Revue	Modèle	Perplexité moyenne
Actualité Économique	LDA 10 thèmes	- 9,986 +/- 0,015
	LDA 40 thèmes	-10,398 +/- 0,022
Études Internationales	LDA 10 thèmes	-9,905 +/- 0,011
	LDA 40 thèmes	-10,753 +/- 0,024
Relations Industrielles	LDA 10 thèmes	-9,709 +/- 0,003
	LDA 40 thèmes	-10,299 +/- 0,022

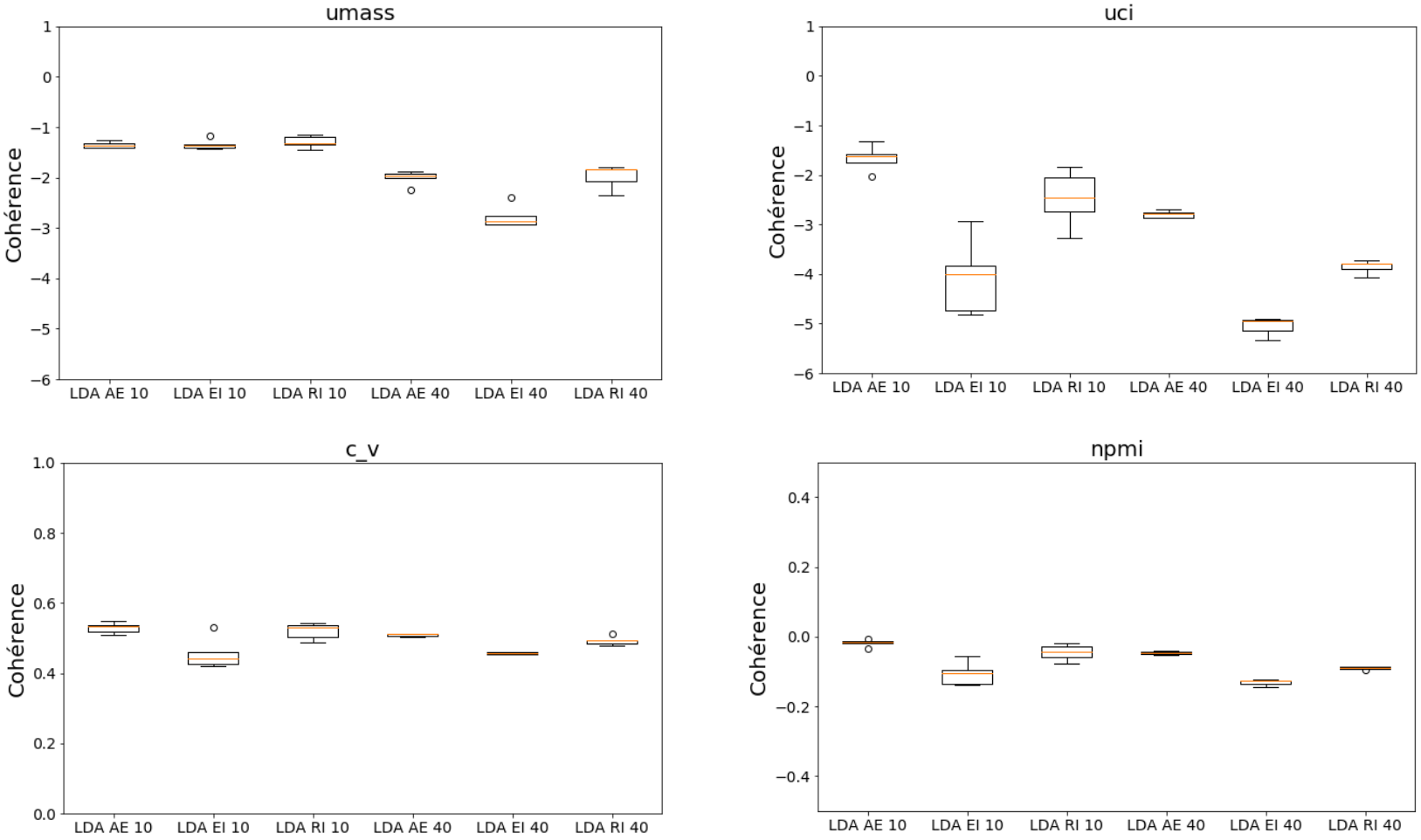

Figure 4.8 Boîte à moustaches des résultats de cohérence pour l'étude de fiabilité

Les modèles LDA évalués présentent donc une bonne fiabilité selon les mesures de cohérence et de perplexité : munis des mêmes hyperparamètres d'entraînement, deux modèles LDA obtiendront des scores semblables.

Qualitativement, on se propose d'étudier la fiabilité des modèles en comparant les termes les plus représentatifs extraits pour l'ensemble des thèmes de chaque version d'un modèle LDA. Autrement dit, on recense l'ensemble des dix jetons représentatifs de chaque thème pour un modèle LDA donné. Puis on compare deux à deux cette liste de jetons entre les modèles avec un même nombre de thèmes. Par exemple, pour les cinq modèles LDA à 10 thèmes appliqués à la revue AE, on extrait les dix jetons les plus représentatifs de chacun des 10 thèmes, ce qui constitue une liste de 100 jetons et on compte ensuite le pourcentage de jetons communs entre chaque paire de modèles. On obtient finalement un pourcentage qu'on nomme « similarité sémantique globale », défini pour un modèle LDA à K thèmes. L'Équation 4.4 en donne la définition mathématique :

$$\%\ similarité\ sémantique\ globale\ (LDA\ avec\ K\ thèmes) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le \binom{n}{2}} \frac{Card\left(S_i^K \cap S_j^K\right)}{top_m \times K} \times 100$$

Équation 4.4 Définition de la similarité sémantique d'un modèle LDA à K thèmes

Où :

- K = nombre de thèmes du modèle LDA considéré
- n = nombre de versions entraînées de ce modèle
- top_m : nombre de jetons représentatifs conservés pour chaque thème
- S_i^K : Liste de tous les jetons représentatifs conservés pour la version *i* du modèle LDA à *K* thèmes

Ici, les valeurs de ces paramètres sont : $\{K \in [10,40]$; $n = 5$; $top_m = 10\}$.

La formule devient pour ce cas particulier :

$$\%\ s.s.g.\ (LDA\ avec\ K\ thèmes) = \frac{1}{\mathrm{K}} \sum_{1 \le i < j \le \binom{n}{2}} Card\left(S_i^K \cap S_j^K\right)$$

Les résultats de similarité sémantique globale sont présentés dans le Tableau 4.3.

Tableau 4.3 Mesures de cohérence et de similarité sémantique pour l'étude de fiabilité

Revue	Modèle	C_{UMass}	C_{UCI}	C_V	C_{NPMI}	Similarité sémantique globale
Actualité Économique	LDA 10 thèmes	-1.35 +/- 0.054	-1.66 +/- 0.233	0.529 +/- 0.014	-0.017 +/- 0.009	70.9% +/- 2,6
	LDA 40 thèmes	-2.001 +/- 0.125	-2.794 +/- 0.065	0.509 +/- 0.005	-0.047 +/- 0.004	69.0% +/- 1,6
Études Internationales	LDA 10 thèmes	-1.341 +/- 0.086	-4.063 +/- 0.689	0.456 +/- 0.04	-0.105 +/- 0.03	69,6% +/- 3,9
	LDA 40 thèmes	-2.778 +/- 0.199	-5.052 +/- 0.166	0.457 +/- 0.003	-0.131 +/- 0.008	66.2% +/- 2,3
Relations Industrielles	LDA 10 thèmes	-1.288 +/- 0.106	-2.475 +/- 0.512	0.52 +/- 0.022	-0.045 +/- 0.021	63,8% +/- 3,5
	LDA 40 thèmes	-1.978 +/- 0.207	-3.852 +/- 0.117	0.493 +/- 0.011	-0.09 +/- 0.004	69,1% +/- 2,2

La similarité sémantique obtenue est comprise entre 63% et 71%. Ce résultat valide qualitativement la capacité des modèles LDA à produire des thèmes similaires une fois fixés les hyperparamètres du modèle. Notons que la nature probabiliste du modèle ne permet pas d'obtenir des thèmes complètement identiques entre les différentes versions d'un même modèle. Cette fluctuation probabiliste doit être connue des chercheurs en sciences sociales qui souhaitent utiliser des modèles LDA pour extraire de l'information d'un corpus.

Les mesures automatiques d'évaluation comme la perplexité et la cohérence aident l'utilisateur des modèles de thèmes à décrire des tendances générales concernant la qualité des thèmes extraits. Toutefois, elles ne donnent aucune information sur l'utilité opérationnelle de la représentation thématique pour des chercheurs en SHS. Pour cela, il est nécessaire d'évaluer le modèle sur une tâche spécifique : la recherche de documents.

CHAPITRE 5 INTÉGRATION DU LDA DANS UNE TÂCHE DE RECHERCHE DE DOCUMENTS

À ce stade de la recherche, différents modèles LDA ont été entraînés sur les trois revues. Les étapes de visualisation et d'évaluation automatique ont permis d'identifier la configuration « corpus sac de mots/sans lemmatisation du vocabulaire » comme la plus adaptée au corpus. Afin de valider les résultats de ces modèles, il est apparu nécessaire de construire une tâche spécifique d'évaluation externe. Cette tâche part de la situation opérationnelle suivante : un chercheur dispose d'un corpus d'articles en SHS et souhaite récupérer certains articles à partir d'une requête spécifique. Le but est de développer un algorithme de recherche qui propose des articles pertinents vis-à-vis de la requête. En particulier, il s'agit de s'appuyer sur la représentation informatique calculée par le modèle LDA pour mener la fouille.

Des modèles LDA entraînés sur les trois revues ont donc été utilisés pour effectuer trois recherches de thématiques au sein de chaque revue, à l'aide de deux approches de fouille différentes. Pour chaque recherche, les deux approches proposent des listes de 5 titres d'articles publiés dans la revue qu'ils estiment pertinents par rapport à la requête de l'utilisateur. Ces listes de titres d'articles ont ensuite été proposées à des chercheurs en SHS qui en ont indiqué le niveau de pertinence. On présente maintenant les deux approches de fouille et les choix méthodologiques de construction de la tâche.

5.1 Construction de la tâche de fouille de documents

Le but de l'évaluation externe est de mesurer indirectement l'utilité opérationnelle des résultats des modèles LDA sur une tâche de fouille de documents. Ce but se scinde en quatre objectifs spécifiques pour répondre à la seconde question de recherche :

(1) Comparer la performance de deux algorithmes de recherche d'information basés sur des modèles LDA sur la pertinence des résultats.
(2) Mesurer l'influence du nombre de thèmes des modèles LDA sur la pertinence des résultats.
(3) Mesurer l'influence du domaine de la revue sur lequel le modèle est entraîné sur la pertinence des résultats.
(4) Mesurer l'influence de la spécificité de la requête sur la pertinence des résultats.

L'objectif (1) vise à texter deux méthodes d'intégration du LDA pour la tâche de recherche de documents : une qui utilise une approche fréquentielle et l'autre une approche vectorielle.
L'objectif (2) vise à tester un hyperparamètre clé du LDA, le nombre de thèmes *a priori* qui peut modifier les thèmes obtenus comme on l'a étudié au chapitre 4.
Les objectifs (3) et (4) sont motivés par la nature du LDA. En effet, le LDA est un modèle qui extrait de l'information globale du corpus, sous forme d'une modélisation thématique. Par conséquent, on peut se demander si la qualité de cette information sera équivalente entre les trois revues d'une part et si d'autre part elle sera à même de répondre à des requêtes très spécifiques.
La tâche de fouille de documents se décompose en 4 étapes, résumées par la Figure 5.1.

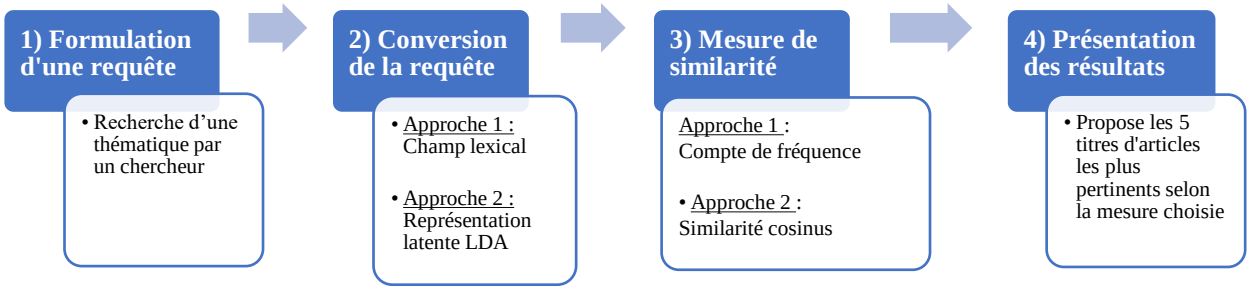

Figure 5.1 Pipeline de la fouille de documents

Chaque étape de cette tâche nécessite de procéder à des choix méthodologiques.

5.1.1 Choix de requête

Le principe est de n'utiliser que le contenu textuel des articles (et donc d'exclure les métadonnées comme les noms d'auteurs, les citations, etc.) afin d'isoler l'information apportée par les modèles de thèmes sur cette tâche de fouille de documents. La requête contiendra donc uniquement des mots-clés liés à une thématique qui intéresse le chercheur, comme le terme « Macroéconomie » par exemple. Dans une situation réelle, c'est le chercheur qui choisirait la thématique avec laquelle il souhaite fouiller le corpus. Mais pour cette expérience, les thématiques ont été choisies en amont de l'évaluation par les chercheurs en SHS pour assurer la présence de la thématique au sein d'une revue donnée et répondre aux objectifs. Ces thématiques ont été choisies pour chaque revue selon deux critères :

Critère 1 : La thématique de la requête doit être traitée dans la revue (Objectif (1))

Critère 2 : Les thématiques de requête doivent présenter des spécificités différentes (Objectif (4))

Le premier critère assure la présence de certains documents parmi le corpus traitant de la thématique recherchée. Le second critère se justifie par le caractère d'extraction thématique global des modèles LDA. En effet les modèles LDA extraient une représentation globale du corpus, en identifiant des groupes de termes de forte co-occurrence : on ne s'attend pas à ce qu'ils soient capables de proposer des informations spécifiques à un unique document.

Pour le premier critère, nous avons identifié les intérêts de recherche des auteurs présents au comité de rédaction de chaque revue à l'aide de leur page institutionnelle, constituant ainsi une liste de thématiques *a priori* pour chaque revue. On qualifie ces thématiques d'*a priori* car on s'attend à ce que certaines d'entre elles soient traitées dans les revues du corpus, puisqu'elles sont des intérêts de recherche des auteurs qui contribuent aux revues et sélectionnent les articles qui y sont publiés. Ensuite, nous avons déterminé un champ lexical pour chacune de ces thématiques à l'aide du site rimesolides.com[9]. Le champ lexical (ou sémantique) désigne ici un ensemble de noms, d'adjectifs et de verbes qui présentent une corrélation sémantique entre eux. Dans le cadre de ce travail, cette corrélation sémantique sera mesurée à l'aide de comptes de cooccurrence. Pour un mot-clé donné, l'outil proposé par ce site renvoie une liste de mots utilisés fréquemment avec le mot-clé dans des bases de données externes (Wikipédia, réseaux sociaux, romans). L'algorithme n'est pas en accès ouvert mais pour les besoins de notre recherche, les résultats qu'il propose nous sont apparus suffisants. Pour chaque thématique (i.e. mot-clé), on constitue ainsi une liste de termes associés qui lui sont proches sémantiquement : cette liste de termes correspond au champ lexical de la thématique. La Figure 5.2 illustre la construction de ces champs lexicaux (ou sémantiques) sur la revue Actualité Économique.

[9] https://www.rimessolides.com/motscles.aspx

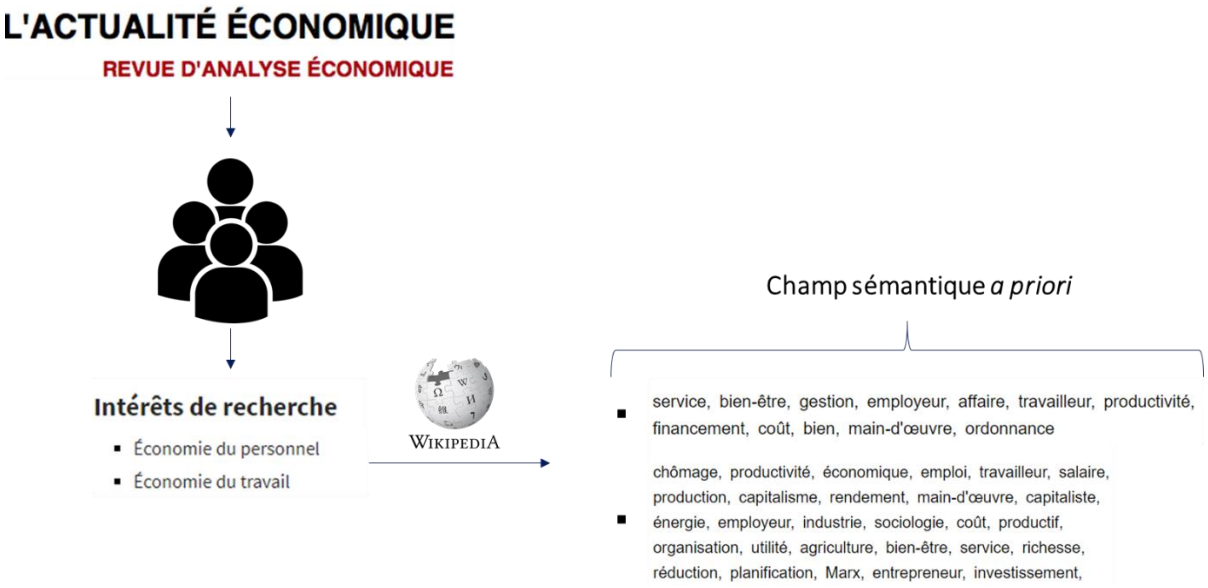

Figure 5.2 Illustration de la construction des champs sémantiques *a priori* présentes dans la revue AE.

À partir de la liste des rédacteurs de la revue, on identifie les intérêts de recherche puis à l'aide d'une mesure de cooccurrences basée sur Wikipédia, on leur associe un champ sémantique *a priori.*

Pour assurer le premier critère, à savoir que les mots-clés utilisés pour la requête sont traités par au moins cinq articles d'une revue donnée, on procède à un compte d'occurrences. Pour l'ensemble des mots composant chacun des champs lexicaux construits précédemment, on compte leur nombre d'occurrences dans chaque article d'une revue. Si pour un article le total d'occurrences dépasse cinq, on considère que l'article traite de cette thématique. Ce choix de cinq articles a été fait de façon heuristique et mériterait d'être considéré davantage lors de l'application de la méthode à d'autres corpus. S'il y a plus de cinq articles qui traitent de cette thématique, alors on considère qu'elle respecte le premier critère. À la fin de cette première étape de filtre, on a conservé pour chaque revue des thématiques qui sont traitées dans au moins cinq articles de la revue correspondante. La première ligne du Tableau 5.1 présente la liste de thématiques pour chaque revue. Le second critère permet de conserver des thématiques de spécificité différente pour mener la recherche de documents. On détaille maintenant ce critère.

Pour le second critère concernant la granularité des mots-clés choisis, le travail s'appuie sur *l'Australian and New Zealand Standard Research Classification* (ANZSRC)[10]. Ce guide de classification datant de 2008 propose une structure hiérarchique positionnant l'ensemble des domaines de recherche et développement. La structure se décompose selon trois niveaux de granularité : DIVISION — GROUPE — DOMAINE. Dans le cas où les thématiques identifiées à l'aide du premier critère ne sont présentes dans aucun de ces trois niveaux, on les considère comme appartenant à un « SOUS-DOMAINE ». Les divisions « *Economics* » et « *Economic Framework* » pour la revue AE et la division « *Law, Politics and Community Services* » pour les revues EI et RI ont été identifiées. Plus spécifiquement, pour la revue AE, ont été retenus le groupe « Macroéconomique » (*Macroeconomics)* ainsi que les domaines « Économie internationale » (*International Economics*) et « Économie monétaire et financière » (*Financial Economics)*. Pour la revue EI, aucune thématique issue du premier critère ne correspond à un groupe ou domaine donc ont été choisis les sous-domaines « Démocratisation », « Dissuasion » et « Sécurité Internationale ». Pour la revue RI, a été retenu le groupe « Travail » et les sous-domaines « Convention Collective » et « Théories de la justice et de l'équité ». Finalement, une fois filtrés les mots-clés de recherche potentiels à l'aide du premier critère, on a conservé des mots-clés ayant une granularité (groupe, domaine, sous-domaine) différente pour chaque revue.

Les trois mots-clés sélectionnés pour chaque revue à l'aide de ces deux critères sont présentés dans le Tableau 5.1. On obtient finalement deux mots-clés avec une granularité de type « Groupe », deux mots-clés avec une granularité de type « Domaine » et cinq mots-clés avec une granularité plus spécifique (« Sous-domaine »).

---

10 https://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/4AE1B46AE2048A28CA25741800044242?opendocument

Tableau 5.1 Choix des thématiques pour les requêtes pour chaque revue.

En gras : thématiques présentes dans la revue. Les autres sont considérées comme non traitées dans la revue.

		ACTUALITÉ ÉCONOMIQUE	ÉTUDES INTERNATIONALES	RELATIONS INDUSTRIELLES
1er critère	Thématiques *a priori* (identifiée à l'aide des pages institutionnelles des rédacteurs de chaque revue) **En gras, thématiques présentes dans la revue**	**Macroéconomie** – **Politiques monétaires** – **cycle économique** – Finances internationales – Économétrie appliquée – **Économie du personnel** – **Économie monétaire et financière** – **Économie du travail** – **Économie internationale** – **Croissance et développement** – **Microéconomie** – Politique incitation – Finance quantitative – **Économétrie** – **Économie** – Gestion des risques financiers – **Méthodes numériques**	Interventions étrangères dans les conflits intra-étatiques – **Méthodes qualitatives** – **Politique étrangère des États–Unis** – **Politique étrangère et de sécurité du Canada** – **Science politique** – Théories des relations internationales – **Afrique subsaharienne** – Alternance politique – **Démocratisation** – Famille et politique – Hérédité en politique – Mobilisations collectives – **Notabilités** – **Politique comparée** – Science politique – **Sénégal** – Socio–histoire de l'État – Sociologie des institutions – Sociologie du religieux – Relations internationales – Gouvernance – Environnement – **Afrique** – **Relations internationales** – **Sociologie de la discipline des relations internationales** – Migration de mariage – **Études critiques de sécurité** – **Émotions, gouvernementalité et sécurité** – **Philosophie politique** – Etudes stratégiques – Politique de défense – **Relation civil-armée** – **Sécurité internationale** – **Conflits civils** – **Sociologie politique des organisations internationales** – **Gouvernance** – Elaboration des politiques – Diplomatie multilatérale – **Théorie sociale** – Relations nord / sud – Organisations internationales – **Coopération internationale** – **Coopération militaires** – **Alliances** – **Dissuasion**	**Travail** – **Syndicalisme** – Gestion des Ressources Humaines – **Politiques publiques du travail et de l'emploi** – **Ergonomie** – Santé et Sécurité au travail – **Individualisme dans les relations d'emploi** – **Relations de travail dans les pays de l'Asie** – **Pacifique** – **Relations industrielles** – Équité salariale et évaluation des emplois – **Gestion de la rémunération** – **Théories de la justice et l'équité** – Économie politique et micro-politique de la santé et de la sécurité du travail – **Intervention en santé et sécurité du travail et en santé mentale au travail** – **Organisation de la prévention dans les milieux de travail** – **Systèmes de gestion de la santé et de la sécurité du travail** – **Industrie de la construction** – **Rapports collectifs du travail** – Action collective et mouvements sociaux – **Protection sociale** – **Transformations du travail** – Travail atypique – Gestion des ressources humaines – Comportement organisationnel – **Conciliation travail-famille** – Culture organisationnelle – Mieux–être au travail – Roulement du personnel – **Formation et développement des compétences** – Gestion du rendement – **Qualité du travail et de l'emploi** – **Sécurisation des trajectoires professionnelles** – **Mondialisation** – **Sociologie du travail, de l'emploi et des professions** – Insertion professionnelle des jeunes – **Restructuration d'entreprise** – Entreprise multinationale – Sous–traitance – **Délocalisation** – **Relations du travail** – **Convention collective** – **Négociation collective** – Altermondialisme – Ententes commerciales internationales – Immigrants et emploi – Interaction patronale, syndicale et gouvernementale – **Nouvelles technologies** – **Organisations syndicales** – **Planification et gestion stratégique** – **Droit du travail** – **Droit international du travail** – Relations du travail – **Droits de la personne** – Travail atypique – **Temps de travail**

Niveaux de granularité :

Groupe
Domaine
Sous-domaine

Une fois que l'on dispose des thématiques à rechercher, on souhaite le convertir pour que l'ordinateur puisse renvoyer une liste d'articles qu'il considère comme sémantiquement proches de la requête.

5.1.2 Conversion de requête

Deux approches de conversion de la requête sont comparées : l'approche 1 repose sur la construction de différents champs lexicaux et l'approche 2 utilise l'espace latent du modèle LDA. Le but de ces deux approches est de déterminer la meilleure modalité d'utilisation des modèles LDA pour cette tâche. L'approche 1 utilise une fréquence de jetons tandis que l'approche 2 utilise une représentation vectorielle.

Approche 1

La première approche de conversion se base sur la même idée que le plus simple des algorithmes de fouille, à savoir compter le nombre d'occurrences du mot-clé dans chacun des articles du corpus et renvoyer ceux où le mot-clé d'intérêt a la plus grande fréquence. L'ajout de l'approche 1 est de ne pas se limiter au mot-clé dans le calcul d'occurrences mais plutôt d'utiliser un « champ lexical » qui lui est rattaché. Le champ lexical est utilisé dans le même sens que précédemment. Cette méthode repose sur l'hypothèse que si un article contient des termes proches sémantiquement du mot-clé de recherche, alors il est susceptible d'être pertinent. De fait, les termes appartenant à un même champ lexical ne sont pas nécessairement synonymes mais sont utilisés dans un même contexte pour désigner un même phénomène. Trois procédés de construction de ce champ lexical sont proposés :

1) Le premier procédé est celui détaillé plus haut dans l'identification des mots-clés des requêtes. On utilise le moteur du site rimessolides.com qui se base sur des statistiques de cooccurrences de mots construites à l'aide de bases de données de références telle que Wikipédia. Pour chaque mot-clé de recherche, on récupère une liste de 30 termes présentant une forte cooccurrence avec le mot-clé. Cette méthode sera représentée par l'acronyme ***a priori***.
2) Le deuxième procédé consiste à établir un lien entre le champ lexical construit à l'aide du premier procédé et les listes de mots représentatifs de chaque thème LDA. À partir des modèles LDA à 10 thèmes, la correspondance est établie en identifiant les termes communs

entre les listes de mots deux à deux. Par exemple, si le champ lexical du thème « Macroéconomie » contient le mot « inflation » et que l'un des thèmes LDA contient le mot « inflation » dans l'un de ses dix jetons les plus représentatifs, alors on considère qu'il y a un lien entre ce thème LDA et le thème *a priori* « Macroéconomie ». Une fois cette correspondance établie, si l'on s'intéresse à la requête « Macroéconomie », on utilisera non pas le champ lexical issu du premier procédé mais plutôt l'ensemble des jetons du thème LDA associé au thème « Macroéconomie ». Cette méthode sera représentée par l'acronyme **LDA.**

3) Le troisième procédé dérive du précédent. Une fois établie la correspondance entre les thèmes *a priori* et les thèmes LDA, on identifie pour chaque thème LDA les cinq articles les plus représentatifs du thème (i.e. avec le poids le plus fort pour ce thème). Puis on constitue une liste de jetons à l'aide des 30 mots les plus fréquents de ces cinq articles. Cette méthode sera représentée par l'acronyme **docmaj**.

On aboutit ainsi à trois champs lexicaux liés au mot-clé de la requête qui seront utilisés pour procéder à la fouille de documents. Le premier est indépendant des thèmes LDA et servira de référence de comparaison. Le second utilise directement les jetons extraits par le modèle LDA à 10 thèmes. Le troisième se sert indirectement des thèmes extraits par le modèle LDA à 10 thèmes en s'appuyant sur le contenu de certains documents du corpus. Le choix d'un modèle LDA à 10 thèmes pour les méthodes 2 et 3 a été fait suite aux résultats des mesures de cohérence.

Approche 2

La seconde approche utilise directement l'espace latent du modèle LDA pour procéder à la fouille. Les modèles LDA calculent en effet une représentation vectorielle pour chaque article du corpus. Chaque article est ainsi représenté à l'aide d'un vecteur dont la dimension est égale au nombre de thèmes du modèle et dont les composantes correspondent au poids affecté à chaque thème par le modèle. Le principe de cette seconde approche est de considérer la requête de l'utilisateur comme un document. Autrement dit, on applique un modèle LDA à la requête afin d'obtenir un vecteur contenant les poids de chaque thème pour la requête. Ce vecteur est ensuite utilisé pour évaluer la pertinence des articles de la revue. Les modèles LDA à 10, 30 et 50 thèmes ont été utilisés, dans le but d'évaluer l'influence du nombre de thèmes sur la pertinence des résultats avec cette approche.

Les méthodes de cette approche seront représentées par les acronymes **Latent LDA 10**, **Latent LDA 30** et **Latent LDA 50.**

5.1.3 Mesure de similarité

Chacune des approches de conversion de la requête nécessite une mesure de similarité spécifique.

Pour l'approche 1, un compte de fréquence a été utilisé comme indicateur de proximité sémantique. En fonction du champ lexical utilisé, on compte pour chaque article de la revue que l'on fouille la fréquence de l'ensemble des jetons composant le champ lexical. Les articles les plus pertinents sont ceux ayant le total d'occurrences le plus élevé parmi les jetons du champ lexical.

Pour l'approche 2, on utilise une mesure de similarité cosinus comme indicateur de proximité sémantique. La similarité cosinus mesure l'angle entre deux vecteurs. Le principe est de comparer le vecteur de chaque article dans l'espace latent LDA avec celui de la requête. Les articles les plus pertinents sont ceux dont le vecteur a la plus petite distance cosinus avec le vecteur requête.

5.1.4 Présentation des résultats

Pour des questions de limitation de temps et de ressources, il a été décidé de ne conserver que les titres des articles estimés comme les plus pertinents lors de la présentation des résultats. Ce choix repose sur la capacité humaine d'extrapolation : à partir de peu d'information, ils sont capables d'en déduire des propriétés plus générales. Finalement, la Figure 5.3 résume les étapes de la tâche de fouille de documents.

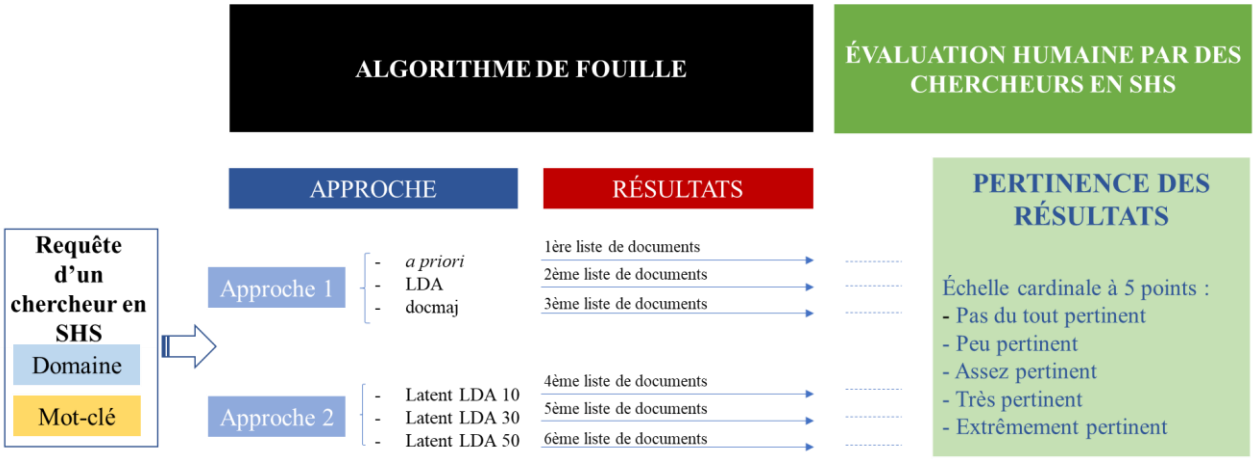

Figure 5.3 Schéma récapitulatif de la tâche de fouille de documents

En résumé, trois requêtes ont été effectuées sur chaque revue. Chacune des requêtes correspond à une thématique de recherche, plus ou moins spécifique en fonction de sa granularité. Deux approches de fouille ont été utilisées pour fouiller chaque revue. Chaque approche de fouille comporte trois méthodes. Trois champs lexicaux distincts sont utilisés par l'approche 1, ce qui donne trois listes de titres d'articles différentes : la liste « *a priori* », la liste « LDA » et la liste « docmaj ». Trois nombres de thèmes distincts pour les modèles LDA ont été utilisés par l'approche 2, ce qui donne également trois listes des titres d'articles différentes : la liste « Latent LDA 10 », la liste « Latent LDA 20 » et la liste « Latent LDA 50 ». Au total, pour chaque revue et chaque requête, on dispose donc de six listes de titres d'articles dont on souhaite faire évaluer la pertinence par des chercheurs en SHS.

5.2 Construction du questionnaire d'évaluation

Une fois obtenues les six listes de titres d'articles pour chaque mot-clé de recherche, on souhaite répondre aux deux objectifs de l'évaluation externe. Pour cela, le choix a été fait de construire un questionnaire pour faire évaluer par des experts en SHS la pertinence des résultats proposés par les différents algorithmes de fouille.

5.2.1 Répondants ciblés

Afin d'évaluer les résultats des algorithmes de fouille de documents pour chaque revue, il est nécessaire d'adresser le questionnaire à des humains ayant une expertise spécifique à chacune des revues. Trois universités québécoises ont été ciblées comme sources de répondants : HEC Montréal, l'Université du Québec à Montréal (UQAM) et l'Université de Sherbrooke. Ces trois universités comportent des départements de SHS et fournissent l'adresse courriel de leurs professeurs et chercheurs.

Ensuite, pour chacune des revues ont été identifiés les chercheurs et professeurs ayant une expertise proche des thématiques abordées au sein de la revue. Ainsi, pour la revue Actualité Économique ont été sélectionnés les professeurs des départements d'« Économie Appliquée et de Finance » du HEC Montréal et de l'Université de Sherbrooke et les professeurs avec une expertise en « Économie » de l'UQAM. Pour la revue Études Internationales, ont été sélectionnés les professeurs du département d'« Affaires Internationales » du HEC Montréal, les professeurs avec une expertise en « International » de l'UQAM (droit, relations, stratégie, etc.) et les professeurs du département de

« Politique Appliquée » de l'Université de Sherbrooke. Enfin, pour la revue Relations Industrielles ont été sélectionnés les professeurs du département de « Gestion des Ressources Humaines » du HEC Montréal et de l'Université de Sherbrooke et les professeurs avec une expertise en « Relations Industrielles » de l'UQAM. Les membres du comité de rédaction de chaque revue ont également été joints à la liste de cibles. Cet exercice a permis d'identifier 332 universitaires comme potentiels répondants au questionnaire.

5.2.2 Construction et administration du questionnaire

Le questionnaire se structure en trois parties. Une première page présente les enjeux de la recherche et ce qui est attendu du répondant. Puis, le répondant est invité à choisir l'expertise qui lui correspond le plus entre « Économie », « Relations Internationales » et « Relations Industrielles ». Enfin lui sont proposées les listes de titres d'articles à évaluer pour les trois mots-clés associés à la revue correspondante à son choix d'expertise. Les questions sont de la forme « Veuillez indiquer si la liste de ces 5 titres d'articles publiés dans la revue X vous semble pertinente par rapport au mot-clé suivant : "Y" ». Chaque évaluateur doit répondre à 18 questions pour compléter le questionnaire, ce qui prend entre 5 et 10 minutes. Le questionnaire est rédigé uniquement en français car les modèles LDA ont été entraînés sur des revues en langue française.

Pour mener l'évaluation, il faut enfin choisir une échelle de pertinence. Manning *et al.* (2009) mentionnent deux types d'échelles de pertinence utilisées pour valider les résultats de systèmes de recherche d'information. Une échelle de pertinence binaire (pertinent/non pertinent) permet d'appliquer des mesures interjuges facilement. Cependant, elle limite le pouvoir de discrimination des résultats obtenus. Une échelle de pertinence cardinale avec plus de 2 points augmente les possibilités d'analyse des résultats du questionnaire. Nous avons donc choisi une échelle cardinale à cinq points : « Pas du tout pertinent — Peu pertinent — Assez pertinent — Très pertinent — Extrêmement pertinent », inspirée de l'échelle développée par l'INEX pour l'évaluation de contenu XML (Pehcevski *et al.*, 2005).

Le logiciel LimeSurvey™ [11] a été utilisé pour construire et administrer le questionnaire. Il permet d'assurer l'anonymat des répondants et la confidentialité des données récupérées. Un certificat de conformité a été obtenu auprès du Comité d'éthique de la recherche avec des êtres humains de Polytechnique Montréal avant d'administrer le questionnaire. Un prétest a été effectué auprès de membres du laboratoire afin de valider la forme du questionnaire et la compréhension des questions. Le questionnaire a ensuite été envoyé une première fois par mail accompagné d'une lettre explicative, suivie d'un rappel 15 jours plus tard. La période d'administration du questionnaire s'est ainsi déroulée sur 1 mois.

5.3 Résultats et analyses statistiques

Le Tableau 5.2 présente les résultats de participation au sondage. 36 experts ont répondu, ce qui donne un taux de participation de 11,5 %.

Tableau 5.2 Statistiques de participation au questionnaire

Revue	Nombre total de cibles	Nombre total de répondants	Pourcentage de participation
ACTUALITÉ ÉCONOMIQUE	172	17	
ÉTUDES INTERNATIONALES	86	7	
RELATIONS INDUSTRIELLES	74	12	
TOTAL	332 (312)	36	11,5%

Le nombre en parenthèses sur la ligne de TOTAL correspond au nombre réel de cibles, une fois retirées les adresses courriel défectueuses. C'est le nombre utilisé pour le calcul du pourcentage de participation.

Chaque point de l'échelle de pertinence a été recodé par un chiffre entre 1 (« pas du tout pertinent ») et 5 (extrêmement pertinent) pour les analyses statistiques. Une réponse incomplète pour la revue AE a été retirée.

[11] Limesurvey GmbH. / LimeSurvey: An Open Source survey tool /LimeSurvey GmbH, Hamburg, Germany. URL http://www.limesurvey.org

Étude de la fidélité interjuges

Une analyse préliminaire consiste à évaluer la fidélité interjuges des résultats. Le but de cette analyse est d'évaluer la probabilité que les évaluateurs (les « juges ») se soient mis d'accord par chance. On calcule pour cela le coefficient α de Krippendorff présenté dans la littérature. Les résultats d'accord interjuges pour les trois revues sont donnés au Tableau 5.3. Rappelons que le seuil d'acceptation en recherche en sciences sociales se situe à 0.67 (Krippendorff, 2004). Seule la revue RI obtient un α supérieur au seuil recommandé avec une valeur de 0.69. Les faibles valeurs d'accord obtenues peuvent s'expliquer d'une part par la multiplicité des points de l'échelle et d'autre part par la subjectivité de la tâche à évaluer. En effet, la notion de pertinence dans le cas de l'évaluation de résultats d'une fouille de documents est connue pour sa subjectivité (Manning *et al.*, 2009). De plus, le fait que l'on ne propose que les titres d'articles pour évaluer le lien avec le mot-clé laisse plus de place à l'interprétation personnelle de l'évaluateur. Il serait intéressant de reproduire l'étude avec une échelle binaire pour vérifier la fiabilité interjuges sur un nombre plus restreint de choix. Enfin, bien qu'ayant une connaissance générale de leur domaine, les juges ont chacun une expertise spécifique qui peut engendrer un désaccord sur la compréhension des mots-clés recherchés. Ce désaccord souligne par ailleurs la difficulté de construire un moteur de recherche qui satisfait tous les utilisateurs, problème fréquemment observé dans les études utilisateurs (Manning *et al.*, 2009).

Tableau 5.3 Résultats d'accord interjuges

COEFFICIENT	ACTUALITÉ ÉCONOMIQUE	ÉTUDES INTERNATIONALES	RELATIONS INDUSTRIELLES
α de Krippendorff	0,45	0,56	**0,69**

Pour la suite des analyses, chaque revue est identifiée par son domaine de recherche. Cette hypothèse permet d'étudier la différence d'influence de la thématique générale de recherche abordée dans une revue (Économie, Relations Internationales, Relations Industrielles) sur la pertinence des résultats proposés par le modèle thématique LDA.

Trois études statistiques ont été menées sur les résultats obtenus afin de répondre aux quatre objectifs spécifiques.

i. Mesure de l'influence du domaine de la revue sur la pertinence moyenne des résultats (objectif spécifique (3)).

ii. Mesure de l'influence des algorithmes de fouille sur la pertinence moyenne des résultats (objectifs spécifiques (1) et (2)).

iii. Mesurer l'influence de la spécificité de la requête sur la pertinence moyenne des résultats (objectif spécifique (4)).

Pour les deux premières analyses, le processus de moyennisation s'effectue sur les notes de pertinence données par tous les juges d'un même domaine, normalisé sur l'ensemble des trois mots-clés. Pour la dernière analyse, il s'effectue sur les notes de pertinence données par tous les juges d'un même domaine pour chaque mot-clé de requête.

Le Tableau 5.4 résume le cadre des trois analyses. Pour les trois études, les résultats de pertinence moyenne supérieurs à 3/5 seront marqués en gras. Aucun seuil de pertinence n'est fourni par la littérature donc le choix de 3 a été fait spécifiquement pour l'étude. Il faut noter que la sélectivité de celui-ci dépendra de l'exigence fixée pour le moteur de recherche. Dans le cadre de ce travail exploratoire, le but n'est pas de construire un moteur de recherche compétitif mais plutôt de déterminer les paramètres qui influent sur la pertinence des résultats obtenus.

Tableau 5.4 Résumé des analyses statistiques

	i. Influence du domaine	**ii. Influence de la méthode**	**iii. Influence du mot-clé**
Variable comparée	Pertinence moyenne par domaine, pour les 3 mots-clés utilisés par une méthode		Pertinence moyenne par méthode
Méthode de comparaison	Pour chaque méthode, comparaison entre domaines 2 à 2	Pour chaque domaine, comparaison entre méthodes 2 à 2	Pour chaque méthode, comparaison entre mots-clés 2 à 2
Test non-paramétrique (p-valeur)	Test de Kruskal-Wallis et Test de Mann-Whitney (échantillons indépendants car évaluateurs différents)	Test de Wilcoxon (échantillons liés car mêmes évaluateurs)	

i. Influence du domaine de la revue sur la pertinence des résultats

Le but de cette première analyse est de déterminer la pertinence moyenne obtenue pour chaque domaine à l'aide des différentes méthodes de fouille. Il s'agit en particulier de déterminer si le modèle LDA est plus pertinent pour une des revues sur cette tâche de fouille de documents. Deux

tests non paramétriques ont été menés pour étudier la significativité des résultats : le test de Kruskal-Wallis et le test de Mann-Whitney. Le test de Kruskal-Wallis s'applique à l'ensemble des groupes indépendants. Son hypothèse nulle est que les médianes des différents groupes sont égales. Elle est rejetée si au moins une des médianes est différente des autres pour une p-valeur inférieure à 0.05. Dans le cas où elle est rejetée, on peut appliquer un test de Mann-Whitney pour mesurer les différences de médianes entre les groupes deux à deux. Ces deux tests sont valables pour des échantillons de petite taille (N<30), ce qui est le cas de la population étudiée.

Le Tableau 5.5 présente les résultats obtenus. Chaque ligne présente le résultat de pertinence moyenné sur les trois mots-clés pour chaque revue.

Le test de Kruskal-Wallis montre une différence inter-domaine globale fortement significative pour les méthodes de l'approche 1 ($p < 0{,}001$).La première ligne du tableau donne les résultats de pertinence moyenne obtenus par la méthode *a priori* sur les trois revues. On voit que cette méthode donne des résultats significativement meilleurs sur la revue AE que sur les deux autres. La même différence est obtenue pour les méthodes LDA et docmaj. Par conséquent, les méthodes utilisant un champ lexical sont fortement dépendantes du domaine de la revue, ce qui pourrait s'expliquer par le fait qu'elle sont construites autour de la thématique recherchée. En revanche, il n'y a pas de différence observée entre EI et RI. On peut penser que les thématiques développées dans la revue AE sont significativement différentes de celles étudiées dans les revues EI et RI.

Au contraire, les méthodes de l'approche 2 ressortent indépendantes du domaine, mise à part la méthode Latent LDA 30 qui montre une différence de résultat significative entre les revues AE et RI (ligne 5). Cette indépendance signifie que les thématiques qui composent les articles d'une revue n'influent pas sur l'algorithme de fouille utilisant la représentation latente calculée par le modèle LDA. Par conséquent, utiliser directement l'espace latent du LDA permet de s'affranchir d'un effet de domaine sur les résultats de pertinence en proposant des résultats similaires pour les différentes revues.

On souhaite maintenant de déterminer si les résultats obtenus par l'approche 2 sont significativement plus pertinents que ceux obtenus par l'approche 1 pour les trois revues.

Tableau 5.5 Influence du domaine sur la pertinence.

		Actualité Économique (AE)	**Études Internationales (EI)**	**Relations Industrielles (RI)**	Moyenne sur les 3 revues	Test de Kruskal Wallis	Test de Mann Whitney AE/EI	Test de Mann Whitney AE/RI	Test de Mann Whitney EI/RI
APPROCHE	**MÉTHODE**	**PERTINENCE MOYENNE**				**P-VALEUR**			
1	apriori	**3,05**	1,43	1,64	2,24	****	****	****	Ø
	LDA	2,64	1,43	1,31	1,94	****	***	****	Ø
	docmaj	2,51	1,33	1,22	1,83	****	****	****	Ø
2	Latent LDA 10	2,83	2,67	**3,00**	2,86	Ø	Ø	Ø	Ø
	Latent LDA 30	2,71	**3,62**	**3,33**	**3,10**	**	Ø	***	Ø
	Latent LDA 50	**3,19**	2,71	**3,25**	**3,11**	Ø	Ø	Ø	Ø

Le tableau se lit ligne par ligne : on compare chaque méthode de fouille indépendamment, sur les trois revues. Les résultats de P-valeur indiquent la significativité des différences observées entre les revues 2 à 2.

Légende du Tableau 5.5 :

P-valeur	Code
p > 0,1	Ø
p < 0,1	*
p < 0,05	**
p < 0,01	***
p < 0,001	****

ii. Influence de la méthode de fouille sur la pertinence

Le but de cette deuxième analyse statistique est de comparer les algorithmes de fouille de document pour chaque revue. Il s'agit de déterminer s'il existe des différences significatives entre les approches de fouille 1 et 2 concernant la pertinence moyenne des résultats. Au sein de chaque approche, l'analyse permettra de déterminer quelle méthode donne les meilleurs résultats. Pour mesurer la significativité des résultats, le test des rangs signés de Wilcoxon a été réalisé (Kruskal, 1957). C'est le test équivalent au test de Student dans le cas d'échantillons appariés. L'hypothèse nulle est que les médianes des différents groupes sont les mêmes. Cette hypothèse sera rejetée si la p-valeur obtenue est inférieure à 0,05.

Le Tableau 5.6 présente les résultats obtenus. Il se lit ligne par ligne, de gauche à droite. Un « + » (respectivement un « − ») est marqué si la pertinence moyenne obtenue par la méthode inscrite sur la ligne est supérieure (respectivement inférieure) à la pertinence moyenne obtenue par la méthode inscrite sur la colonne. Le nombre de « + » et de « − » dépend de la significativité de la différence. Par exemple, pour la revue AE, à l'intersection de la ligne « apriori » et de la colonne « Latent LDA 30 », on lit un « +++ » car la différence de pertinence moyenne entre les deux méthodes est significative avec une p-valeur inférieure à 0,01 et que la pertinence moyenne de la méthode « apriori » est supérieure à celle de la méthode « Latent LDA 30 ». Le symbole Ø est utilisé pour indiquer une absence de significativité.

Dans l'ensemble, les résultats montrent que les algorithmes les plus performants de façon significative sont les méthodes de l'approche 2 utilisant l'espace latent du LDA. C'est le cas de la méthode Latent LDA 50 pour AE (score de 3,19), Latent LDA 30 pour EI (score de 3,62) et des trois méthodes latentes pour RI (scores de 3,00 ; 3,33 et 3,25). Une exception est à noter pour la revue AE où la méthode apriori montre un score de pertinence moyen de 3,05, qui est meilleur ou équivalent aux scores obtenues par les algorithmes LDA. Pour la revue EI (respectivement RI), les méthodes de l'approche 1 produisent des résultats significativement moins pertinents que les méthodes de l'approche 2 pour un seuil $p < 0.05$ (respectivement $p < 0.01$). Pour la revue AE, les méthodes apriori et Latent LDA 50 donnent les meilleurs résultats de pertinence, sans différence significative entre elles.

Parmi les trois méthodes de l'approche 1, le champ lexical apriori est meilleur que les deux autres pour les revues AE et RI, mais équivalent pour la revue EI. L'apport des modèles LDA n'est donc

pas intéressant dans cette tâche si l'on n'utilise qu'un compte de fréquence des termes représentatifs extraits par le modèle pour chaque thème. Utiliser un champ lexical externe lié au mot-clé (méthode *a priori)* donne des articles considérés comme plus pertinents par les experts en SHS. De plus, les méthodes LDA et docmaj donnent des résultats non significativement différents pour les trois revues. Cette absence de différence indique que les jetons représentatifs des thèmes LDA et les termes les plus fréquents des articles ayant un poids élevé pour ces thèmes LDA apportent une information similaire pour la fouille de documents. Cependant, cette information ne permet pas de proposer des résultats pertinents.

L'effet du nombre de thèmes a priori du modèle LDA sur les résultats (objectif spécifique (3)) se lit à l'aide des résultats des méthodes de l'approche 2. L'effet du nombre de thèmes sur la pertinence n'est pas évident. Il apparaît que le modèle à 10 thèmes n'est meilleur pour aucune des trois revues. Pour AE, c'est le modèle avec 50 thèmes qui donne les meilleurs résultats de façon significative par rapport aux modèles à 10 ($p < 0,1$) et 30 thèmes ($p < 0,01$). Pour EI, c'est le modèle à 30 thèmes qui est meilleur que les modèles à 10 et 50 thèmes (pour un $p < 0,05$). Pour RI, les modèles à 30 et 50 thèmes donnent les meilleurs résultats, sans différence significative entre eux. Autrement dit, le nombre de thèmes du modèle LDA n'est pas un critère suffisant pour assurer la qualité de la méthode de fouille.

Tableau 5.6 Influence de la méthode de fouille sur la pertinence

ACTUALITÉ ÉCONOMIQUE								
APPROCHE	**MÉTHODE**		apriori	LDA	docmaj	Latent LDA 10	Latent LDA 30	Latent LDA 50
		Pertinence moyenne	**3,05**	2,64	2,51	2,83	2,71	**3,19**
1	a priori	**3,05**		++	+++	Ø	+++	Ø
	LDA	2,64			Ø	– –	Ø	– –
	docmaj	2,51				– –	Ø	– – –
2	Latent LDA 10	2,83					Ø	–
	Latent LDA 30	2,71						– – –
	Latent LDA 50	**3,19**						
ÉTUDES INTERNATIONALES								
APPROCHE	**MÉTHODE**		apriori	LDA	docmaj	Latent LDA 10	Latent LDA 30	Latent LDA 50
		Pertinence moyenne	1,43	1,43	1,33	2,67	**3,62**	2,71
1	a priori	1,43		Ø	Ø	– –	– –	– –
	LDA	1,43			Ø	– –	– –	– –
	docmaj	1,33				– –	– –	– –
2	Latent LDA 10	2,67					– –	Ø
	Latent LDA 30	**3,62**						++
	Latent LDA 50	2,71						
RELATIONS INDUSTRIELLES								
APPROCHE	**MÉTHODE**		apriori	LDA	docmaj	Latent LDA 10	Latent LDA 30	Latent LDA 50
		Pertinence moyenne	1,64	1,31	1,22	**3,00**	**3,33**	**3,25**
1	apriori	1,64		+++	++	– – –	– – –	– – –
	LDA	1,31			Ø	– – –	– – –	– – –
	docmaj	1,22				– – –	– – –	– – –
2	Latent LDA 10	**3,00**					– –	Ø
	Latent LDA 30	**3,33**						Ø
	Latent LDA 50	**3,25**						

Tableau 5.6 Influence de la méthode de fouille sur la pertinence (suite et fin)

Légende du Tableau 5.6 :

Ligne > Colonne	P-valeur	Ligne < Colonne
Ø	p > 0,1	Ø
+	p < 0,1	–
++	p < 0,05	– –
+++	p < 0,01	– – –
++++	p < 0,001	– – – –

iii. Influence du mot-clé sur la pertinence

Le but de cette dernière analyse est de déterminer s'il existe un effet de la granularité des mots-clés utilisés dans la recherche. Autrement dit, il s'agit d'étudier l'influence de la requête sur les résultats obtenus. Un moteur de recherche idéal propose des résultats pertinents indépendamment de la requête de l'utilisateur. Cependant, les modèles LDA capturent des propriétés générales du corpus, donc on peut s'attendre à ce qu'ils performent mieux sur les recherches de thématiques plus générales. En particulier, la représentation thématique calculée par le LDA pourrait s'avérer trop grossière pour capturer la spécificité d'une requête portant sur un sujet peu abordé dans la collection.

Le

Tableau 5.7 présente les résultats obtenus. Les mots-clés sont coloriés en fonction de leur niveau de granularité selon la classification de l'ANZSRC. Il se lit encore une fois ligne par ligne, et permet de comparer le résultat obtenu par chacune des méthode entre les 3 requêtes effectuées. Le test non paramétrique de Wilcoxon a été mené sur les paires de mots-clés de chaque domaine pour étudier les différences de résultats de pertinence.

Le premier mot-clé avec une granularité de type « Groupe » est le terme « Macroéconomie » pour la revue AE. Les deux autres mots-clés testés pour cette revue ont une granularité de type « Domaine » donc sont plus spécifiques. Le mot-clé avec une granularité « Groupe » donne des résultats significativement plus pertinents que les deux autres pour la plupart des méthodes de fouille. Les scores de pertinence obtenus avec ce mot-clé sont supérieurs à 3 pour les six méthodes. Pour les méthodes de l'approche 1, il donne de meilleurs résultats que le mot-clé 2 sur toutes les configurations et de meilleurs résultats que le mot-clé 3 sur deux des trois configurations. Pour les méthodes de l'approche 2, il obtient des résultats similaires au mot-clé 2. En revanche, le mot-clé 2 performe bien mieux que le mot-clé 3 sur les modèles latents, malgré leur granularité identique.

L'autre mot-clé de type « Groupe » est le mot « Travail » testé pour la revue RI. Les deux autres mots-clés testés pour cette revue ont une granularité de type « Sous-domaine », donc sont très spécifiques. L'effet de la granularité est moins marqué sur cette revue. Le mot-clé « Travail » se montre significativement plus pertinent que les deux autres sur quatre méthodes sur six. Cependant, les résultats de pertinence obtenus sont globalement faibles, avec seulement deux méthodes sur six qui donnent un score supérieur à 3. D'autre part, les deux mots-clés de granularité « Sous-domaine » ne montrent qu'une seule différence significative sur les six méthodes (sur la méthode Latent LDA

50). Cela indique que des mots-clés avec une granularité très spécifique donneront des résultats similaires indépendamment de la méthode de fouille. Ce résultat se confirme sur la revue EI. En effet, les trois mots-clés testés sur la revue EI sont de type « Sous-domaine » et seul un des dix-huit tests de Wilcoxon ressort significatif.

Enfin, pour les trois revues, l'approche 2 confirme obtenir de meilleurs résultats de pertinence pour la majorité des mots-clés testés. En effet, 15 des 19 scores de pertinence supérieurs à 3 sont obtenus à l'aide des méthodes de l'approche 2 et les 4 scores restants supérieurs à 3 ont été obtenus sur la revue AE.

Tableau 5.7 Influence du mot-clé sur la pertinence

		ACTUALITÉ ÉCONOMIQUE					
		Mot-clé 1	**Mot-clé 2**	**Mot-clé 3**	**Test de Wilcoxon**		
		Macroéconomie	Économie Internationale	Économie Monétaire et Financière	Mot-clé 1 Mot-clé 2	Mot-clé 1 Mot-clé 3	Mot-clé 2 Mot-clé 3
APPROCHE	**MÉTHODE**	**PERTINENCE MOYENNE**			**P-VALEUR**		
1	a priori	**3,41**	2,75	**3,00**	**	Ø	Ø
	LDA	**3,65**	1,94	2,40	***	***	*
	docmaj	**3,53**	1,38	2,73	****	**	***
2	Latent LDA 10	**3,65**	**3,25**	1,60	Ø	****	***
	Latent LDA 30	**3,00**	**3,50**	1,60	Ø	***	***
	Latent LDA 50	**4,06**	**3,50**	1,93	*	****	***

Légende du Tableau 5.7:

P-valeur	Code
p > 0,1	Ø
p < 0,1	*
p < 0,05	**
p < 0,01	***
p < 0,001	****

Niveaux de granularité

Groupe

Domaine

Sous-domaine

Tableau 5.7 Influence du mot-clé sur la pertinence (suite et fin)

		ÉTUDES INTERNATIONALES					
		Mot-clé 1	**Mot-clé 2**	**Mot-clé 3**	**Test de Wilcoxon**		
		Démocratisation	Dissuasion	Sécurité Internationale	Mot-clé 1 Mot-clé 2	Mot-clé 1 Mot-clé 3	Mot-clé 2 Mot-clé 3
APPROCHE	**MÉTHODE**	**PERTINENCE MOYENNE**			**P-VALEUR**		
1	a priori	1,58	1,33	2,00	**	Ø	Ø
	LDA	1,17	1,08	1,67	Ø	Ø	Ø
	docmaj	1,17	1,08	1,42	Ø	Ø	Ø
2	Latent LDA 10	2,83	2,75	**3,42**	Ø	Ø	Ø
	Latent LDA 30	**3,42**	**3,67**	2,92	Ø	Ø	Ø
	Latent LDA 50	**4,00**	2,00	**3,75**	Ø	Ø	Ø
		RELATIONS INDUSTRIELLES					
		Mot-clé 1	**Mot-clé 2**	**Mot-clé 3**	**Test de Wilcoxon**		
		Convention Collective	Théories de la justice et de l'équité	Travail	Mot-clé 1 Mot-clé 2	Mot-clé 1 Mot-clé 3	Mot-clé 2 Mot-clé 3
APPROCHE	**MÉTHODE**	**PERTINENCE MOYENNE**			**P-VALEUR**		
1	a priori	1,71	1,14	1,43	Ø	Ø	**
	LDA	1,71	1,29	1,29	Ø	**	**
	docmaj	1,57	1,14	1,29	Ø	Ø	**
2	Latent LDA 10	2,43	2,57	**3,00**	Ø	*	Ø
	Latent LDA 30	**3,57**	**3,71**	**3,57**	Ø	*	**
	Latent LDA 50	2,57	2,86	2,71	***	Ø	***

CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS

Les expériences menées dans ce travail ont été motivées par la volonté d'améliorer l'accessibilité de la modélisation thématique de documents auprès de chercheurs en SHS et ainsi de contribuer à la valorisation de collections croissantes de documents textuels numérisés.

Pour cela, le travail a consisté à implémenter, visualiser et évaluer des modèles LDA sur trois revues de sciences humaines et sociales.

En premier lieu ont été exposées et expliquées différentes méthodes de visualisation qui permettent aux chercheurs en SHS d'appréhender les thèmes extraits par les modèles. La représentation sous forme de nuage de mots a pour avantage de proposer un portrait général et concis des thèmes extraits. Cependant, elle ne donne aucun moyen de comparer les différents thèmes extraits. La lemmatisation n'a eu aucun effet sur ce poids moyen. L'étude de la variabilité des poids moyens a en outre montré que les modèles répartissent de façon homogène les poids sur chaque groupe de mots. L'interface PyLDAvis permet d'explorer de façon interactive les termes extraits pour chaque thème et l'importance relative de chaque thème au sein du corpus. L'algorithme t-SNE s'est montré plus performant pour discriminer les différents thèmes sur une projection 2D. Enfin, les documents d'une revue ont pu être regroupés sur une projection 2D en appliquant un algorithme t-SNE sur leur vecteur de thèmes, permettant une exploration thématique de la collection.

Puis, une première évaluation s'est basée sur les mesures de perplexité et de cohérence. Elle a permis de quantifier l'influence du nombre de thèmes des modèles et l'effet de la lemmatisation du vocabulaire Les résultats de perplexité ne corrèlent pas avec l'interprétabilité des thèmes. Pour les trois revues, on constate une tendance décroissante de la perplexité avec le nombre de thèmes. La lemmatisation du vocabulaire n'a aucun effet significatif sur la mesure de perplexité. Les résultats de cohérence montrent qu'un nombre de thèmes supérieur à 10 diminue la cohérence des modèles LDA. De plus, les valeurs de cohérence se stabilisent à partir d'un nombre de thèmes égal à 50 pour les mesures C_{UCI}, C_{UMass} et C_{NPMI} et égal à 10 pour la mesure C_V. Les quatre courbes obtenues pour les revues EI et RI suivent des tendances proches, avec des valeurs légèrement supérieures à celles obtenues pour la revue AE. Enfin, une étude de fiabilité originale a été proposée pour les modèles LDA. Les mesures de perplexité et de cohérence montrent une faible variabilité sur des modèles ayant un même nombre de thèmes. À partir de la création d'un score de similarité

sémantique globale permettant de comparer des thèmes LDA deux à deux, nous avons montré une fiabilité qualitative des modèles LDA mais qui reste sujette au caractère probabiliste des modèles.

Ces deux étapes décrites dans le chapitre 4 ont permis de répondre à la première question de recherche.

Afin de valider l'utilité opérationnelle des modèles LDA, une évaluation externe des modèles a été conduite avec une tâche de fouille de documents et un questionnaire envoyé à des chercheurs en SHS. L'accord interjuges et les résultats de pertinence obtenus sont globalement faibles : ceci s'explique probablement par le fait que l'on ne présente que les titres des articles à l'évaluation, que la généralité des thématiques recherchées favorise la subjectivité des notes de pertinence, que la dispersion de l'échelle tend à éliminer les choix très pertinents et que les sujets ciblés peuvent avoir des expertises différentes malgré leur appartenance à des départements d'étude similaires. La première analyse statistique démontre un avantage de fiabilité de l'approche 2 sur l'approche 1, qui est de donner des résultats de pertinence similaires indépendamment de l'expertise de recherche. L'approche 1 utilisant des champs lexicaux a donné des résultats de pertinence particulièrement faibles pour les revues EI et RI. La deuxième analyse démontre que les méthodes de l'approche 2 s'appuyant directement sur l'espace latent des modèles LDA obtiennent des résultats de pertinence significativement meilleurs pour les trois revues. Seule la méthode apriori appliquée à la revue AE a obtenu des résultats de pertinence supérieurs à 3. L'influence des thèmes pour l'approche 2 n'est pas évidente : les modèles à 30 et 50 thèmes se montrent légèrement meilleurs que le modèle à 10 thèmes. La troisième analyse montre que la granularité des mots-clés participe à la pertinence des résultats mais que son effet dépend de la revue. Pour AE, la thématique générale donne de meilleurs résultats que les deux autres thématiques plus spécifiques sur la plupart des méthodes de fouille. Pour les autres revues, les thématiques ayant un même niveau de spécificité donnent des résultats de pertinence similaires. Cette seconde étape, décrite dans le chapitre 5, a permis de répondre à la seconde question de recherche.

Les contributions du mémoires sont donc méthodologiques. Nous proposons d'abord une méthode d'évaluation de la validité et de la fiabilité des modèles LDA à l'aide de mesures d'évaluation automatique de la littérature. Puis, nous proposons une méthode de recherche d'information spécialisée, basée sur le LDA et évaluée par des experts sur un corpus d'articles en sciences humaines et sociales. Ces premiers résultats indiquent que le LDA peut être utilisé dans une tâche

de recherche de documents dans la mesure où l'on utilise directement l'espace latent pour augmenter la requête.

Cette recherche comporte des limites théoriques et expérimentales, qui découlent de son caractère exploratoire. En effet le nombre de documents composant le corpus et de répondants au questionnaire est limité, donc les méthodes proposées doivent être testées sur d'autres données et être évaluées par d'autres chercheurs pour confirmer la capacité de généralisation des résultats obtenus. En particulier, on pourrait s'appuyer sur des ontologies extérieures pour analyser et étiqueter les thèmes extraits par le modèle LDA (mode d'entraînement semi-supervisé). L'évaluation s'est faite uniquement sur des titres ce qui enlève une partie importante de l'information sémantique des articles qui permettrait aux juges de préciser leur jugement concernant l'adéquation avec chacune des requêtes.

Le travail de cette maîtrise appelle donc à être étendu par différentes possibilités. À l'occasion d'une nouvelle évaluation externe, on pourrait comparer d'autres algorithmes avec le LDA, comme une extraction par n-grammes, l'algorithme des K-moyennes ou bien des variantes récentes du LDA comme le lda2vec qui incorpore une représentation vectorielle des mots au modèle LDA (Moody, 2016). Ensuite, le modèle de recherche de documents et de comparaison des requêtes peut être affiné. En effet, les scores de fidélité interjuges et de pertinence obtenus sont dans l'ensemble assez faibles. Une piste d'amélioration serait d'intégrer des métadonnées à l'algorithme de fouille. Le système pourrait aussi ajuster ses résultats en fonction des choix de l'utilisateur plutôt que de fonctionner sur un mode « statique » de fouille (Deveaud, 2014). Pour cela, il faudrait créer une interface visuelle pour que des chercheurs en SHS puissent utiliser intuitivement l'algorithme de recherche sur leurs propres données. Enfin, il serait intéressant d'évaluer les méthodes de fouille proposées dans ce mémoire sur un corpus plus standard comme le TREC.

Par ses contributions, ce travail de ce mémoire participe à valoriser des algorithmes d'analyse de textes auprès d'une communauté de chercheurs en SHS. Cette étape de valorisation est cruciale dans le développement d'outils fonctionnels robustes mis à disposition d'une communauté d'utilisateurs. Cependant, de nombreux défis subsistent pour faciliter la communication et la collaboration entre les experts en modélisation mathématique et les experts en sciences sociales. La définition commune de problématiques de recherche afin d'identifier les outils techniques

adéquats pour y répondre nous semble déterminante pour permettre aux humains d’appréhender le contenu numérique de demain.

BIBLIOGRAPHIE

Aggarwal, C. C. (2015). *Data Mining : the textbook*. Springer.

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Aggarwal, C. C., & Zhai, C. (2012). *A Survey of Text Classification Algorithms*.

Aletras, N., & Stevenson, M. (2013). *Evaluating Topic Coherence Using Distributional Semantics*. 9.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining : Classification, Clustering and Extraction Techniques. *arXiv:1707.02919* .

Angell, R. C., Freund, G. E., & Willett, P. (1983). Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, *19*(4), 255-261.

Anick, P. G., & Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. *ACM SIGIR Forum*, *31*, 314-323. ACM.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2007). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *ArXiv:0710.0845*

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference : A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877.

Blei, D. M., & Lafferty, J. (2009). *Topic models. Text mining : Theory and applications*. Taylor and Francis London.

Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic models*. Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control VII*, 235-255.

Boyd-graber, J. L., & Blei, D. M. (2009). *Syntactic Topic Models*. In Advances in neural information processing systems (pp. 185-192).

Buckley, C., & Voorhees, E. M. (2017). Evaluating evaluation measure stability. *ACM SIGIR Forum*, *51*, 235-242. ACM.

Callan, J. P., Lu, Z., & Croft, W. B. (2017). Searching distributed collections with inference networks. *ACM SIGIR Forum*, *51*, 160-167. ACM.

Carbonell, J. G., & Goldstein, J. (1998). *The Use of MMR and Diversity-Based Reranking for Reodering Documents and Producing Summaries*.

Carlo, C. M. (2004). Markov chain monte carlo and gibbs sampling. *Notes for EEB, 581*.

Chang, J., & Blei, D. M. (2010). *Relational Topic Models for Document Networks*. The Annals of Applied Statistics, 4(1), 124-150.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2010). *Reading Tea Leaves : How Humans Interpret Topic Models*. 10.

Christanti Mawardi, V., Susanto, N., & Santun Naga, D. (2018). Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method. *MATEC Web of Conferences*, *164*, 01047.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite : Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12*, 74.

Clark, A. S., Fox, C., & Lappin, S. (2010). *The handbook of computational linguistics and natural language processing*. Wiley Online Library.

Cutting, D. R., Karger, D. R., & Pedersen, J. O. (1993). Constant interaction-time scatter/gather browsing of very large document collections. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 126-134. ACM.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Knowledge discovery through directed probabilistic topic models : A survey. *Frontiers of Computer Science in China*, *4*(2), 280-301.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61-84.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dickey, J. M. (1983). Multiple hypergeometric functions : Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, *78*(383), 628-637.

Dietz, L., Bickel, S., & Scheffer, T. (2007). *Unsupervised prediction of citation influences*. 233-240.

Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication*, *27*(2), 121-140.

Evans, D. A., Ginther-Webster, K., Hart, M., Lefferts, R. G., & Monarch, I. A. (1991). Automatic indexing using selective NLP and first-order thesauri. *Intelligent Text and Image Handling-Volume 2*, 624-643. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 524-531. IEEE.

George E. P. Box. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, *115*(13), 3308-3313.

Gerrish, S., & Blei, D. M. (2011). Predicting legislative roll calls from text. *Proceedings of the 28th international conference on machine learning (icml-11)*, 489-496.

Gerrish, S. M., & Blei, D. M. (2010). *A Language-based Approach to Measuring Scholarly Impact. In ICML (Vol. 10, pp. 375-382).*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Supplement 1), 5228-5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544)..

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts : Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1-35.

Grimmer, J., & Stewart, B. M. (2013). Text as Data : The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(03), 267-297.

Gupta, G., & Malhotra, S. (2015). Text documents tokenization for word frequency count using rapid miner (taking resume as an example). *International Journal of Computer Applications*, *975*, 8887.

Hantler, S. L., Laker, M. M., Lenchner, J., & Milch, D. (2017). *Methods and apparatus for performing spelling corrections using one or more variant hash tables*. Google Patents.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77-89.

He, J., Hu, Z., Berg-Kirkpatrick, T., Huang, Y., & Xing, E. P. (2017). Efficient Correlated Topic Modeling with Topic Embedding. *arXiv:1707.00206*

Hillard, W. D., Purpura, S., & Wilkerson, J. (2007). *WITP Computer-Assisted Topic Classification for Mixed-Methods Social Science Research*.

Hoffman, M. D. (2013). *Stochastic Variational Inference*. *The Journal of Machine Learning Research*, *14*(1), 1303-1347..

Hoffman, M. D., & Blei, D. M. (2010). *Online Learning for Latent Dirichlet Allocation*. In advances in neural information processing systems (pp. 856-864)..

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1-2), 177-196.

Hopkins, D. J., & King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, *54*(1), 229-247.

Hopkins, D., & King, G. (2007). Extracting systematic social science meaning from text. *Manuscript available at http://gking. harvard. edu/files/words. pdf*, *20*(07).

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, *37*(2), 183-233.

Kataria, S., Mitra, P., Caragea, C., & Giles, C. L. (2011). *Context Sensitive Topic Models for Author Influence in Document Networks*. 7.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data : An introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Kent, A., Berry, M. M., Luehrs Jr, F. U., & Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American documentation*, *6*(2), 93-101.

Kim, J., Kim, D., & Oh, A. (2017). *Joint Modeling of Topics, Citations, and Topical Authority in Academic Corpora*. 14.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv:1312.6114*.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text Classification Algorithms : A Survey. *Information*, *10*(4), 150.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, *30*(3), 411-433.

Kruskal, W. H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association*, *52*(279), 356-360.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79-86.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine Reading Tea Leaves : Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

Li, L.-J., Wang, C., Lim, Y., Blei, D. M., & Fei-Fei, L. (2010). Building and using a semantivisual image hierarchy. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3336-3343. IEEE.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, *12*(3), 141-151.

Lu, K., Cai, X., Ajiferuke, I., & Wolfram, D. (2017). Vocabulary size and its effect on topic representation. *Information Processing & Management*, *53*(3), 653-665.

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.

Manning, C., Raghavan, P., & Schuetze, H. (2009). *Introduction to Information Retrieval*. In Proceedings of the international communication of association for computing machinery conference (p. 260)..

Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River.

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec : Learning generic context embedding with bidirectional lstm. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 51-61.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

Mimno, D., & Blei, D. (2011). Bayesian Checking for Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 227–237.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models*. 11.

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. *Annual review of computer science*, *4*(1), 417-433.

Mitra, M., Buckley, C., Singhal, A., & Cardie, C. (1997). An analysis of statistical and syntactic phrases. *Computer-Assisted Information Searching on Internet*, 200-214. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *arXiv:1605.02019*

Mori, S., Nishida, H., & Yamada, H. (1999). *Optical character recognition*. John Wiley & Sons, Inc.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The computer journal*, *26*(4), 354-359.

Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media.

Newman, D., Chemudugunta, C., & Smyth, P. (2006). Statistical entity-topic models. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 680-686. ACM.

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. *Proceedings of the 10th Annual Joint Conference on Digital Libraries - JCDL '10*, 215.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88-102.

Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). Identification of imminent suicide risk among young adults using text messages. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 413. ACM.

Ofoghi, B., & Verspoor, K. (2017). Textual Emotion Classification : An Interoperability Study on Cross-Genre Data Sets. *Australasian Joint Conference on Artificial Intelligence*, 262-273. Springer.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

Pehcevski, J., Thom, J. A., & Vercoustre, A.-M. (2005). Users and Assessors in the Context of INEX : Are Relevance Dimensions Relevant? *arXiv:cs/0507069*.

Pennington, J., Socher, R., & Manning, C. (2014). *Glove : Global Vectors for Word Representation*. 1532-1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365*

Ponte, J. M., & Croft, W. B. (1998). *A language modeling approach to information retrieval*. University of Massachusetts at Amherst.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2006). An automated method of topic-coding legislative speech over time with application to the 105th-108th US Senate. *Midwest Political Science Association Meeting*, 1-61.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). *Topic Modeling for the Social Sciences*. In NIPS 2009 workshop on applications for topic models: text and beyond (Vol. 5, p. 27).

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the Space of Topic Coherence Measures*. 399-408.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). *The Author-Topic Model for Authors and Documents*. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (pp. 487-494). AUAI Press..

Schofield, A., & Mimno, D. (2016). Comparing apples to apple : The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, *4*, 287-300.

Schrodt, P. A. (2000). Pattern Recognition of International Crises using Hidden Markov Models. *Hidden Markov Models*, 37.

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political science : KEDS—a program for the machine coding of event data. *Social Science Computer Review*, *12*(4), 561-587.

Sebastiani, F. (2002). Information retrie. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Sharma, A., & Paliwal, K. K. (2007). Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, *28*(10), 1151-1155.

Sievert, C., & Shirley, K. (2014). *LDAvis : A method for visualizing and interpreting topics*. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70)..

Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, *48*(2), 157-217.

Socher, R., Gershman, S., Sederberg, P., Norman, K., Perotte, A. J., & Blei, D. M. (2009). A Bayesian analysis of dynamics in free recall. *Advances in neural information processing systems*, 1714-1722.

Sra, S., Nowozin, S., & Wright, S. J. (2012). *Optimization for machine learning*. Mit Press.

Srivastava, A., & Sutton, C. (2017). Autoencoding Variational Inference For Topic Models. *ArXiv:1703.01488 [Stat]*.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.

Swanson, D. R. (1988). Historical note : Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, *39*(2), 92-98.

Tang, D., Qin, B., & Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1422–1432.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581.

Titov, I., & McDonald, R. (2008). A Joint Model of Text and Aspect Ratings for Sentiment Summarization. *Proceedings of ACL-08: HLT*, 308–316.

Turtle, H. R., & Croft, W. B. (1991). *Inference networks for document retrieval*. University of Massachusetts at Amherst.

van Rijsbergen, C. J., Crestani, F., & Lalmas, M. (2012). *Information Retrieval : Uncertainty and Logics : Advanced Models for the Representation and Retrieval of Information* (Vol. 4). Springer Science & Business Media.

Van Rossum, G., & Drake, F. L. (2011). *The python language reference manual*. Network Theory Ltd.

Viégas, F. B., & Wattenberg, M. (2008). Timelines tag clouds and the case for vernacular visualization. *interactions*, *15*(4), 49-52.

Wallach, H. M. (2006). *Topic modeling : Beyond bag-of-words*. 977-984.

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in neural information processing systems*, 1973-1981.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). *Evaluation methods for topic models*. 1-8.

Wang, C., Blei, D., & Heckerman, D. (2012). *Continuous Time Dynamic Topic Models*. 8.

Wang, S., & Manning, C. D. (2012). Baselines and bigrams : Simple, good sentiment and topic classification. *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, 90-94. Association for Computational Linguistics.

Wang, X., McCallum, A., & Wei, X. (2007). *Topical N-Grams : Phrase and Topic Discovery, with an Application to Information Retrieval*. In Seventh IEEE International Conference on Data Mining (ICDM 2007) (pp. 697-702). IEEE..

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06*, 178.

Wong, S. M., Ziarko, W., & Wong, P. C. (1985). Generalized vector spaces model in information retrieval. *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, 18-25. ACM.

Yu, C. T., & Salton, G. (1975). *Precision weighting-an effective automatic indexing method*. Cornell University.

Zeng, J., Liu, Z.-Q., & Cao, X.-Q. (2016). Fast Online EM for Big Topic Modeling. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 675-688. https://doi.org/10.1109/TKDE.2015.2492565

Zhang, H. (2004). The optimality of naive Bayes. *AA*, *1*(2), 3.

ANNEXE A RÉSULTATS DES MODÈLES LDA À 10 ET 20 THÈMES

Tableau A.1 Thèmes extraits pour la revue AE avec un modèle LDA à 10 thèmes

REPRÉSENTATION	Sac de mots	
LEMMATISATION	NON	OUI
LDA AVEC 10 THÈMES		
Thème 1	exportations matières_premières commerce_extérieur balance_paiements taux_change	test choc courbe taux_chômage prévision
Thème 2	dépenses_publiques salaire_réel taux_chômage taux_croissance théorie_générale	firme rente joueur stratégie probabilité
Thème 3	taux_chômage marché_travail salaire_minimum capital_humain statistique_canada	contrat portefeuille crédit financement firme
Thème 4	marché_travail capital_humain banque_mondiale taux_rendement pauvreté	logement ville transport tonne zone
Thème 5	politique_monétaire banque_centrale monnaie taux_change banque	ménage enfant femme famille probabilité
Thème 6	canadiens_français province_québec population_active per_capita science_économique	province provincial subvention planification agricole
Thème 7	firme coût_marginal firmes risque_moral second_rang	taux_change exportation pib balance pétrole
Thème 8	taux_change variables_explicatives taux_croissance péché_originel test	technologie pollution technologique manufacturier entrant
Thème 9	millions_dollars gouvernement_fédéral fédéral provinces tonnes	exportation européen france français importation
Thème 10	demande_finale taux_croissance valeur_ajoutée progrès_technique facteurs_production	monnaie keyes prêt crédit dépôt

Tableau A.2 Thèmes extraits pour la revue AE avec un modèle LDA à 20 thèmes

LDA AVEC 20 THÈMES		
Thème 1	coût_marginal firme second_rang risque_moral contrat contrats	soin pauvreté santé comté médecin
Thème 2	marché_travail temps_partiel baie_james système_métrique heures_travail	planification ville blé conseil commission
Thème 3	théorie_générale salaire_réel keyes taux_marge taux_profit	scénario épargne retraite pauvre pension
Thème 4	banque monnaie banques banque_centrale banques_charte	pollution agricole terre gestion firme
Thème 5	facteurs_production politique_commerciale protection_tarifaire logements services_santé	enfant famille probabilité simulation prévision
Thème 6	pauvreté revenu_personnel revenu_moyen niveau_vie revenu_net	firme taux_change dette taxe domestique
Thème 7	planification politique_économique science_économique développement_économique pouvoirs_publics	producteur pétrole compagnie européen mine
Thème 8	taux_croissance dépenses_publiques taux_rendement demande_finale stock_capital	contrat ménage joueur stratégie probabilité
Thème 9	probabilité fonction_coût variable_aléatoire incertain_information valeurs_propres	keyes monnaie smith hayek théorie_générale
Thème 10	test variables_explicatives tests variable_dépendante séries_chronologiques	bois tonne papier usine transport
Thème 11	banque_mondiale firmes taux_change commerce_international firme	economie firme division journal for
Thème 12	prix_pétrole pays_membres activité_économique zone_euro taux_inflation	exportation importation échange flux tarif
Thème 13	taux_change politique_monétaire taux_croissance	prêt dépôt crédit

Tableau A.2 Thèmes extraits pour la revue AE avec un modèle LDA à 20 thèmes (suite et fin)

	balance_paiements dette_publique	logement financement
Thème 14	taux_chômage marché_travail population_active canadiens_français salaire_minimum	femme étudiant éducation homme jeune
Thème 15	croissance_productivité technologies_information statistique_canada industries_manufacturières secteur_manufacturier	test échantillon régression pib performance
Thème 16	matières_premières commerce_extérieur millions_dollars marché_commun produits_agricoles	choc rente cycle saisonnier durée
Thème 17	péché_originel langue_maternelle capital_humain langue ressources_renouvelables	technologie taux_chômage taux_croissance courbe technologie_information
Thème 18	revenu_national taux_croissance développement_économique règles_budgétaires per_capita	monnaie crise crédit banque_centrale devise
Thème 19	millions_dollars gouvernement_fédéral provinces fédéral gouvernement_central	québécois ville toronto régional transport
Thème 20	capital_humain cycle_vie offre_travail marché_travail espérance_vie	province provincial français subvention exportation

Tableau A.3 Thèmes extraits pour la revue EI avec un modèle LDA à 10 thèmes

REPRÉSENTATION	Sac de mots	
LEMMATISATION	NON	OUI
LDA AVEC 10 THÈMES		
Thème 1	communiqué québécois globe mail the_globe	javier faiseurs commencé espionnage construit
Thème 2	trade canadiens mexique politique_commerciale convention	assouplissement george_bush inspirer conseil_national canadian
Thème 3	communiste partis socialiste socialistes communistes	éradiquer fil conseiller glenn tomates
Thème 4	monétaire monnaie nucléaires dollar droit_international	commencé proposent manques faiseurs organisationnelle
Thème 5	identité violence terrorisme otan irak	ancrage prochaine_section consolidé formée spirale
Thème 6	pétrole rfa allemande compagnies soviétiques	attardons syndicale traditionnelles cadeaux appuyait
Thème 7	brésil industries tableau gouvernance banque	voient respectées québécoises écrasante menée
Thème 8	variables théories système_international modèles university_press	éradiquer sociopolitiques devrions freedom fourchette
Thème 9	mer pêche japonais eaux soviétiques	pertes méliens continu inertie brosser
Thème 10	africains africain cee agricole russie	traités_bilatéraux nouveaux_acteurs instantanées défavorablement mettrait

Tableau A.4 Thèmes extraits pour la revue EI avec un modèle LDA à 20 thèmes

LDA AVEC 20 THÈMES		
Thème 1	dir security régionales united states	faiseurs javier étudie évaluer biens_services
Thème 2	variables pêche mer eaux systémique	éradiquer souverain sociopolitiques faiseurs javier
Thème 3	québécois aron nationalisme peuple dandurand	tomates ancrage pris_individuellement mieux_cerner capacités_militaires
Thème 4	communiqué globe mail the_globe and_mail	commencé espionnage recommencé décennies brosse
Thème 5	identité onu gouvernance agit terrorisme	pertes conseiller encouragent imaginée participé
Thème 6	africains africain africaine sanctions banque	continu inertie brosser homo frontalières
Thème 7	trade mexique politique_commerciale importations santé	manques consolidé ancrage commencé sociopolitiques
Thème 8	canadiens canadiennes canadian toronto provinces	méliens continu inertie brosser homo
Thème 9	théories capitalisme discipline système_international capitaliste	glenn québécoises respectées voient usage_force
Thème 10	droit_international convention protocole migration cour	prochaine_section formée macroéconomique spirale instantanées
Thème 11	communiste chinois partis communistes socialiste	traités_bilatéraux participé nouveaux_acteurs défavorablement lasers
Thème 12	russie maintien_paix otan conseil_sécurité forces_armées	george_bush réalités believe woodrow exemplaire
Thème 13	arctique cuba norvège	devrions freedom semences

Tableau A.4 Thèmes extraits pour la revue EI avec un modèle LDA à 20 thèmes (suite et fin)

	groenland autochtones	plan_commercial impulser
Thème 14	pétrole compagnies énergétique gaz producteurs	syndicale attardons traditionnelles conceptuellement récit
Thème 15	japonais nucléaires québécois subventions producteurs	marais évaluer vérifie lloyd_axworthy fixait
Thème 16	monétaire monnaie dollar banques banque	fourchette globe dizaines person kenneth
Thème 17	allemande allemagne rfa allemands soviétiques	proposent internationalistes colonies organisationnelle potentielle
Thème 18	brésil amérique_latine latin america argentine	conseil_national mandat proposent bosnie séparément
Thème 19	cee communautaire socialiste parlement élections	tierce_partie rejetée exécutions aptitude moratoire
Thème 20	dissuasion nucléaires armements blancs doctrine	relevait fil paramètres éradiquer réserve

Tableau A.5 Thèmes extraits pour la revue RI avec un modèle LDA à 10 thèmes

REPRÉSENTATION	Sac de mots	
LEMMATISATION	NON	OUI
LDA AVEC 10 THÈMES		
Thème 1	agreement act board strike any	devoir liant lié décrète requiert
Thème 2	chômage productivité rendement produits coût	syndicats_américains bureaux richard possession législation_provinciale
Thème 3	emplois jeunes marché_travail	législation_provinciale antisyndicale forêt

Tableau A.5 Thèmes extraits pour la revue RI avec un modèle LDA à 10 thèmes (suite et fin)

	femmes chômage	syndicats_américains capter
Thème 4	syndicaliste parti langue fédération français	blancs recherches_futures renouvellement mis_pied accéder
Thème 5	recensé acteurs ibn stratégie régulation	requiert responsabilités_familiales matière_différends adhérents conformément
Thème 6	cour sutra supra_note salarié suprême	party caractérisent axis façons_faire valoir
Thème 7	art salarié tribunal code_travail cour	allait propriété individualisées rapport_salarial engendré
Thème 8	conciliation go prévention accidents règlement	pouvoir_négociation bâtiment décisions_stratégiques mis_pied oldham
Thème 9	variables performance engagement confiance satisfaction	prévoyance basil_blackwell communication insurance jean_marchand
Thème 10	hâte bargaining would our should	possession accident communautaire ver gérard

Tableau A.6 Thèmes extraits pour la revue RI avec un modèle LDA à 20 thèmes

LDA AVEC 20 THÈMES		
Thème 1	négociation_collective représentation structures régulation syndicalisation	décisions_stratégiques impact bâtiment participation_active graphiquement
Thème 2	ouvrières pro tiens com ouvrière	blancs recherches_futures oldham maîtriser grande_majorité
Thème 3	syndicaliste parti mouvement_syndical labor tribunal_travail	party caractérisent bureaux richard façons_faire
Thème 4	our must have level should	syndicats_américains antisyndicale législation_provinciale seem publiques
Thème 5	salarié cour tribunal art juge	possession forêt maître vitrage législation_provinciale
Thème 6	santé_sécurité prévention sexuel accidents variable	politics jean_marchand ducats longues communication
Thème 7	décret art associations fonction_publique décrets	chartier_roger gestionnaires corps_enseignant retardataire pouvoir_négociation
Thème 8	évaluation participants équipes conception phase	mis_pied pouvoir_négociation statut_professionnel organisationnelle situe
Thème 9	bill go ministre compagnie gagné	accident gérard firmes enquêtes_visant per_hour
Thème 10	variables journal engagement organizational stress	bâtiment axis décisions_stratégiques gathering publiés
Thème 11	bargaining agreement hâte would collective_bargaining	déloyale considérée contrats désigne motifs
Thème 12	compétences carrière performance variables justice	engendré accident mercy motivés lisait
Thème 13	femmes jeunes partiel	allait propriété individualisées

Tableau A.6 Thèmes extraits pour la revue RI avec un modèle LDA à 20 thèmes (suite et fin)

	temps_partiel emplois	changement_technologique progresser
Thème 14	chômage revenu revenus fédéral productivité	rapport_salarial biens_services capter arrière grèves
Thème 15	rendement productivité machine machines fatigue	conformément accord relatifs oui adhérents
Thème 16	confiance france recrutement coopérative centres	retirer communautaire combinant tribu bousculer
Thème 17	profession professions professionnel professionel ingénieurs	prévoyance basil_blackwell réduits embauchés communication
Thème 18	langue français french canadiens affiliés	kruger améri protègent colline ouvrier
Thème 19	recensé ibn conseil_canadien canadian pierre	devoir ver system capter freed
Thème 20	act règlement congé safety prévoit	renouvellement accéder matière_différends responsabilités_familiales asbestos_corporation

ANNEXE B RÉSULTATS DE VARIABILITÉ DES POIDS MOYENS ATTRIBUÉS PAR LES MODÈLES LDA

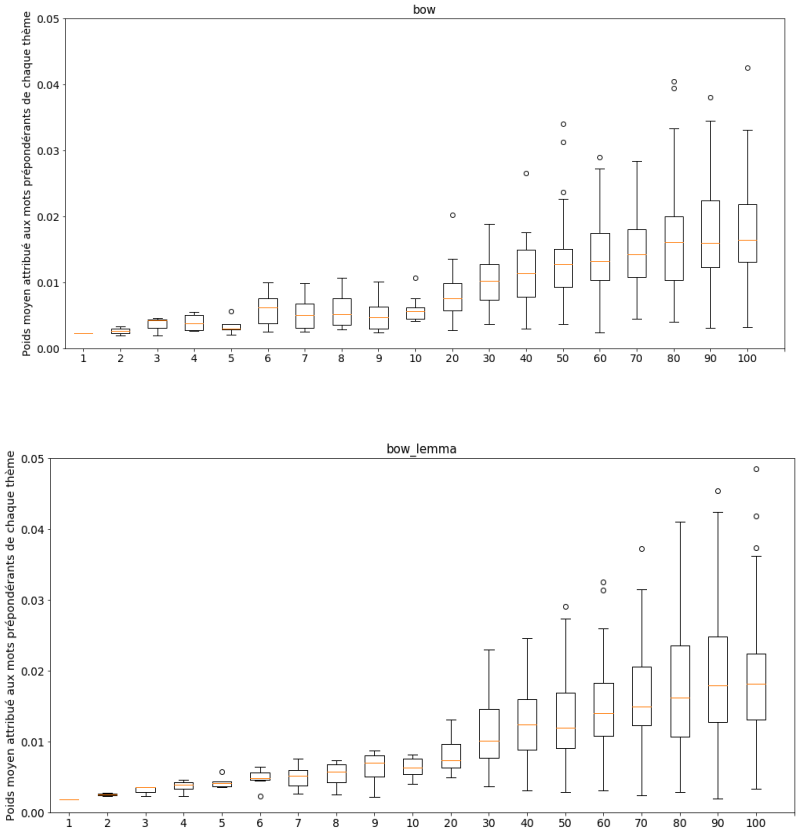

Figure B.1 Résultats de variabilité pour la revue AE

Ligne du haut : sans lemmatisation. Ligne du bas : avec lemmatisation.

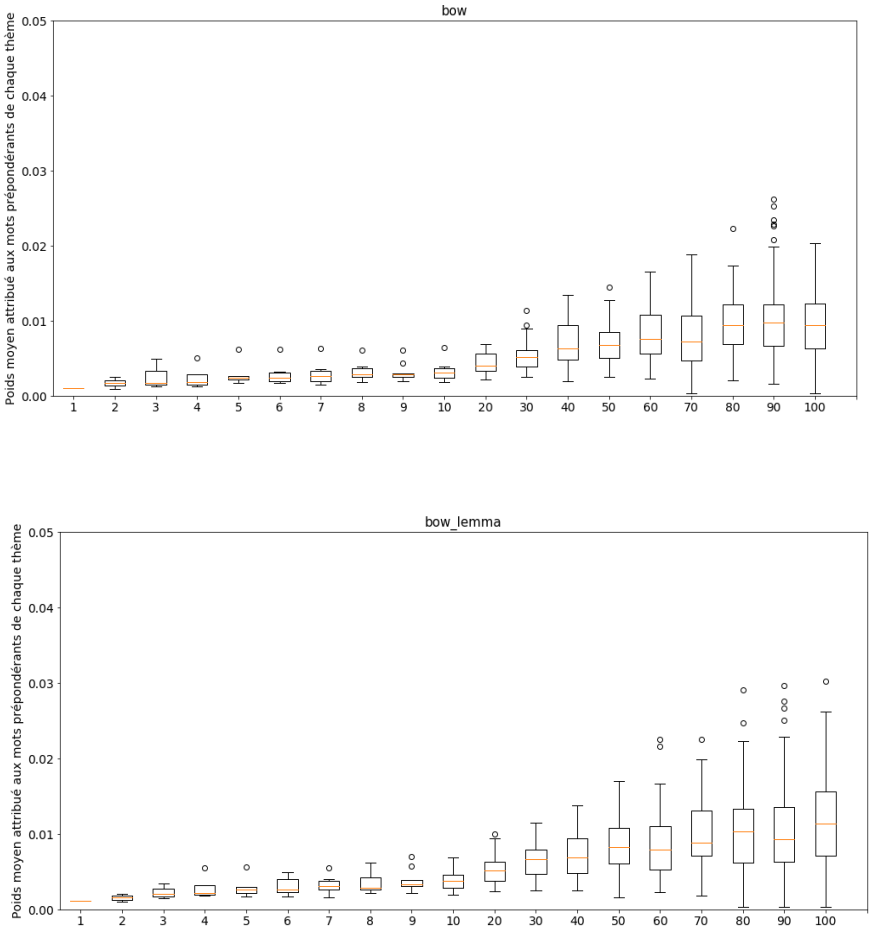

Figure B.2 Résultats de variabilité pour la revue EI

Ligne du haut : sans lemmatisation. Ligne du bas : avec lemmatisation

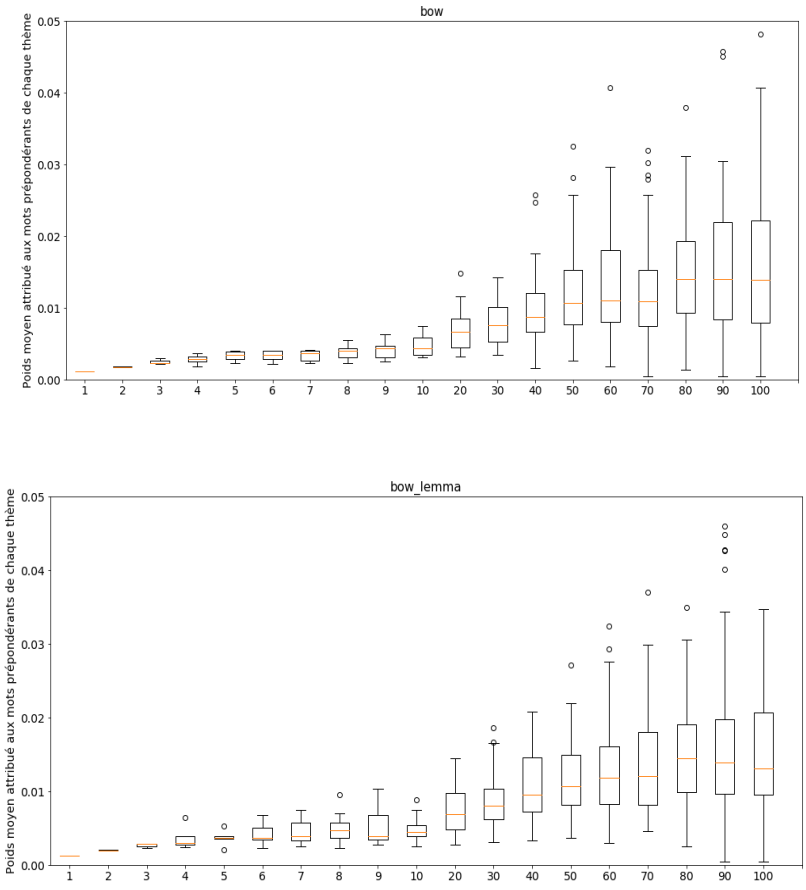

Figure B.3 Résultats de variabilité pour la revue RI

Ligne du haut : sans lemmatisation. Ligne du bas : avec lemmatisation.

ANNEXE C ÉTUDE EXPLORATOIRE : MODÈLES DTM ET DIM

En plus de l'étude de l'intégration du LDA sur un tache de fouille de documents, une étude exploratoire a été menée sur deux modèles dynamiques construits à partir du modèle LDA : Document Topic Model (DTM) et le Document Influence Model (DIM). Ces deux modèles ont été développés avec l'intention d'incorporer une information de contexte au contenu textuel des articles. En effet, ses articles de recherche sont souvent accompagnés de données structurées comme la date de publication, ou le nom des auteurs et de la revue. La motivation du modèle DTM est d'intégrer le facteur temporel dans la modélisation thématique du corpus afin d'obtenir une description plus complète du corpus. La motivation du modèle DIM est d'incorporer un facteur d'influence sémantique au modèle DTM. Le but est de construire un modèle d'influence d'un article en fonction de son vocabulaire et ainsi de proposer une nouvelle mesure d'impact bibliométrique en complément des mesures actuelles basées sur des citations.

C.1 Introduction

Le code source non parallélisé des articles ayant proposé les modèles DTM et DIM a été utilisé pour l'implémentation. Pour limiter la complexité computationnelle, des modèles à 10 thèmes ont été entraînés, avec la configuration « sac de mots — sans lemmatisation ». Un modèle DTM et un modèle DIM ont été entraînés sur chaque revue séparément. Un nœud à 32 cœurs a été utilisé sur le serveur de calcul Cedar mis à disposition par Calcul Canada pour entraîner les modèles. L'entraînement sur les trois revues a duré 5 jours.

C.2 Résultats

C.2.1 DTM

Le modèle DTM produit pour chaque année de publication dix thèmes représentés par le groupe de mots ayant le plus de poids pour ce thème. Aucune évaluation quantitative n'a été menée sur ces résultats, mais plutôt une approche de visualisation dynamique. Le but est de fournir une première interprétation de l'apport de la modélisation dynamique sur les résultats obtenus par le modèle LDA.

Visualisation statique

Une première façon d'approcher ces résultats est de moyenner le poids accordé à chaque jeton pour les différents thèmes sur l'ensemble de la durée de publication d'une revue. Cette moyennisation permet de comparer qualitativement les thèmes obtenus par le modèle dynamique avec ceux obtenus par le modèle statique. Les termes extraits sont assez généraux, avec quelques mots qui reviennent dans plusieurs thèmes comme « taux » pour la revue AE et « travail » pour la revue RI. Quelques thèmes sont redondants ou non porteurs d'information, en particulier pour les revues AE et RI. Peu de poids forts sont attribués à des bigrammes comparativement aux modèles LDA à 10 thèmes. Dans l'ensemble, les thèmes extraits en moyenne par le DTM apparaissent cohérents avec ceux obtenus par les modèles LDA. Une méthode d'évaluation externe similaire à celle présentée dans le chapitre précédent pourrait être menée pour vérifier ce résultat.

Visualisation dynamique : évolution des idées

L'intérêt essentiel du modèle DTM se situe cependant dans la mesure évolutive des thématiques. En particulier, il procure une information sur la façon dont la composition de chaque thème évolue dans le temps. Les évolutions pour les 10 thèmes de chaque modèle sont détaillées dans le fichier « *Résultats DTM-DIM — jetons.xlsx* » disponible sur le GitHub. Le Tableau C.1 présente ici deux exemples d'évolution de thèmes pour chaque revue. Les jetons avec le plus grand poids pour un thème à une date donnée sont indiqués. On observe que certains termes restent représentatifs d'un thème tout au long de la période de publication comme « pays » ou « prix » pour AE, « chine » et « conflit » pour EI, « personnel » et « travail » pour RI. Il est difficile d'associer l'analyse de l'évolution des termes avec le domaine de recherche de chacune des revues sans être expert de ces domaines. Cependant, certains cas particuliers peuvent être interprétés assez simplement. Par exemple, le deuxième thème présenté pour la revue EI contient les termes « soviétique » pendant la période de la guerre froide jusqu'en 1990 (où le terme « gorbatchev » apparaît). Puis les chercheurs s'intéressent davantage au développement de la Chine à la suite de l'effondrement de l'URSS. Aussi, pour le premier thème de la revue EI qui traite de conflit international, on repère le conflit entre Cuba et les États-Unis datant des années 60, puis apparaît explicitement la Russie à partir des années 2000. D'autres thèmes comme le premier présenté pour la revue AE et le deuxième de la revue RI traitent de thématiques plus générales (comme la politique monétaire ou les lois régissant le travail) dont les termes clés changent peu avec le temps. Le modèle DTM n'est

pas assez spécifique pour dégager une évolution explicative de ce type de thèmes. Enfin, il arrive que des thèmes perdent visiblement en cohérence à certaines périodes comme le thème 1 de la revue RI en 1980. Cette incohérence passagère peut être due à la présence de peu d'articles publiés autour de cette période dans la revue. Toutefois, ce phénomène n'empêche pas les thèmes des années ultérieures de redevenir cohérents.

D'autre part, il est possible de suivre l'évolution du poids individuel de chaque jeton représentatif d'un thème. Cette analyse permet de déterminer l'évolution de popularité de certains concepts ou sujets au sein des différentes thématiques abordées dans une revue. Les Figure C.1, Figure C.2 et Figure C.3 présentent des exemples d'évolution particulièrement notables. Chacun des graphes correspond à l'évolution des cinq jetons les plus représentatifs du thème à une date précise. Pour AE, la Figure C.1 montre que le concept « d'industries » est devenu moins populaire à partir des années 1980, au profit du concept de « technologies » et « d'information » qui apparaît en 2005. La Figure C.2 illustre pour EI le pic de discussion concernant l'union soviétique en 1990 puis la montée en popularité du sujet de la Chine à partir des années 2000. Enfin pour la revue RI, au sein d'un thème portant sur la loi du travail, on constate sur la Figure C.3 l'apparition d'un intérêt de recherche centré sur la prévention et la sécurité au travail dans les années 2000.

Les modèles d'évolution thématique peuvent donc être des indicateurs de l'évolution historique des intérêts de recherche dans un domaine. Contrairement à un suivi classique de n-grammes, ils permettent d'isoler la popularité des concepts au sein de leur thématique. Cependant, l'interprétation de ces graphes reste exploratoire, car les résultats dépendent fortement des sujets abordés dans les revues de l'étude. De plus, il faudrait confronter ces graphes à des experts en SHS (historiens, économistes etc.) pour en valider le contenu informatif et leur utilité réelle.

Tableau C.1 Exemples d'évolution de thèmes pour chaque revue

	ACTUALITÉ ÉCONOMIQUE					
	1960	1970	1980	1990	2000	2010
Thème 1	banque, politique, monétaire, canada, pays	pays, monétaire, prix, monnaie, taux	pays, prix, taux, monétaire, monnaie	taux, prix, monétaire, politique, monnaie	taux, monétaire, prix, monnaie, chocs	taux, prix, inflation, monétaire, taux_intérêt
Thème 2	production, prix, industries, valeur, cas	production, prix, demande, modèle, fonction	prix, production, capital, modèle, demande	prix, fonction, modèle, cas, coût	modèle, cas, fonction, capital, marché	consommation, marché, modèle, firmes, prix

	ÉTUDES INTERNATIONALES				
	1970	1980	1990	2000	2010
Thème 1	latine, région, pays, cuba, amérique	latine, cuba, amérique, pays, région	états, conflit, crise, conflits, régime	paix, sécurité, russie, conflits, conflit	russie, conflit, états, conflits, intervention
Thème 2	chine, soviétique, pays, chinois, socialiste	soviétique, pays, chine, soviétiques, politique	pays, soviétique, soviétiques, gorbatchev, politique	chine, the, pays, migration, chinois	chine, chinois, chinoise, migration, développement

	RELATIONS INDSUTRIELLES						
	1950	1960	1970	1980	1990	2000	2010
Thème 1	personnel, points, fonction, employés, mérite	personnel, points, profession, groupe, professionnels	professionnels, profession, groupe, and, professionnel	and, stress, niveau, vol, the	humaines, ressources, gestion, professionnels, carrière	gestion, performance, ressources, pratiques, humaines	engagement, employés, travail, résultats, organisation
Thème 2	travail,loi,québec,province,commission	loi,travail,québec,service,province	loi,publique,travail,fonction,fonction_publique	loi,travail,construction,public,sécurité	loi,travail,sécurité, règlement,canada	santé,travail,prévention,règlement,sécurité	santé,travail,services,prévention,sécurité

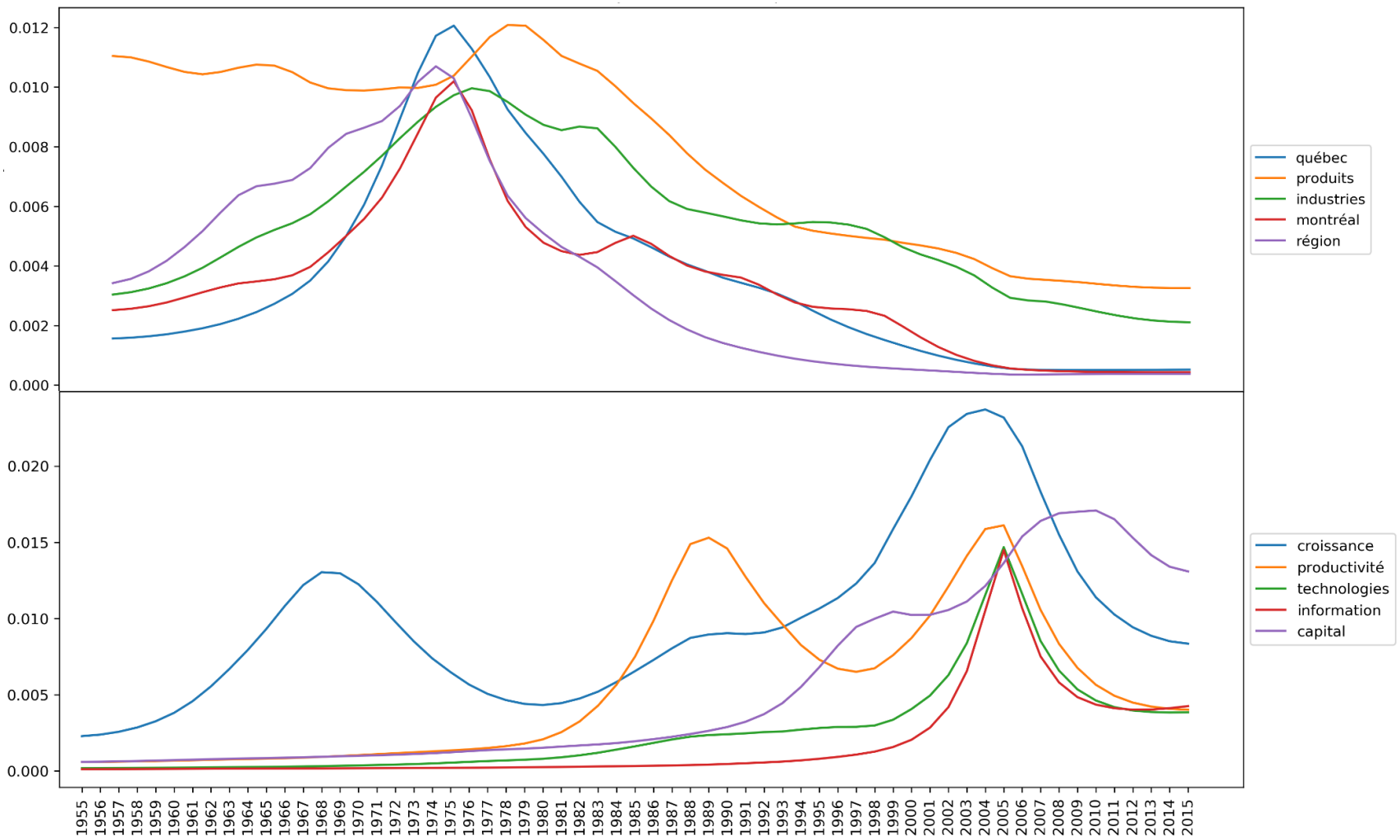

Figure C.1 Évolution temporelle du poids de jetons spécifiques d'un thème extrait pour la revue AE

En haut, jetons les plus représentatifs du thème en 1975. En bas, jetons les plus représentatifs du thème en 2005.

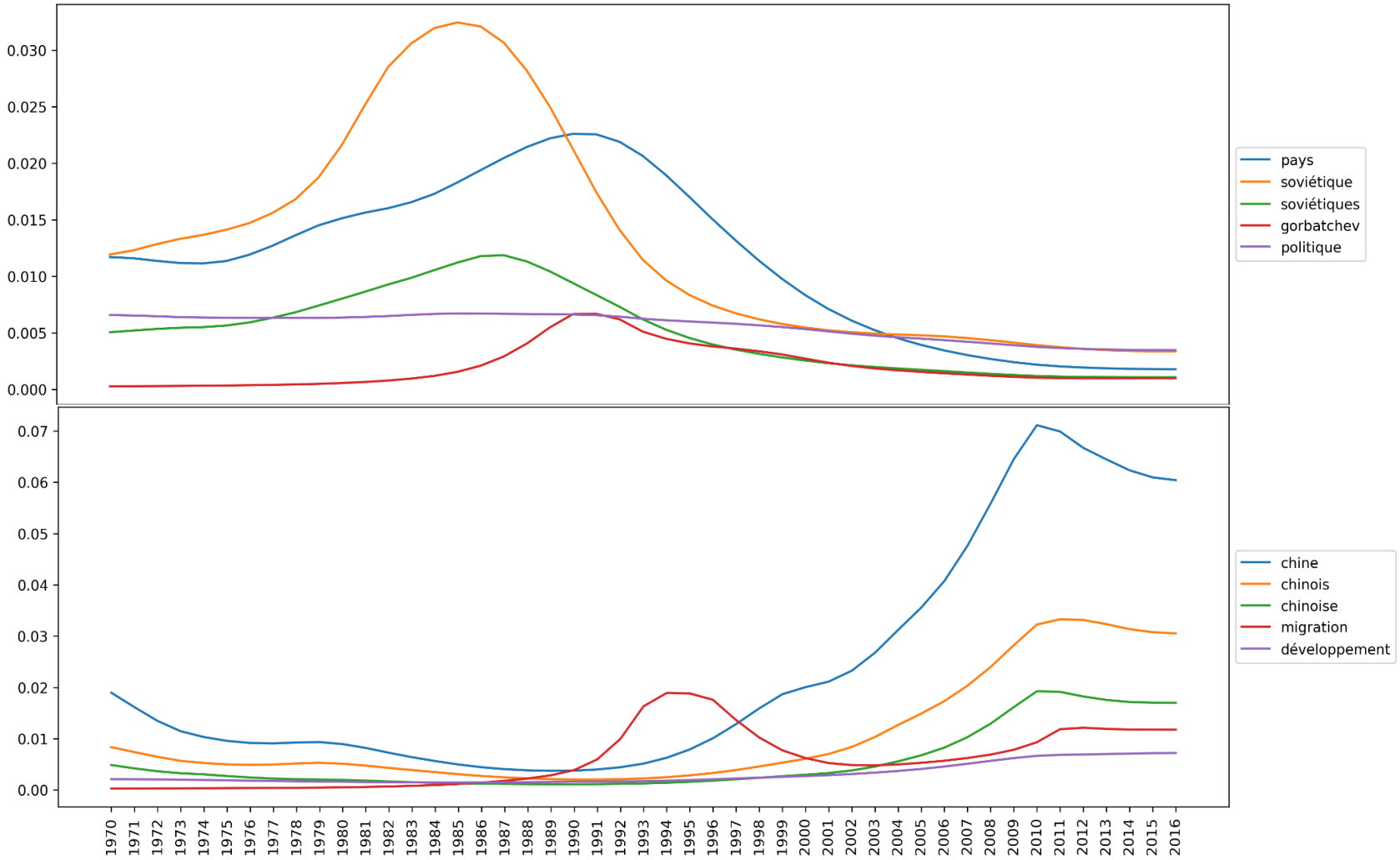

Figure C.2 Évolution temporelle du poids de jetons spécifiques d'un thème extrait pour la revue EI

En haut, jetons les plus représentatifs du thème en 1990. En bas, jetons les plus représentatifs du thème en 2010.

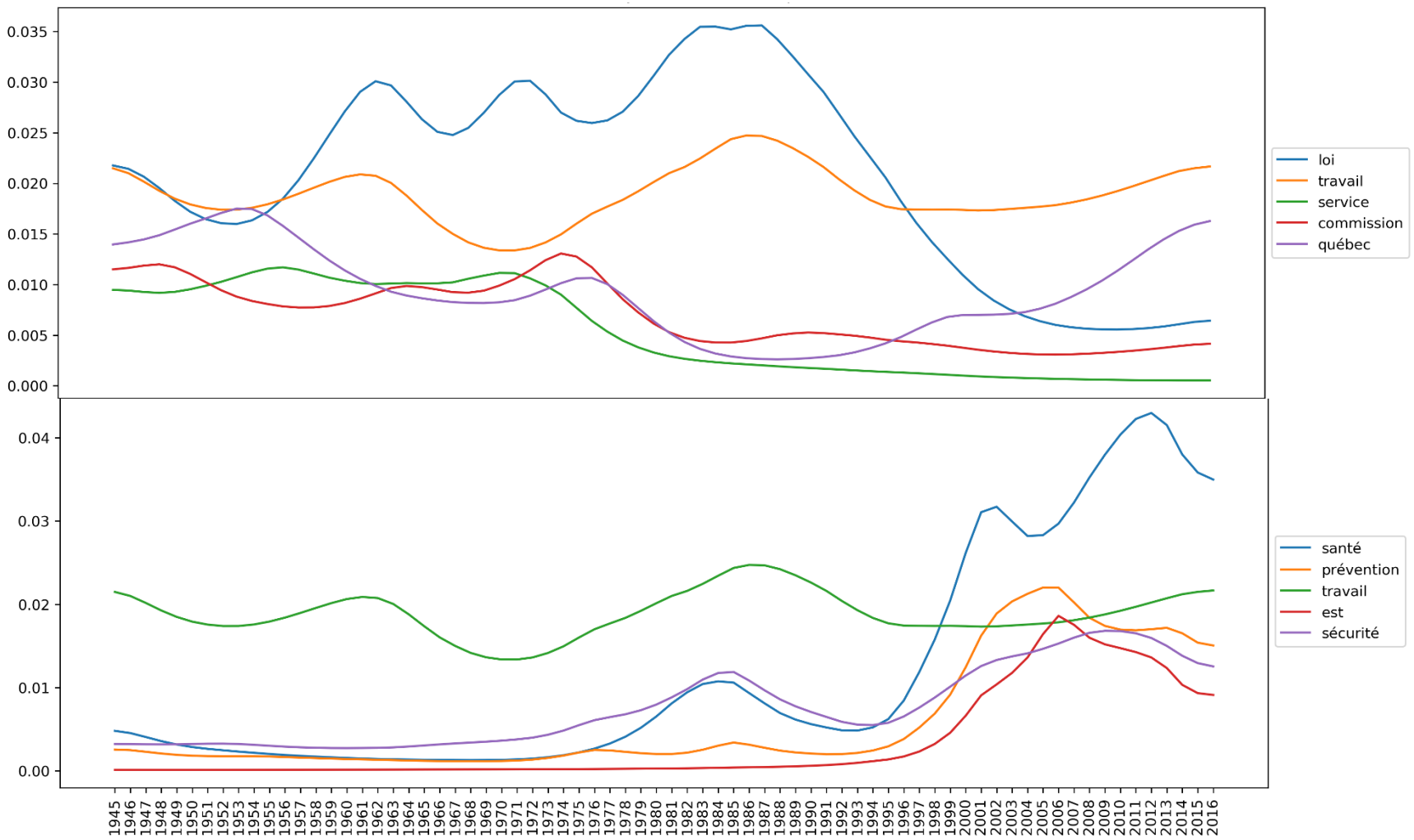

Figure C.3 Évolution temporelle du poids de jetons spécifiques d'un thème extrait pour la revue RI

En haut, jetons les plus représentatifs du thème en 1965. En bas, jetons les plus représentatifs du thème en 2005.

C.2.2 DIM

Les résultats obtenus par les modèles DIM sont décevants. La visualisation des jetons représentatifs de chaque thème indique que les thèmes extraits sont tous incohérents. L'analyse des poids attribués à ces jetons montre également que chaque mot du vocabulaire a reçu un poids similaire dans chaque thème, ce qui signale une non convergence du modèle.

Des courbes d'influence ont toutefois été tracées. Pour rappel, le modèle DIM calcule, en plus d'une distribution thématique qui évolue dans le temps, un paramètre d'influence. Ce paramètre est attribué à chaque triplet (thème, année, article). À chaque article publié à une année donnée est ainsi attribué une influence qui mesure son impact sémantique sur le contenu textuel des articles publiés aux dates ultérieures. La Figure C.4 présente les courbes d'influence annuelle pour chaque thème et chaque revue. L'influence annuelle d'un thème est calculée par la somme des paramètres d'influences attribués aux articles d'une année. L'ordre de grandeur d'influence correspond à la littérature (Gerrish et Blei, 2010) mais les courbes ne montrent aucune tendance significative.

Plusieurs raisons sont possibles pour expliquer cette non convergence des modèles d'influence. En particulier, le temps d'entraînement et le nombre de points d'entraînement peuvent être insuffisants, menant à un comportement de sous-apprentissage.

C.3 Synthèse

L'exploration dynamique des thèmes enrichit le portrait obtenu pour les données étudiées. Certaines tendances d'évolution thématique sont apparues sensées mais la discrimination rigoureuse des évolutions significatives de celles inutiles devrait être menée par des experts du domaine concerné. Les modèles dynamiques restent toutefois des modèles complexes, difficiles à entraîner et à analyser. L'échec du modèle d'influence compte tenu des ressources et du temps disponibles pour mener ce travail a montré qu'il existe encore des limites d'application de certains modèles probabilistes au corpus d'étude.

AE

EI

RI

Figure C.4 Évolution de l'influence annuelle de chaque thème pour chaque revue