Titre: Title:	Algorithmes de recherche de comparables en finance
Auteur: Author:	Kasra Dadkhah-Hadi
Date:	2020
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Dadkhah-Hadi, K. (2020). Algorithmes de recherche de comparables en finance [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5539/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Algorithmes de recherche de comparables en finance

KASRA DADKHAH-HADI
Département de génie informatique et génie logiciel

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

Décembre 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Algorithmes de recherche de comparables en finance

présenté par **Kasra DADKHAH-HADI**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Marios-Eleftherios FOKAEFS, président
Michel DESMARAIS, membre et directeur de recherche
Catherine BEAUDRY, membre

DÉDICACE

À tous mes amis du labos,
vous me manquerez...

REMERCIEMENTS

Ce travail est le résultat d'effort de plusieurs personnes sans lesquelles je n'aurais pas pu accomplir cet ouvrage. J'aimerais donc remercier toutes ces personnes ayant contribué directement ou indirectement au projet.

Je souhaite remercier mon directeur de recherche, le professeur Michel Desmarais qui m'a encouragé à entreprendre et à mener ce projet de maîtrise lors de notre cours de «système de recommandation». Sans son soutien et ses recommandations, ce travail n'aurait pas été réalisé.

J'aimerais remercier tous mes collègues de laboratoire, mes amis et ma famille pour leur conseil et leur aide précieuse tout au long ce projet d'envergure.

Je remercie également Catherine Beaudry et Marios-Eleftherios Fokaefs, membres du jury d'évaluation de mon mémoire.

Enfin, la réalisation de ce projet n'aurait pas non plus été possible sans la présence, les encouragements, l'appui et l'amour de ma mère, Afkham Hadi.

RÉSUMÉ

Parmi les tâches multiples auxquelles se consacrent les acteurs en finance, l'évaluation des entreprises est au premier plan, en ce sens que cette tâche est préalable à toute transaction d'achat, de fusion ou de vente. La valorisation d'une entreprise permet la prise de décision relative à l'achat ou à la vente, et plus précisément à la valeur donnée de celles-ci. L'approche préconisée est l'évaluation relative, qui, dans des contextes réels, consiste en une comparaison de compagnies similaires et comparables entre elles. Or, il est primordial de choisir des compagnies qui présentent un maximum de similarités, tant financièrement que dans leur structure opérationnelle et leur modèle d'affaires. Cette tâche nécessite d'acquérir une connaissance des activités plus précises de l'entreprise à valoriser. À cet égard, les analystes financiers orientent leurs pistes de recherche vers les compétiteurs et souvent à l'aide de standards de classification des compagnies et des bases de données financières, permettant de retracer les compagnies similaires.

Cependant, ces systèmes de classifications — tels que les codes SIC — ne permettent pas toujours de retracer l'ensemble des compagnies comparables. Ces regroupements sont avant tout basés sur diverses caractéristiques — par exemple, le secteur d'activité, la classe des clients de la principale source de revenus — et quelquefois sont composés de compagnies hétérogènes par centaines. Ainsi peut-on retrouver, dans un même groupe, une vaste gamme de produits, une variété de modes d'opérations, et en somme, des divergences de tailles.

Le projet de ce mémoire se consacre au développement d'algorithme de recherche des compagnies comparables pour un analyste investisseur, en utilisant les sources d'information suivantes : la capitalisation des entreprises, le groupe d'industrie, le secteur d'activité, ainsi que la description des entreprises. L'approche proposée repose sur des outils de traitement de la langue, tels que l'étiquetage sémantique. Nous présentons des modèles d'algorithmes où se combinent des données textuelles et la capitalisation des entreprises.

Plusieurs travaux portent sur la recommandation basée sur de contenu similaire, par exemple, des algorithmes de cours ou de films similaires. Or, nous en avons identifié peu permettant la recherche de compagnies comparables. De plus, nous utilisons aussi les descriptions des entreprises, ce qui a été peu traité dans les travaux précédents. Ce projet ajoute dans son originalité par la combinaison de différentes données afin d'effectuer des recommandations. En définitive, nous utilisons une combinaison de similarités de capitalisation, de descriptions textuelles et des catégorisations des entreprises.

Pour chaque type de donnée, à savoir la description, la capitalisation des entreprises, les caté-

gories d'industrie et les secteurs d'activité, des modèles de combinaison selon différents algorithmes sont développés pour obtenir des recommandations. L'utilisation des noms dans les descriptions selon une transformation TF-IDF «term frequency–inverse document frequency» avec la capitalisation de l'entreprise est un exemple de combinaison. La capitalisation, information quantitative qui varie d'une entreprise à l'autre, est normalisée et utilisée dans le calcul de proximité. Les mots définissant le groupe d'industrie sont regroupés avec les mots du secteur d'activité. Les mots regroupés du groupe d'industrie et du secteur d'activité sont classés par proximité selon la transformation TF-IDF. Les descriptions des entreprises sont réduites selon les types de mots, par exemple en retenant uniquement les noms. Nous utilisons la racine des mots, ou encore l'étiquetage sémantique pour sélectionner des mots spécifiques selon la classe des mots et selon les classes grammaticales des mots avec l'utilisation de la fonction POS (part-of-speech). Nous utiliserons aussi le graphique de connaissance «Wordnet» afin d'améliorer nos algorithmes avec la désambiguation des noms et verbes dans les descriptions et d'utiliser la similarité entre les mots dans le graphique de connaissance.

Des analyses supplémentaires sont aussi présentées, notamment, la vectorisation des documents «Doc2Vec», une technique connue sous le nom de plongement de mots «word embeddings» en utilisant une méthode augmentation des descriptions avec des mots supplémentaires. Notre modèle combinant la similarité TF-IDF et sac-de-mots des descriptions, la similarité des catégories des entreprises, la similarité des capitalisations financières, la similarité des sacs de mots des définitions désambiguées des mots des descriptions et la similarité des descriptions sur le graphique de connaissance «Knowledge-Graph» permet d'avoir une meilleure performance lorsque nous analysons les résultats.

Nos modèles d'analyse des descriptions utilisent la corrélation de Pearson. Par la suite nous combinons les analyses des données textuelles et la capitalisation pour créer la recommandation des compagnies comparables. Notre validation repose sur un ensemble de données étiquetées. Cet ensemble de données étiqueté n'existait pas auparavant, afin de pouvoir comparer les compagnies avec leurs comparables nous devions le préparer avec l'aide d'un expert. Les résultats montrent que nos modèles incluant la description dans les informations des entreprises fournissent des recommandations plus précises qu'avec la catégorie et la capitalisation uniquement.

ABSTRACT

In the financial world, during the process of merger and acquisition, one of the most complicated tasks is the company valuation. Company valuation is critical for acquisition decisions. This task is complex and invokes many methods. One of the most common approaches is relative valuation, also known as comparative valuation. This method requires a search for comparable companies. Before being able to execute the evaluation, it's important to choose the right companies with the most similar financial profile, operation and business model. Finding the right comparable company is a very long and complicated task due to the required knowledge and understanding of the company operation and other companies. Financial analysts often use shortcuts. For example, they would search online competitors or do research based on the industry classification index that allows having a very broad group of companies that are in the same industry.

Unfortunately, these industry classification lists do not always allow finding comparable companies quickly, because their classification is based on a large array of attributes often relevant to other purposes than the company valuation. For this reason, some industry classification sector regroups more than a hundred companies inside them that do not offer the same product or have different operation management styles and are of different sizes. These classifications are also made manually. Moreover, it often happens that a company operates in many industries at the same time and it also happens that companies don't have competitors or direct similar companies due to their specialized activities or technology.

The aim of this study is to develop algorithms to find comparable companies with the use of natural language processing tools such as word semantics, word embedding, knowledge graphs in order to be able to create a recommendation of companies based on their publicly available content. In order to obtain comparable company recommendations, we will use different sources of information. We will use their market capitalization, industry classification, sector activity, and company descriptions.

This research presents many novelties. In previous research, similarity algorithms for example to have university course recommendations, or movie recommendations have been presented, but not for comparable companies' recommendations for valuation purposes. Another part of the research novelty comes from the chosen information of the companies and the target comparable validation dataset preparation with an expert. The final part of novelty is the mixture of financial and textual similarity finding regrouped for recommendations of comparable companies using different natural language processing methods.

For each type of data, for example, the description, the market capitalization, the industry sector, the industry activity, a model of analysis is developed and combined to obtain recommendations. Each type of data requires different manipulations, for example, the capitalization is numbers that have high variation between them, hence are log-scaled and normalized. The industry category is analyzed with the words that are describing the category with TF-IDF term frequency method.

The textual similarity analysis on the company's descriptions is focusing on natural language processing techniques like word labelling, stemming, part-of-speech tagging, knowledge graph definition disambiguation, knowledge-graph path similarity, word embeddings and document embeddings. When we present our different models, these techniques will be detailed.

Our results show that the novel model with categories, description and capitalization allows to make recommendations that are relatively precise. The model contains capitalization similarity, category similarity and multiple description similarity models. The description similarity models included raw description bag of words TF-IDF similarity, TF-IDF similarity of disambiguated word from raw description based on trained knowledge-graph and finally description path similarity of disambiguated definitions.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

La recherche d'entreprises comparables pose des défis aux acteurs du domaine financier. Pour ce faire, ils utilisent plusieurs outils tels que les moteurs de recherche, les systèmes de classification d'industries, les rapports annuels des compagnies, la recherche manuelle de compétiteurs et la recherche dans des bases de données financières. Cette méthode présente des limites, car elle nécessite beaucoup de temps de recherche d'entreprises et beaucoup de connaissances sur la compagnie. Ce mémoire présente des algorithmes de recherche de compagnies comparables avec la combinaison d'informations textuelles et la capitalisation des entreprises. Les informations textuelles des entreprises contiennent des informations importantes sur les opérations et les produits ou services des entreprises. Dans ce premier chapitre, nous présentons l'originalité, les objectifs et le plan du mémoire.

1.1 Définitions et concepts de base

Ce mémoire s'appuie sur de nombreux concepts tels que la finance, l'analyse linguistique et la recherche de similarité, qui sont définis au fil du mémoire. À ces concepts s'ajoutent les notions financières utilisées pour la recherche d'entreprises comparables. Les notions de système de recommandation permettront de comprendre les méthodes de recommandation de contenu similaire. Les notions d'analyse linguistique décrieront les méthodes de transformation de texte, de filtrage et d'analyse de texte.

1.2 Éléments de la problématique

Le défi que rencontre un analyste financier lors de sa recherche des entreprises comparables afin de réaliser l'évaluation relative d'une entreprise est considérable et nous en présenterons le détail dans le prochain chapitre. Ainsi, la problématique du sujet de ce mémoire est la conception d'algorithmes permettant le regroupement d'entreprises comparables. Nous trions les données accessibles des entreprises publiques, parmi lesquelles nous sélectionnons les plus pertinentes. L'éventail des données d'entreprises, telles que leur rapport annuel, leurs informations comptables et leurs communications, présente une diversité importante. Il faut utiliser des méthodes d'analyse sémantique pour des données textuelles ou des méthodes numériques pour le traitement de données financières. À la suite de ce travail préliminaire de sélection des informations, nous pouvons dès lors établir un algorithme de recommandation des entreprises comparables.

1.3 Originalité et objectif de recherche

L'objectif de cette recherche est d'appliquer des techniques, empruntées à celles du domaine du traitement de la langue et à celui des algorithmes d'apprentissage pour la recherche de compagnies comparables. Les données des entreprises utilisées sont leur capitalisation, les mots de leur catégorie d'industrie et de leur secteur d'activité, et enfin, les mots de leur description d'entreprise.

Nous avons développé plusieurs approches permettant la recherche de compagnies comparables dont une de nos démarches est de combiner certaines informations textuelles avec la capitalisation des entreprises. Nous avons aussi étiqueté et validé à l'aide d'un expert les comparables cibles des entreprises de nos données. Une autre démarche est dans le filtrage des descriptions des entreprises, avec lesquelles nous utilisons des méthodes d'analyse sémantique de la librairie connues de Python «natural language toolkit». Ainsi, nous filtrons certains types de termes ou certains groupes verbaux provenant de chaque phrase des descriptions. Une autre approche est l'intégration des graphiques de connaissances préentraînées «WordNet» pour trouver des descriptions similaires. Nous utilisons des données réelles d'entreprises et nos sources de données financières proviennent des compagnies disponibles de la base de données «US Company Fundamentals» d'Intrinio. Chaque compagnie est représentée par sa description, les mots de sa catégorisation d'industrie regroupés avec les mots de son secteur d'activité et sa capitalisation.

Pour évaluer la performance de nos modèles, nous utiliserons des valeurs cibles d'entreprises comparables dont on comparera les recommandations des modèles en utilisant le taux de rappel global. Nous comparons aussi nos modèles par rapport à un modèle de base contenant uniquement les catégories des entreprises et leur capitalisation financières.

À partir des données des compagnies, plusieurs modèles présentés dans le chapitre 5 sont créés. Nous étudierons plus spécifiquement les techniques comme la transformation TF-IDF, la vectorisation des descriptions «Doc2Vec», un mécanisme d'augmentation de données et l'utilisation de graphique de connaissances «knowledge-graphs». L'utilisation des graphiques de connaissances «knowledge-graphs» permettra d'évaluer la réduction de l'ambiguïté des mots des descriptions selon leur définition dans le contexte d'une phrase. L'utilisation des graphiques de connaissances «knowledge-graphs» permettra aussi trouver des descriptions similaires selon la distance du chemin dans le graphique de connaissances «knowledge-graph path similarity» entre les mots des descriptions.

1.4 Plan du mémoire

Ce mémoire est organisé de la manière suivante. Un premier chapitre se consacrera au contexte d'utilisation de l'analyse financière. Par la suite, nous présenterons une introduction aux techniques de recherche d'information et de traitement de texte. Nous y réviserons les techniques les plus courantes, tel que la fréquence inverse des termes, l'analyse sémantique, les modèles d'espaces vectoriels et les graphiques de connaissances. Ces chapitres nous permettront l'exploration de travaux portant sur les méthodes actuelles de recherche d'entreprises similaires. Nous implémenterons et évaluerons plusieurs algorithmes et techniques pour effectuer des recommandations de compagnies similaires. Ainsi pourrons-nous comparer les résultats des différents algorithmes. Pour chaque étude des modèles, nous utiliserons le rappel sur nos compagnies cibles acceptées par des professionnels du domaine. Avant de conclure, nous présenterons nos résultats issus de tous les algorithmes utilisés et du modèle optimal qui en sera ressorti.

CHAPITRE 2 L'ANALYSE DE COMPAGNIES COMPARABLES

La recherche d'entreprises publiques comparables est une tâche souvent assignée à un analyste financier avec des ressources variables en temps et en outils technologiques tels que des sources de données financières. Selon les limites de temps, de ressources humaines et technologiques, un analyste financier doit chercher des entreprises comparables et en faire une valorisation. En considération du temps et des ressources alloués pour sa recherche d'entreprises comparables, un analyste financier cherche des résultats satisfaisants et adéquats plutôt qu'idéaux et parfaits. La «satisfiction» (Simon, 1955) réfère à ce concept de compromis entre l'idéal et les ressources disponibles. Un exemple de cas de figure idéal d'entreprise comparable est une situation où une compagnie similaire et comparable apparaît d'emblée, en ce sens que la similarité des produits ou services et la comparabilité entre les deux entreprises sont évidentes et adéquates. Souvent les compagnies similaires «Peers» selon quelques critères sont utilisées dans la littérature. D'une part, les analystes du côté de la vente «Sell-Side» analysent et publient dans leurs rapports «Sell-side reports» des recommandations, qui couvrent généralement les entreprises dans la même industrie doivent faire des recommandations d'achat et choisissent des compagnies similaires «Peers» selon le secteur d'activité, le groupe d'industrie ou une activité spécifique telle que le média (Bradshaw et al., 2009). Les comités de compensations utilisent des compagnies paires similaires «Peers» pour déterminer et justifier les compensations des membres exécutifs (Albuquerque et al., 2013). Les auditeurs dans le domaine de la comptabilité des entreprises utilisent des compagnies similaires «Peers» selon une perspective comptable afin d'effectuer leurs analyses sur la relation entre informations financières et non financières (Hoitash et al., 2006). Notre recherche se penche non seulement sur des compagnies similaires «Peers», mais plus précisément sur la recherche des compagnies comparables qui demande davantage de critères similaires pour être comparables.

Les investisseurs utilisent les comparables pour déterminer le mérite d'un investissement permettant d'allouer efficacement et de faciliter la confiance de l'allocation de capital (De Franco et al., 2011). Un exemple où l'on atteint l'optimal est lorsqu'on cherche une entreprise comparable à «Coca-Cola Company». Nous pouvons retrouver son plus grand rival «PepsiCo, Inc.» qui est une entreprise similaire «Peer» et comparable. Le cas de figure davantage contraignant nécessite d'extrapoler la recherche d'entreprise comparable dans d'autres secteurs d'activité et dont la comparabilité n'est pas évidente à première vue. À cet égard, nous considérons qu'une liste d'entreprises est composée de comparables dans la mesure où elle répond, autant que possible, aux conditions qui s'articulent autour des facteurs principaux suivants (Fris and Gonnet, 2010). Premièrement, le facteur des caractéristiques de la propriété telles que

les actifs de l'entreprise, les produits et les services offerts par l'entreprise. Le deuxième facteur est celui des résultats de l'analyse fonctionnelle tels que les opérations de l'entreprise. Troisièmement le facteur des contrats tels que les brevets et les contrats avec les autres entreprises. Quatrièmement le facteur des circonstances économiques qui affectent l'entreprise, et finalement le facteur de la stratégie de l'entreprise.

2.1 Les méthodes de valorisation

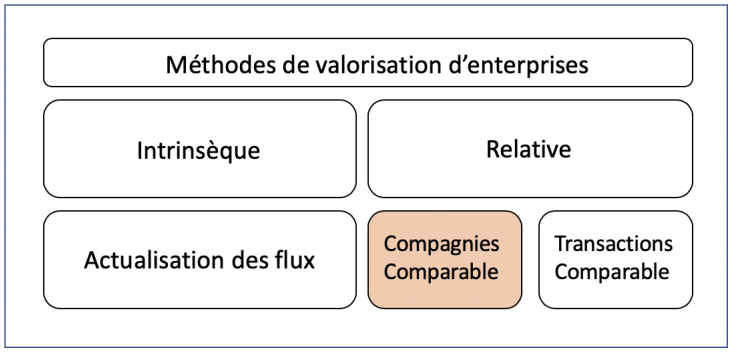

Figure 2.1 Principales méthodes de valorisation

Dans la figure 2.1 on présente les principales méthodes de valorisation des entreprises qui sont soit intrinsèques soit relatives. Dans les méthodes intrinsèques, on retrouve l'actualisation des flux et dans les méthodes relatives on retrouve la méthode avec les compagnies comparables et la méthode avec les transactions comparables. L'actualisation des flux connue en anglais sous le nom de «Discounted Cash Flow» est basée sur des profits anticipés. Ces méthodes sont couramment utilisées par les banques d'investissement, l'équité privée et les entreprises dans leurs acquisitions (Deev, 2011). Cela dit, dans un cas comme dans l'autre, le point de départ est le même, l'étude les actifs de l'entreprise, afin de prendre en compte la situation spécifique de l'entreprise. La tâche d'un analyste financier, lors d'une valorisation d'entreprise, est de donner une valeur à l'entreprise selon la méthode qui est adéquate à l'entreprise en question. Celle-ci, permets d'y assigner une valeur et de positionner la compagnie par rapport aux avantages compétitifs et aux plans d'avenir qu'elle développe. Prenons l'exemple de la vente d'une maison avec des maisons comparables. Pour obtenir la valeur du marché, il faut chercher des transactions récentes de maisons comparables selon leur taille, le type de construction, leur quartier, l'entourage, etc. C'est à partir des maisons comparables, basées sur les critères tout juste évoqués, qu'un vendeur détermine par exemple le prix au pied carré et la multiplie par la grandeur de sa maison. Ainsi il obtient la valorisation de sa maison au prix du marché.

Les entreprises ont divers concepts de leur valeur. D'abord, selon la comptabilité la valeur de l'entreprise connue en anglais sous le nom de «Enterprise Value» représente toutes les sources de capitaux pour toutes les opérations accessibles aux actionnaires, comme la dette, les actifs et les actions. Il existe néanmoins des actifs dont la valorisation est difficile à estimer ou intangible — c'est-à-dire qu'elles ne sont ni représentables ni échangeables — telles que le nombre d'utilisateurs d'une application web. Le site de visualisation de vidéo «www.youtube.com» est un bon exemple. Pourtant, les actifs non tangibles ont un impact important sur la valorisation d'une entreprise. La capitalisation boursière, donnée qu'on utilise dans nos algorithmes, représente la valeur de l'entreprise au marché. Nous estimons que la capitalisation de l'entreprise permet de mieux quantifier la taille d'une entreprise, car elle est actualisée au marché présent et prend en compte son contexte économique et le prix que les investisseurs sont prêts à payer.

2.2 L'utilisation des ratios financiers après la recherche de compagnies comparables

La recherche de compagnies comparables est un une tâche qui impose une compréhension d'une entreprise. Chaque compagnie est unique. Il est important de noter qu'on évite d'utiliser des données financières qui se résumeraient à la recherche donnée fondamentale similaire. Or nous n'utilisons pas les ratios ou données fondamentales pour retrouver des entreprises comparables dans nos algorithmes, car ces données sont utilisées lors de la valorisation des entreprises une fois qu'on a déterminé nos comparables. Les analystes, à défaut de disposer d'informations financières identiques entre les entreprises, ont généralement recours à des multiples et des ratios entre les valeurs financières des entreprises. Une fois les entreprises comparables trouvées, les analystes utilisent les valeurs comptables des différentes entreprises pour créer des ratios financiers qui permettent d'analyser et de valoriser la compagnie cible. Pour déterminer la valeur juste d'une entreprise, la méthode d'évaluation relative s'intéresse particulièrement à des ratios — par exemple au multiple de la valeur de l'entreprise sur les résultats d'exploitation avant intérêts, les impôts et amortissements (EV/EBITDA) — des autres entreprises au même moment. L'efficacité et la facilité d'usage de la méthode de valorisation relative ont déjà été rapportées par nombre de travaux (Kaplan and Ruback, 1995; Gilson et al., 2000).

Les ratios financiers présentent un avantage, en ce qu'ils facilitent la comparaison entre deux entreprises. En effet, en utilisant les ratios financiers, il n'est pas nécessaire de faire un grand nombre d'ajustements. Par exemple, chaque entreprise a des niveaux d'endettement différents ou des méthodes d'amortissement de ses actifs différentes. La comparaison directe

de ces différentes méthodes pour chaque entreprise affecte la vraie valeur des actifs totaux. C'est pourquoi l'utilisation de ratios — tels que le ratio d'immobilisation de l'actif — dispense du besoin de détailler les différences entre les deux entreprises.

L'évaluation relative d'une entreprise avec des compagnies comparables est très répandue. Elle est d'abord accessible et rapide, simple a utiliser, et possède une grande concordance avec les prix du marché (Meitner, 2006). L'application de la méthode est directe, dès lors que les compagnies comparables et le modèle de valorisation sont choisis. Et enfin, elle ne nécessite nulle spécialisation particulière ni ne requiert de connaissance dans le domaine de la finance.

Plusieurs ratios financiers existent pour différentes tâches d'analyse des entreprises tels que la profitabilité, la dette, l'efficacité et le flux de trésorerie. Les principaux ratios financiers pour l'analyse comparative sont le prix sur l'équité (P/E), la valeur de l'entreprise sur les résultats d'exploitation avant intérêts, les impôts et amortissements (EV/EBITDA) et la valeur de l'entreprise sur les revenus (EV/Revenu).

2.3 La recherche de compagnies comparables

Si les analystes investisseurs doivent effectuer la recherche de compagnies comparables afin de pouvoir déterminer la valeur d'une entreprise, il arrive qu'une compagnie requière une étude davantage approfondie pour la recherche de compagnies comparables. Cette situation survient notamment lorsqu'un nombre trop important de compagnies possèdent la même cote de classification d'industrie. C'est le cas, par exemple, des industries de technologie numériques. Dans ces cas particuliers, la recherche des analystes financiers doit être relayée à un travail manuel, appliqué aux activités de la compagnie, c'est-à-dire la recherche de compétiteurs ou de compagnies dans d'autres industries, possédant néanmoins un modèle d'affaire similaire.

Les analystes font par ailleurs usage d'outils davantage spécialisés en vue d'effectuer ce travail tel que des bases de données privées. Ces outils mettent à disposition les divers secteurs d'activité, créés et destinés aux analystes. Ces outils sont notamment disponibles du terminal Bloomberg et de la base de données CapitalIQ. Ces logiciels permettent de détailler les nouvelles, les communications, les livres financiers des entreprises, la description des entreprises et d'obtenir leur classification d'industrie. Ici peut alors avoir lieu, manuellement, la recherche des compagnies les plus similaires et comparables. Ainsi savons-nous que les outils actuels sont agrégés par des professionnels qui ont préalablement classifié et départagé les secteurs des compagnies. De plus, ces classifications sont continuellement mises à jour, ainsi que les

listes de leurs compagnies similaires ou compétitrices. Par conséquent, c'est sur ces listes que les professionnels se penchent : elles constituer essentiellement la source dont ils font usage. Dans le cas qu'un analyste financier ne trouverait pas de compagnies comparables adéquat, cette recherche est entreprise manuellement.

2.4 Les comparables par standards de classifications (Code SIC, NAICS et GICS)

Dans nos modèles, nous utiliserons les mots représentant les différents groupes des standards de classification tels que le secteur d'activité et le groupe d'industrie. Ces mots permettent d'avoir une information dans nos algorithmes sur le secteur principal ou le groupe d'industrie principal des activités de l'entreprise. Les standards de classification (SIC, NAICS, GICS) se traduisent par un nombre de quatre à six chiffres, dont l'utilité est la suivante. D'abord, de catégoriser l'industrie dans laquelle une compagnie s'exerce, et de plus, organiser les industries en fonction de leurs activités commerciales. Le standard de classification SIC «Standard industry classification», créée par les États-Unis en 1937, apporte une aide considérable quant à l'analyse des activités économiques, et ce, pour plusieurs de leurs agences gouvernementales.

Toutefois, le standard de classification SIC a été remplacé en 1997 par un système à six chiffres ; nommé le North American Industry Classification System «NAICS» ou en français le «SCIAN» Système de classification des industries de l'Amérique du Nord. Son objectif initial est de mettre à disposition un standard de classification aux entreprises de différents pays, tels que les États-Unis, le Canada, ou encore le Mexique. Ce système de classification est révisé manuellement par rapport aux clients de la source de revenus principale des compagnies tous les quatre ans, de sorte qu'il tienne compte des mises à jour des données de l'industrie pour les classifier (US Bureau of labor statistic, 2019).

Par ailleurs, la «Security and Exchange Commission» (SEC) est une agence gouvernementale américaine. Celle-ci régule le marché en faisant encore usage des codes SIC. Mais un nombre important de compagnies, de banques et même d'agences gouvernementales utilisent encore ces codes de classification pour de multiples fins, par exemple dans le cadre d'une classification des compagnies en regard aux taxes, ou pour la recherche de comparable, pour l'attribution d'un score de crédit et pour le comité exécute afin de déterminer la compensation des dirigeants.
Le tableau 2.1 met en évidence les principales industries déterminées par les deux premiers chiffres du code de six chiffres du NAICS. Nous pouvons toutefois remarquer que le nombre de compagnies dans chaque regroupement varie considérablement, allant de 32 950 pour l'in-

2 premier code NAICS	Nom de l'industrie	Nombre de compagnies
11	Agriculture, Forestry, Fishing and Hunting	389 628
21	Mining, Quarrying, and Oil and Gas Extraction	32 950
22	Utilities	41 783
23	Construction	1 488 851
31-33	Manufacturing	645 157
42	Wholesale Trade	706 904
44-45	Retail Trade	1 825 399
48-49	Transportation and Warehousing	589 292
51	Information	356 941
52	Finance and Insurance	792 434
53	Real Estate and Rental and Leasing	869 567
54	Professional, Scientific, and Technical Services	2 229 824
55	Management of Companies and Enterprises	72 727
56	Administrative and Support and Waste Management and Remediation Services	1 797 527
61	Educational Services	428 654
62	Health Care and Social Assistance	1 693 539
71	Arts, Entertainment, and Recreation	369 183
72	Accommodation and Food Services	901 202
81	Other Services (except Public Administration)	1 895 018
92	Public Administration	265 129
	TOTAL DES ENTREPRISES	17 391 709

Tableau 2.1 Catégories d'industrie avec 2 premiers chiffres du code SCIAN (NAICS) (NAICS Association, 2019)

dustrie minière à 2 229 824 pour l'industrie des services techniques et professionnels.

La grande différence s'explique par les fondements des codes de classification, plus précisément les bases sur lesquelles les classifications sont faites. En effet, la création du code de six chiffres du NAICS se fait en fonction du secteur d'activité principal de la compagnie, mais d'abord et avant tout, à partir de ses sources de revenus. Ce modèle de classement présente des limites lorsque les compagnies sont très grandes ou ont de très diverses sources de revenues. Le code comprend un interstice, divisant dès lors ses deux premiers chiffres des quatre autres. Tandis que les deux premiers chiffres déterminent l'industrie primaire en 20 catégories, les quatre autres chiffres ont quant à eux 1066 catégories d'industrie. Pour plus de précision, le troisième chiffre indique la sous-section ; le quatrième indique le groupe de sous-industrie dans lequel la compagnie opère ; le cinquième chiffre renvoie à l'industrie particulière ; quant au sixième chiffre, il indique l'industrie nationale spécifique.

Examinons par exemple le code 111110 avec la figure 2.2 : la compagnie se trouve dans le secteur 11, sous-secteur 111, groupe d'industrie 1111 et industrie 11111. Ce code global (111110) représente l'industrie de la culture du soja «Soybean farming, field and seed production». Le dernier chiffre représente l'industrie nationale (ETATS-UNIS, CANADA, MEXIQUE) dont le 0 indique qu'il n'y a pas plus de détails national pour le code en question.

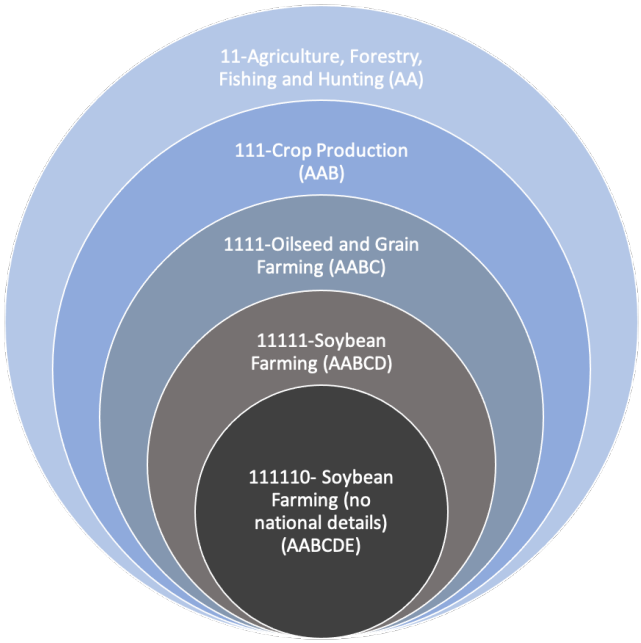

Figure 2.2 Exemple NAICS (111110) culture du soja en format AABCDE

Un autre standard de classification privé s'appelant le «Global Industry Classification Standard» (GICS Code) a été développé par SP Dow Jones Indices et MSCI une compagnie de produits financiers. L'objectif de ce standard de classification est de permettre aux analystes investisseurs d'avoir une meilleure forme de classification que les précédents. La structure du GICS a 11 secteurs, 24 secteurs d'industries et 69 groupes d'industries. Ce code de classification est considéré universel, fiable, flexible et évolutif par des changements annuels qui permettent une transparence et une efficience dans un processus d'investissement par secteurs.

La classification des compagnies est refaite à chaque année pour le Code GICS selon les rapports annuels des entreprises, les états financiers et les rapports des analystes.

2.5 Les outils et les méthodes des analystes financiers

Les analystes financiers utilisent une gamme d'outils qui accélèrent l'accès aux informations des compagnies. Certains sont voués à l'analyse comparative des compagnies par les analystes, permettant ainsi de faciliter et d'ajouter de la précision au travail des analystes — dont l'objectif ultérieur consiste en une prise décision d'investissement. Parmi ces outils, les plus connus sont les suivants : Reuteurs, CapitalIQ et Bloomberg. Au-delà des informations quant à l'analyse fondamentale des entreprises, ces outils s'engagent également à l'actualité et aux

communications des entreprises, de sorte que les analystes financiers aient un accès fiable, simple et complet aux entreprises qui les intéressent. Voici, dans ces conditions, quelques exemples d'informations publiquement accessibles :

1. Les rapports annuels et communications des entreprises
2. Leurs secteurs d'activités
3. La capitalisation
4. Le prix courant des actions de l'entreprise

D'autres exemples d'informations accessibles avec les logiciels spécialisés sont :

1. Les données comptables des entreprises (dettes, actifs, ventes, etc.)
2. Les données comptables historiques (dettes, actifs, ventes, etc.)
3. Les actionnaires des entreprises avec leurs parts et historique
4. Les consensus de prédictions de leurs résultats
5. Le prix courant des actions de l'entreprise

Une fois que les analystes financiers obtiennent toutes les informations dont ils ont besoin, leur travail se poursuit sur le support des logiciels Microsoft Excel et Powerpoint. Ce faisant, ils modélisent d'abord l'information par rapport à la compagnie qui leur semble prometteuse en vue d'investissements, et en dernier lieu, ils exposent l'information aux parties prenantes.

2.6 Récapitulation

Dans ce chapitre, nous avons présenté les bases de la recherche de compagnies publiques comparable en expliquant le contexte dans lequel cette activité est effectuée, les méthodes de valorisations supplémentaires qui existent pour le contexte des compagnies publiques, l'importance des ratios financiers dans l'évaluation relative d'une entreprise, les sources de donnée des comparables pour les analystes financiers tels que les standards de classifications (SIC, NAICS, GICS). Maintenant que nous avons un contexte financier de la recherche de comparable, dans le prochain chapitre nous expliquerons les techniques et algorithmes pour la recherche d'informations similaire et le traitement de la langue naturelle.

CHAPITRE 3 TECHNIQUE DE RECHERCHE D'INFORMATION SIMILAIRE ET TRAITEMENT DE LA LANGUE NATURELLE

Dans le chapitre précédent, nous avons présenté l'analyse de compagnies comparables. Comme nous utiliserons dans nos modèles des méthodes de traitement de la langue pour les secteurs d'activités, les groupes d'industries et les descriptions des entreprises, nous détaillerons les mesures de similarité et les méthodes de traitement de langue naturelle dans cette section.

3.1 Méthodes de recherche par similarité de texte

La recherche d'information procède en sondant les documents qui semblent pertinents pour un utilisateur. Nous possédons des informations textuelles, telles que la description des entreprises, à partir duquel nous devons retrouver des compagnies comparables. Nous devons tout d'abord effectuer des transformations sur les descriptions et appliquer des méthodes de similarité pertinentes dont nous détaillons dans ce chapitre.

3.1.1 La similarité dans l'espace vectoriel avec sac de mots

La représentation par sac de mots, connue en anglais sous le terme de «bag-of-words», d'un document est très répandue lors de la recherche d'information. Le principe général est de représenter le document par une matrice qui identifie l'occurrence des mots d'où chaque mot est représenté sur les lignes et les documents sur les colonnes. Le défaut majeur de la représentation par sac de mots est que la dimension des colonnes devient trop grande à cause du nombre de mots présent dans tous les documents et crée des matrices creuses.
Dans le tableau 3.1, nous pouvons observer une représentation vectorielle de type sac de mot des 2 documents suivants :
Document 1 : «La compagnie ABC distribue des fruits. Les fruits sont des pommes et des bananes. »
Document 2 : « ABC distribue des pommes et des bananes»

Les mots de tous les documents sont sur les lignes dont l'occurrence par document est présentée dans les colonnes selon le document. Par exemple, le mot compagnie est présent une fois dans le document 1 et aucune fois dans le document 2. En considérant une représentation vectorielle d'un corpus de documents, on peut calculer la similarité entre les documents. Plusieurs mesures de similarité peuvent être utilisées pour quantifier la proximité sémantique

MOT	Document 1	Document 2
la	1	0
compagnie	1	0
ABC	1	1
distribue	1	1
des	3	2
fruits	2	0
les	1	0
sont	1	0
pommes	1	1
et	1	1
bananes	1	1

Tableau 3.1 Exemple de matrice terme-document

entre différents documents.

Les principales mesures de similarité sont les suivantes : la distance euclidienne, le cosinus et la corrélation de Pearson. Nous n'étendrons pas notre étude sur les mesures de similarité, mais pour plus de méthodes et de détails sur les méthodes d'extractions d'information et d'analyse de contenu veuillez vous référer au livre de Mr.Ricardo Baeza-Yates (Baeza-Yates et al., 1999).

La similarité dans l'espace vectoriel avec une représentation sac de mots part de l'hypothèse que l'ordre des mots ne compte pas et que la présence d'un mot détermine le contenu sémantique du texte. Ainsi, les descriptions des entreprises sont représentées par des vecteurs. Ces vecteurs, inscrits dans un espace vectoriel, donnent lieu à des mesures de similarités.

3.1.2 La mesure de similarité cosinus

$$\text{SimCos}(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \tag{3.1}$$

Dans la figure 3.1 nous présentons la distance entre les points A, B et C selon la similarité cosinus et la distance euclidienne. Le calcul de cosinus normalise le résultat par rapport à la longueur des documents respectifs. Tandis que la valeur du produit scalaire tend à augmenter proportionnellement à la longueur d'un document, le cosinus, quant à lui, pénalise les documents de longueur trop importante, en favorisant la proportion relative de termes communs. Autrement dit, plus l'angle est fermé, plus les documents se trouvent à proximité.

Prenons les données du tableau 3.1 pour calculer la similarité cosinus entre les 2 documents :

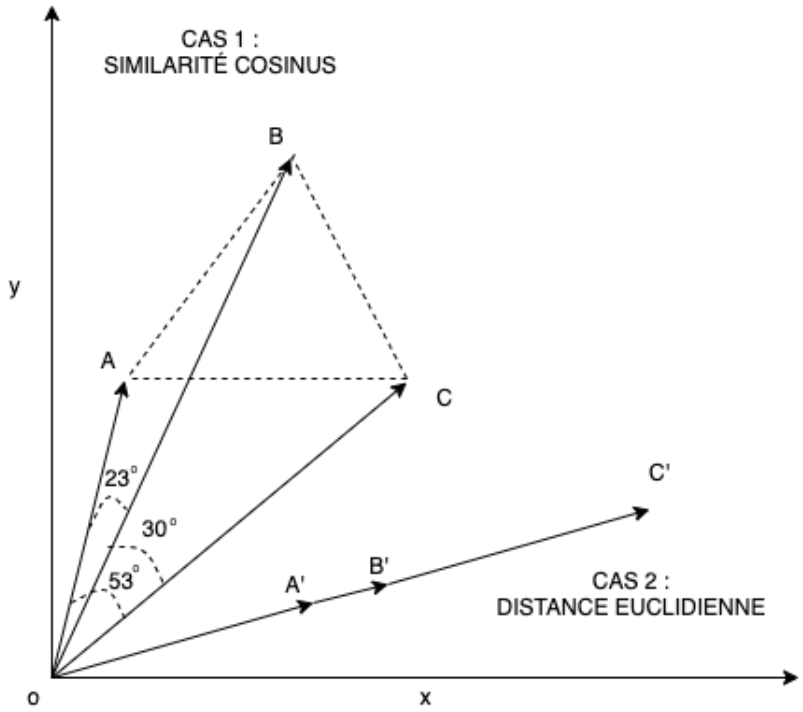

Figure 3.1 Similarité cosinus et distance euclidienne

le vecteur du document 1 est : [1,1,1,1,3,2,1,1,1,1,1]
le vecteur du document 2 est : [0,0,1,1,2,0,0,0,1,1,1]

La similarité cosinus est :

$$\text{SimCos}(\text{D}_1, D_2) = \frac{\sum_i D_1 D_2}{\sqrt{\sum_i {D_1}_i^2}\sqrt{\sum_i {D_2}_i^2}}$$

$$= 0.7817$$

3.1.3 La distance euclidienne

La distance euclidienne de deux vecteurs est une distance géométrique bien connue. Dans le contexte des similarités vectorielles, elle est utilisable comme la similarité cosinus, cependant chaque mesure de similarité est appropriée pour certains contextes.

La figure 3.1 illustre deux situations où la différence entre le cosinus et la distance euclidienne est apparente. On voit que pour des vecteurs colinéaires, la distance euclidienne permet de déterminer correctement que A est plus près de B que ne l'est C (Cas 2), alors que le cosinus est à 1 pour tous les pairs. Aux contraires, pour le cas 1, la distance euclidienne est égale,alors que le cosinus révèle une proximité plus grande entre A et B qu'entre B et C.

Prenons les données du tableau 3.1 pour calculer la distance euclidienne entre les 2 documents :
le vecteur du document 1 est : [1,1,1,1,3,2,1,1,1,1,1,1]
le vecteur du document 2 est : [0,0,1,1,2,0,0,0,1,1,1]

La distance euclidienne est :

$$\text{Dist}(\text{D}_1, D_2) = \sqrt{\sum (D_{1i} - D_{2i})^2}$$
$$=3$$

3.1.4 La corrélation de Pearson

Le coefficient de corrélation de Pearson, pouvant aller de -1 à 1, est une autre mesure de similarités fréquemment employée. Cette mesure de similarité entre 2 vecteurs se calcule selon la formule suivante :

$$Corr(D_1, D_2) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \tag{3.2}$$
$$= SimCos(x - \bar{x}, y - \bar{y}) \tag{3.3}$$

Comme la formule 3.3 le démontre, on remarque qu'il y a un lien entre la corrélation de Pearson avec la mesure de similarité cosinus. En effet, la corrélation de Pearson revient à mesurer la similarité cosinus en centrant les occurrences des mots par rapport à la moyenne.

Prenons les données du tableau 3.1 pour calculer la corrélation de Pearson entre les 2 documents présentés en exemple dans la section :
le vecteur du document 1 est : [1,1,1,1,3,2,1,1,1,1,1,1]
le vecteur du document 2 est : [0,0,1,1,2,0,0,0,1,1,1]

Le coefficient de corrélation est :

$$\text{Corr}(\text{D}_1, D_2) = \frac{\sum_i (D_{1i} - \bar{D_1})(D_{2i} - \bar{D_2})}{\sqrt{\sum (D_{1i} - \bar{D_1})^2} \sqrt{\sum (D_{2i} - \bar{D_2})^2}} = 0.4795$$

3.2 Le traitement de la langue naturelle

Le traitement de la langue naturelle ou en anglais connus sous le nom de «Natural language processing» a débuté dans les années cinquante et représente un domaine informatique de compréhension de la langue. Ce domaine se départage en subdivisions. Il rassemble une multiplicité de sous-domaines technologiques pour un spectre varié d'applications (Nadkarni et al., 2011). L'association de la computation linguistique définit la computation linguistique comme l'étude scientifique de la langue d'une perspective d'ordinateur. Le traitement du langage naturel (NLP) est un domaine informatique qui cherche à rendre le langage naturel compréhensible pour un ordinateur et d'en ressortir des utilités (Mitkov, 2003).

Les applications les plus utilisées de ce domaine sont : l'extraction d'information, la traduction automatique, les synthèses et résumés de textes, la recherche et enfin, les interfaces entre humains et ordinateurs (Collobert and Weston, 2008). À ce sujet, la littérature rapporte des recherches structurées ainsi : les chercheurs considèrent que la compréhension sémantique est encore un objectif lointain, et que pour l'atteindre, une approche de subdivision de la complexité serait nécessaire. La compréhension sémantique comprend les principaux sous-domaines suivants : l'étiquetage sémantique, l'extraction des entités nommées et la résolution anaphore (Xing et al., 2018).

Par ailleurs, la recherche de textes similaires est particulièrement active. À cet égard, les principaux domaines de recherche s'appliquent à la collecte d'informations, la classification de textes, de regroupements de documents, de recherche de sujets dans un texte, de formulations de réponses ou de questions, ou encore de résumés, de traductions et bien d'autres sous-applications (Gomaa and Fahmy, 2013). Nous concentrerons notre travail sur la recherche de textes similaire par modèle sac de mots combiné à la transformation TF-IDF, par l'utilisation de graphique de connaissances «Knowledge-graph», par des méthodes de filtrage et de réduction de textes et par les plongements lexicaux.

3.2.1 Transformation TF-IDF

Une méthode de représentation vectorielle de texte est la représentation par sac de mots (bag-of-words) expliqué dans la section 3.1.1. Lors de la recherche d'informations sur les compagnies, il est primordial de moduler le poids des termes en fonction de leur fréquence, ce dont la transformation TF-IDF permet d'obtenir.

En effet, les termes ne s'équivalent pas tous. Il faut donner un poids aux mots en fonction de la pertinence du mot entre les documents. Si un déterminant est utilisé dans chaque document, il n'est pas très pertinent pour différencier la similarité entre les documents d'où son poids

sera réduit.

Plus un terme est rare plus sa cooccurrence entre les documents est un indice de similarité, bien qu'il puisse exister de nombreux contre-exemples à cette logique. L'attribution d'un poids aux termes permet de mettre en ordre en vertu des termes les plus importants de chaque description.

Afin d'attribuer un poids pour chaque terme, nous mobilisons la fréquence inverse du document. Voici, donc, comment s'érigent les formules pour le calcul de la fréquence inverse du document :

$$\text{IDF}(T_i) = \log \frac{D}{D_i} \tag{3.4}$$

$$\text{TF-IDF}(T_{x,i}) = \text{TF}(T_{x,i}) \times \text{IDF}(T_i) \tag{3.5}$$

où :

$T_{x,i}$: Fréquence du terme i dans le document x
D : Nombre de documents
D_i : Nombre de documents avec le terme i
$\text{TF}(T_{i,x})$: Fréquence du terme i dans document x
$\text{IDF}(T_i)$: Fréquence inverse du terme i dans les documents

La méthode TF-IDF est utilisée sur la matrice terme-documents. Prenons comme exemple le tableau 3.1.1. Nous cherchons à transformer cette matrice avec la méthode TF-IDF. Commençons avec la première ligne avec le mot "la" qui est présent une fois dans le document 1 et 0 fois dans le document 2 :

$$\text{IDF}(T_i) = \log \frac{2}{1} = 0.301 \tag{3.6}$$

$$\text{TF-IDF}(T_{x,i}) = \frac{1}{14} \times 0.301 = 0.021 \tag{3.7}$$

Ainsi pour touts les termes des documents en exemple, nous obtenons la matrice dans la figure 3.2.

On remarque que dans le tableau 3.2, comme nous avons un corpus de deux documents, seuls les mots qui sont présents uniquement dans le premier document ont une valeur. Dans notre exemple on utilise des mots présents dans le premier document pour le deuxième document d'où toutes les valeurs sont nulles pour le document 2.

MOT	Document 1	Document 2
la	0.021502	0
compagnie	0.021502	0
ABC	0	0
distribue	0	0
des	0	0
fruits	0.043004	0
les	0.021502	0
sont	0.021502	0
pommes	0	0
et	0	0
bananes	0	0

Tableau 3.2 Exemple de matrice terme-document avec transformation TF-IDF

3.2.2 Réduction de vocabulaire, lemmatisation et racinisation

Pour des compagnies comparables, nous cherchons concrètement les similarités de leurs produits, de leurs opérations, de leurs marchés géographiques et de leurs modèles d'affaires. La réduction du vocabulaire permet d'éviter les matrices creuses qui nuisent aux résultats. Une première approche consiste à éliminer les termes courants tels que les déterminants ou les mots dénués d'importance, généralement appelés mots vides ou en anglais, «stop-words». Ces mots peuvent induire un bruit nuisible.

Deux autres façons de réduire le vocabulaire sont la racinisation et la lemmatisation. Ces méthodes sont couramment utilisées afin d'éliminer la conjugaison, les majuscules et pluriels des mots. D'une part, la racinisation est l'algorithme par lequel la forme radiale des mots est obtenue, selon des règles de modifications permettant de retrouver la même idée associée à différents mots de même forme radiale. D'autre part, la lemmatisation transforme les mots en leur forme canonique, grâce à une base de données de forme canonique des mots.

La racinisation et la lemmatisation permettent de réduire la dimension des vecteurs de fréquence de termes, car elle évite d'avoir une dimension pour chaque forme grammaticale dans lequel le mot pourrait être écrit, par exemple au pluriel. L'utilisation de cette méthode de traitement de texte permet de réduire la distance entre des documents qui auraient le même mot selon différentes formes, car elles seraient considérées comme deux mots différents. Par exemple si un document contient le mot «machine», à la forme au singulier et un autre document contient le mot «machines» à la forme au pluriel, la forme canonique des deux mots est «machine» permettant de trouver une similarité entre les deux documents contenant tous les deux la forme canonique «machine». Les autres méthodes utilisées dans notre projet sont

présentées ci-dessous afin de filtrer les descriptions des entreprises. Nous utiliserons l'étiquetage de mots afin de réduire le vocabulaire et ne retenir que les noms. Nous verrons que cette pratique améliore sensiblement les résultats.

3.2.3 L'étiquetage des mots

L'étiquetage consiste à associer à chaque mot une étiquette d'information telle que le type de mot. Il y a principalement neuf classes grammaticales de mots communément utilisées dans une phrase. Certaines classes sont variables et d'autres sont invariables. D'une part, les classes variables regroupent les noms, les déterminants, les adjectifs qualificatifs, les pronoms et enfin, les verbes. D'autre part, les classes invariables regroupent les adverbes, les prépositions, les conjonctions, les pronoms et les interjections.

En faisant usage des fonctions trouvées dans la librairie NLTK de python (Bird et al., 2009), nous pouvons utiliser la base de donnés du «Part-Of-Speech tagger», laquelle est déjà entraînée, conditionnée à reconnaître et à classifier les mots d'une phrase. Autrement dit, elle contient déjà, à même sa base de données, les classes de mots évoquées (classes variables et invariables). Le «Part-Of-Speech tagger», appliqué aux descriptions des entreprises, permet de réduire ces derniers à un contenu strictement essentiel et retenu par la sélection de certaines classes uniquement tels que les noms et les verbes.

Dans l'exemple à la figure 3.2, un modèle de filtre de mots est présenté. Ce dernier extrait l'information importante pour son analyse, en utilisant l'étiquetage des mots. L'exemple étiquette les «Tweets», des messages à 140 caractères selon le type de mot (Nom, verbe, autre) et fait une recherche du mot dans un corpus de mot-clé pour retrouver des «Tweets» similaires.

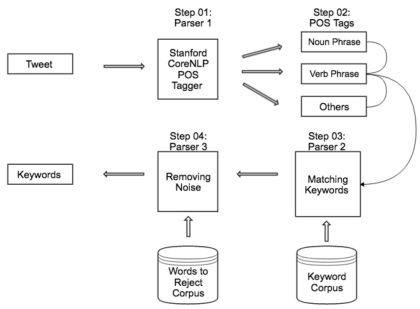

Figure 3.2 Méthodes d'extraction d'information des tweets utilisant le "Part-Of-Speech tagger" (Weerasooriya et al., 2017)

3.2.4 L'étiquetage des fonctions grammaticales des mots

Nous savons que les mots d'une phrase n'appartiennent pas tous aux mêmes groupes grammaticaux. En effet, pour toute phrase, chaque groupe de mots mobilisé possède une fonction qui lui est propre, permettant à la phrase de transmettre une idée ou un message. Ces groupes grammaticaux sont les suivants : le sujet et le groupe sujet, le complément d'agent, les attributs du sujet et enfin, le complément. Chaque fonction grammaticale permet de saisir un vaste champ d'informations dont nous pouvons explorer les possibilités au moyen d'une sélection d'information appliquée aux phrases.

Sur le support des fonctions mises à disposition par la librairie NLTK de Python, nous pouvons mettre en exercice le filtrage de phrases par fonctions grammaticales (Manning et al., 2014). Avec cette librairie, nous pouvons retrouver les différentes fonctions grammaticales dans une phrase et filtrer les phrases si le sujet de la phrase est relié à ce qu'on veut retenir dans un paragraphe. Par exemple, dans le paragraphe « ABC Company is in the USA. It provides fruits. The USA is a country.» on retrouverait le sujet de chaque phrase qui sont «ABC Company» pour la première phrase, «It» qui réfère à «ABC Company» pour la seconde phrase et «The USA» pour la dernière phrase d'où on retiendrait uniquement les phases avec le sujet en référence «ABC Company».

3.2.5 Plongements lexicaux et apprentissage machine

Le plongement lexical est une représentation de mot par un vecteur de petite dimension. Il permet de réduire la dimension de l'espace vectoriel où chaque mot du corpus correspond à un nombre reel permettant de représenter des mots par des vecteurs proches lorsqu'ils ont des contextes similaires. Nous détaillerons uniquement certains modèles de plongement lexicaux pour nous limiter aux sujets en lien avec nos modèles d'expérimentation.

Mikolov et al. (Mikolov et al., 2013b) présente un modèle de plongement lexical appelé «Word2Vec». Le principe est d'utiliser un reseau de neurones sur un corpus de document. Il y a principalement deux types d'architecture, présentés dans la figure 3.3, utilisées pour créer le modèle de plongement lexical. La première architecture s'appelle «Continuous Bag Of Word (CBOW)» où le modèle cherche à predire un mot à partir des mots qui l'entourent dans une fenêtre «Window Size» de taille donnée. L'autre architecture s'appelle «Skip-Gram», qui va chercher à prédire les mots en contexte à partir d'un mot central. L'un des inconvenients de «Word2Vec» est qu'il ne tient pas compte du contexte du mot dans le document. Afin de remédier au problème d'ambiguïté, Le and Mikolov (Le and Mikolov, 2014) présentent un plongement «Doc2Vec» permettant de prendre en compte du contexte du document. Le

modèle «Doc2Vec» va generer une liste de vecteurs de document en plus de la liste de vecteur mots. Cela permet de ne pas avoir besoin de prendre la moyenne des vecteurs de mots pour créer le vecteur du document. La similarité entre deux documents est calculée avec leur vecteur inféré et la similarité cosinus. Pour notre recherche, nous entraînerons nos propres plongements sur les descriptions avec «Doc2Vec» et nous utiliserons le modèle «Continuous Bag Of Words» pour créer les vecteurs de documents et calculerons la similarité cosinus.

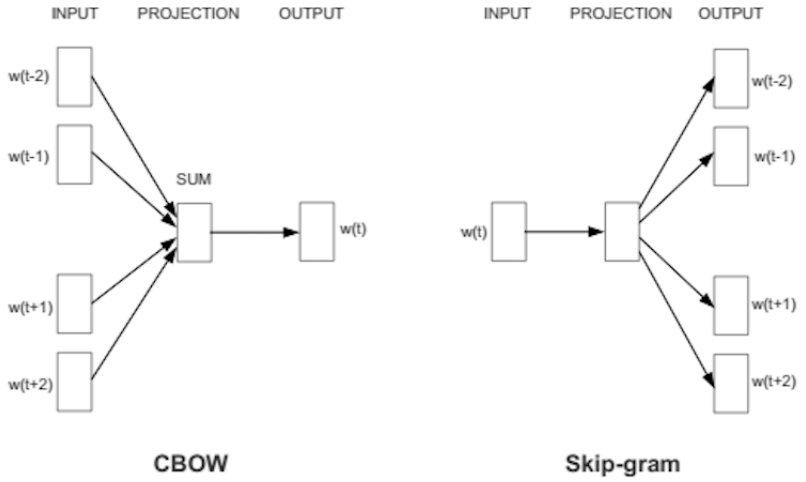

Figure 3.3 Deux Architectures du plongement lexical «Word2vec» (Mikolov et al., 2013a)

3.3 Graphique de connaissances «Knowledge-graphs»

Les graphiques de connaissances «Knowledge-graphs» existent dans la littérature depuis 1972 et sont devenus populaires depuis que Google à présenté en 2012 le «Google Knowledge graph» (Singhal, 2012). Cette base de donnée sous forme de graphique permet de faciliter la recherche. Par exemple, le nom «Taj Mahal» peut représenter un monument ou un endroit ou un musicien. Le graphique de connaissance permet de créer des relations et de tous les retourner lors d'une requête tel le Taj Mahal plutôt que chercher des mots. Un graphique abstrait de connaissance présente de nombreux avantages lorsqu'on le compare à une base de données relationnelle. Les graphiques de connaissances permettent une représentation abstraite de différents domaines dans le même graphique ou les arêtes représentent une relation entre les entités. Plus spécifiquement un graphique de données permet d'accumuler et de transférer le savoir du vrai monde où les nœuds représentent les entités d'intérêt et les arrêtent représente les relations entre les entitées (Hogan et al., 2020).
Nous ne nous étendrons pas sur les méthodes de création d'un graphique de connaissances,

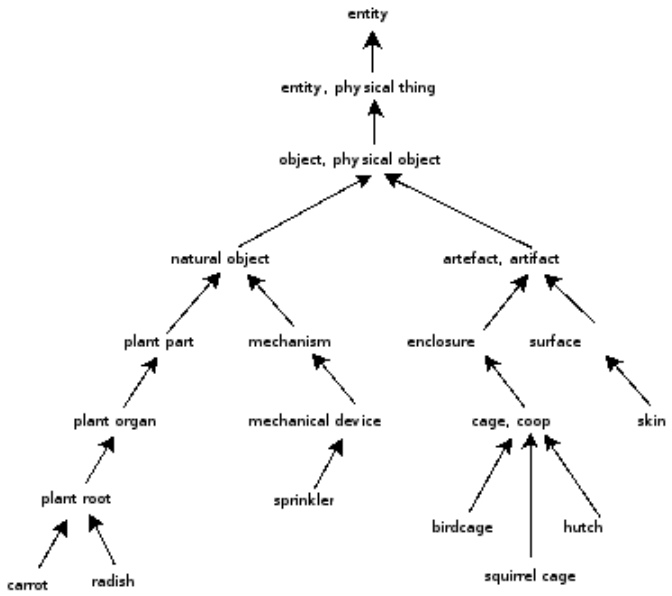

Figure 3.4 Représentation graphique du Wordnet (Princeton, 2007)

mais plus sur l'utilisation du graphique de connaissance WordNet (Fellbaum, 2012) permettant de trouver des similarités dans les descriptions des entreprises. Dans la figure 3.5 nous pouvons observer un exemple de représentation graphique et des abstractions pour le mot «carrot» qui est relié à «plant root» qui est relié à «plant organ» qui est relié à «natural object» et ainsi de suite. Dans notre recherche, nous utiliserons certaines fonctions de la librairie NLTK permettant d'importer et d'utiliser le graphique de connaissance très populaire Wordnet. Le Wordnet est une base de données lexicale qui a été développée par un laboratoire de l'université de Princeton. Le Wordnet permet de répertorier, de classifier et de mettre en relation le contenu lexical et sémantique de la langue anglaise. Le graphique de connaissance est composé de «synsets» ou synonyme qui permettent de dénoter un sens particulier a un mot ou group de mot dont nous présentons un exemple pour le mot «bank».

Nous utiliserons les graphiques de connaissances à plusieurs fins. La première utilité du graphique de connaissance Wordnet est la désambiguation des mots dans leur contexte avec la fonction «Lesk». Cette fonction permet de retourner la définition la plus probable du mot, dont voici un exemple d'utilisation dont les définitions possibles sont les suivantes :

```
Synset('bank.n.01') sloping land (especially the slope beside a body of water)
Synset('bank.n.03') a long ridge or pile
Synset('bank.n.04') an arrangement of similar objects in a row or in tiers
Synset('bank.n.06') the funds held by a gambling house or the dealer in some gambling games
Synset('bank.n.07') a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force
Synset('savings_bank.n.02') a container (usually with a slot in the top) for keeping money at home
Synset('bank.n.09') a building in which the business of banking transacted
Synset('bank.v.01') tip laterally
Synset('bank.v.02') enclose with a bank
Synset('bank.v.03') do business with a bank or keep an account at a bank
Synset('bank.v.04') act as the banker in a game or in gambling
Synset('bank.v.05') be in the banking business
Synset('deposit.v.02') put into a bank account
Synset('bank.v.07') cover with ashes so to control the rate of burning
Synset('trust.v.01') have confidence or faith in
```

Nous pouvons remarquer dans les définitions possibles précédentes tous les contextes dans lesquelles le mot «bank» peut faire référence. Parmi les définitions possibles, on remarque qu'on dénote le type de mots («n» pour nom et «v» pour verbe) et que des fois on retrouve des définitions de groupe de mots ou de mots différents («deposit» ou «saving_bank»). On remarque aussi que parfois le mot «bank» fait référence à un autre «synonyme set» tel que «Deposit.v.02» qui est défini avec «put into a bank». Avec la fonction «Lesk» on cherche le «synset» le plus plausible selon les mots des définitions dont on compare avec les mots de la phrase. Voici un exemple :

```
from nltk.wsd import lesk
sent = ['I', 'went', 'to', 'the', 'bank', 'to', 'deposit', 'money', '.']
print(lesk(sent, 'bank', 'n'))
> > >Synset('savings_bank.n.02')
```

Dans cet exemple, nous pouvons remarquer l'utilisation de la fonction «Lesk» permettant de

trouver la définition de «bank» dans le contexte de la phrase retournant «savings_bank.n.02».

À partir d'un graphique de connaissance, nous pouvons utiliser des méthodes de similarité entre les nœuds. Plusieurs notions de similarité existent dans la lexicographie du graphique de connaissance du Wordnet, dont les trois majeures basées sur les parcours dans le graphique de connaissance. Les trois principales sont la méthode de similarité de lch (Leacock et al., 1998), wup (Wu and Palmer, 1994) et le parcours par plus court chemin «path similarity».

La méthode de similarité de lch (Leacock et al., 1998) se calcule ainsi :

$$\mathrm{Lch}_{\mathrm{sim}}(a,b) = -\log \frac{D(A,B)}{2\mathrm{Dept}} \tag{3.8}$$

où :

D(A,B) : 1 + La plus courte distance entre le noeud A et le noeud B
Dept : Profondeur maximale de la taxonomie

La méthode de similarité de wup (Wu and Palmer, 1994) se calcule ainsi :

$$\mathrm{Wup}_{\mathrm{sim}}(a,b) = \frac{2 * D(R_{gl}, R_{ab})}{D(a, R_{ab}) + D(b, R_{ab})} \tag{3.9}$$

où :

- D(R_{gl}, R_{ab}) : La distance entre le noeud racine global et le noeud racine commun entre le noeud a et le noeud b
- D(a,R_{ab}) : La distance entre le noeud A et le noeud racine commun entre le noeud a et le noeud b
- D(b,R_{ab}) : La distance entre le noeud B et le noeud racine commun entre le noeud a et le noeud b

La méthode de similarité selon le parcours dans le graphique «Path Similarity» se calcule ainsi :

$$\mathrm{Path}_{\mathrm{sim}}(a,b) = \frac{1}{D(A,B)} \tag{3.10}$$

pour la recherche de similarité entre les mots avec la fonction «Path_Similarity (wn.path_similarity(synset1,synset2, simulate_root=False)) » entre 2 «synset» comme présenté avec le code sui-

vant :

```
dog = wn.synset('dog.n.01')
cat = wn.synset('cat.n.01')
wn.path_similarity(dog, cat,simulate_root=False)
=0.2
```

3.4 Les systèmes de recommandation

Un contexte dans lequel la recherche de similarité est très répandue est dans le domaine des systèmes de recommandations. Les systèmes de recommandation visent à modéliser les préférences des utilisateurs dans un contexte où il y a plusieurs items tels que des cours, des vidéos, des articles de nouvelles. En d'autres termes, en fonction des contextes, ces prédictions mettent en application des algorithmes qui suggèrent les items conformes aux préférences des utilisateurs. Le principe de base est de trouver des items similaires ou utilisateurs similaires pour prédire les préférences. Parmi ces techniques, les plus exploitées sont les suivantes : les filtres collaboratifs, les filtres basés sur le contenu et les méthodes de forage de données (Lops et al., 2011). Ce thème est très étudié dans le domaine des systèmes de recommandation et nous le présenterons brièvement, mais il déborde le cadre de cette recherche et nous ne nous étendrons davantage sur une approche basée sur du contenu textuel.

Les premiers systèmes de recommandations basées sur le contenu textuel utilisaient des méthodes simples telles que la représentation par sac de mots. D'autres systèmes de recommandations récents combinent plusieurs techniques ensemble qui sont connues sous le nom de méthodes hybrides. Les méthodes hybrides permettent de surmonter les limitations de l'utilisation d'une seule méthode. Par exemple, avec la combinaison de méthodes basées contenues et des méthodes de filtres collaboratifs, on parvient à obtenir de meilleures recommandations (Sun et al., 2019). D'autres méthodes plus récentes utilisent des graphiques de connaissances de multiples domaines «Knowledge Graphs» (Ehrlinger and Wöß, 2016). Ces systèmes de recommandations utilisent comme information autre que le contenu les graphiques de connaissances et le filtrage collaboratif permettant d'avoir des recommandations d'items plus précises. Nous utiliserons des approches de recherche de contenu et de l'utilisation de graphique de connaissance.

3.5 Récapitulation

Nous avons détaillé le concept des systèmes de recommandation général et l'avons adapté à notre cas d'utilisation pour en retrouver les besoins telles la recherche d'information et la réduction de texte. Ainsi avons-nous explicité, les principales méthodes et librairies utilisées dans notre objectif d'analyse textuelle tel que TF-IDF, les plongements lexicaux et les graphiques de connaissances. C'est donc à la lumière de ce qui précède que s'amorce notre prochain chapitre, «l'état de l'art de la recherche de compagnies comparables». Avant de présenter nos modèles d'algorithmes et nos techniques pour la recommandation de compagnies, nous étudierons les méthodes les plus utilisées par les analystes financiers.

CHAPITRE 4 L'ÉTAT DE L'ART DE LA RECHERCHE DE COMPAGNIES COMPARABLES

Les chapitres précédents présentent les concepts clés de la valorisation d'entreprises ainsi que le contexte de travail de l'analyste. Par la suite, dans le but de présenter les bases d'un modèle d'analyse de données pour la recherche de compagnie comparable, nous avons explicité les techniques de recherche d'information et les méthodes d'analyse de texte. Dans ce chapitre, nous présentons les enjeux de la recherche de compagnies comparables et les recherches en lien avec la recherche de similarité pour les entreprises. Dans les chapitres suivants, nous présenterons les algorithmes proposés et les résultats de notre expérimentation.

4.1 Données et méthodes utilisées dans la recherche par la similarité

Dans les sous-sections qui suivent, nous présenterons les principales méthodes, données et algorithmes qui ont été étudiés. Nous débuterons par les sujets sur la recherche de compagnies similaires qui montrent les différentes données qui ont été utilisées. Par la suite, nous présenterons les méthodes de similarités qui ont été les plus utilisées.

4.1.1 Les sources de données des entreprises

Dans les articles des sujets reliés à la similarité des entreprises, les chercheurs ont exploré différentes sources de données. Les chercheurs ont étudié les activités et les décisions des analystes financiers du côté de la vente par rapport à leur sélection de compagnies similaires «Peers» à partir de leurs rapports de vente «Sell-side reports». Dans leurs études ils concluent que les analystes choisissent des compagnies dont les similarités sont sur le plan des tailles, de la croissance, de la dette, de la classification d'industrie, de la volatilité ou du volume de transaction de leurs actions (De Franco et al., 2015). De ces conclusions nous avons inclus la classification d'industrie et la taille avec la capitalisation, car l'ajout de ces données dans nos modèles nous permet de créer un modèle général de recommandation de compagnies comparables sur lequel des travaux futurs pourrons s'ajouter au modèle pour des contextes précis tels que des compagnies avec un haut taux de croissance ou un taux de dette comparable.

Parmi les travaux en lien avec notre sujet, plusieurs se concentrent sur la comparaison des multiples et des comparaisons de ratios financiers (Dittmann and Weiner, 2005). Alford (Alford, 1992) en est un bon exemple, dans la mesure où il propose des méthodes de sélection de

compagnies similaires en se basant sur les trois premiers chiffres des codes SIC des secteurs d'activités et de différents indicateurs de croissance et des risques — une démarche qui, selon lui, conduit à de bons résultats de comparables.

Les travaux précédents présentent des modèles pour classifier les compagnies dans les standards de classification qui existent qu'ils valident avec les données des standards de classifications préexistants, ce qui diffère de notre travail, car nous présentons des modèles pour retrouver des compagnies comparables et validons selon leur rappel avec nos données cibles des comparables pour chaque entreprise. Gkotsis (Gkotsis et al., 2018) pour le domaine de la technologie utilise les brevets des entreprises pour retrouver celles qui sont en compétition avec une méthode de regroupement «K-means clustering» et «Ward's criterion». Plus précisément ils comparent les résultats de regroupement de leurs modèles de regroupement et leur modèle avec le critère de Ward (Ward Jr, 1963) en utilisant une représentation vectorielle des classifications des brevets associés à une entreprise selon leur «International Patent Classication (IPC)» et valident avec le «Industry classification benchmark ICB» et concluent qu'avec 38 regroupements ils peuvent définir les marchés technologiques dans lesquelles les entreprises sont en compétition.

Taeyoung Kee a présenté dans son mémoire de maîtrise (Kee et al., 2018) une méthode de classification des entreprises utilisant une liste de vecteurs de mots entraînée avec «Word2Vec» sur des articles des nouvelles sur une période de 10 ans pour classifier les entreprises dans le GICS. Son travail présente une comparaison entre la classification du GICS des entreprises du SP 500 et de la classification par similarité cosinus entre le nom de l'entreprise et le nom de la classification d'industrie dans la liste de vecteurs préentraînée. Par exemple, on calcule la similarité cosinus entre le vecteur du mot «Facebook, inc» avec les mots des classifications d'industrie comme «Information Technology» et «Consumer Staple» pour identifier la classification du GICS la plus proche. Dans ses résultats, 43% des compagnies sont classifiées dans le même groupe d'industrie du GICS.

4.1.2 Les algorithmes de représentation sémantique

Dans la littérature, plusieurs recherches en lien avec des méthodes alternatives de l'utilisation des standards de classification des entreprises ont été effectuées afin de classifier les entreprises. Parmi les travaux de recherche, Taeyoung Kee a présenté (Kee et al., 2018) (Kee, 2019) une méthode de classification des entreprises dans le GICS en utilisant un modèle «Word2Vec» déjà préentraîné sur des articles de nouvelles sur 10 ans jusqu'à l'an 2012. Ils utilisent la similarité cosinus pour déterminer la ressemblance entre le mot «Facebook, inc» et «Google, Inc». Il effectue ce calcul pour toutes les entreprises permettant de déterminer

les compagnies les plus similaires. Leur méthode est très limitée, tout d'abord elle dépend des donnés présentes dans le «Word2Vec» pré-entraîné sur 10 ans jusqu'à l'an 2012 ce qui n'est pas accessible pour des années plus récentes et de nouvelles compagnies qui n'existaient pas ou qui ne seraient pas dans les articles de nouvelles.

L'inconvénient de la méthode «Word2Vec» est qu'il ne tient pas compte du contexte du document et que cela crée des imprécisions sur des mots qui sont employés dans différents contextes, ce dont le modèle «Doc2Vec» permet d'éviter, car le modèle va générer des vecteurs des listes des documents et des vecteurs des listes de mots. En effet, chaque vecteur document permettra de représenter le sens général et sera associé avec les vecteurs des mots du document. Les limitations de ce modèle est qu'il nécessite un long entraînement et optimisation des paramètres. L'entraînement de ce modèle doit aussi être répété à chaque fois que nous avons de nouvelles descriptions des entreprises afin de représenter les vecteurs des documents et des mots, car leurs dimensions changent en fonction du nombre de mots ou documents présent. Pour plus de détail sur «Word2Vec» et son fonctionnement, vous pouvez vous référer à l'article de Mikolov (Mikolov et al., 2013b) sur le sujet.

Dans notre recherche nous utiliserons des données différentes afin de retrouver des entreprises comparables. En effet, nous utiliserons une description pour chaque entreprise dont nous explorerons la méthode «Doc2Vec» qui est une méthode plus récente du «Word2Vec». Nous entraînerons nous-mêmes l'espace vectoriel et n'utiliserons pas des modèles préentraînés afin d'avoir un espace vectoriel adapté au langage de la finance pour les descriptions des entreprises. Nous détaillerons ces modèles dans la section 5.6 de notre mémoire. En raison de l'inaccessibilité des données utilisées dans les recherches de Taeyoung Kee (articles de nouvelles des entreprises sur 10 ans) et de notre choix d'utiliser moins de données et des données différentes (descriptions des entreprises, groupe d'industrie, secteur d'activité et capitalisation des entreprises), des entreprises qui sont différentes dans nos données, il ne nous est pas possible de comparer nos résultats, mais il serait possible d'utiliser leur mesure de similarité pour remplacer notre utilisation de groupe d'industrie et de secteur d'activité qui sont issus de standards de classification ou d'ajouter la mesure dans nos algorithmes. De manière générale, les deux recherches de Taeyoung Kee assument un lien entre les classifications des entreprises et le contenu des articles de nouvelles qui leur sont liées. Dans notre recherche, nous approchons la recherche de compagnies comparables avec des données différentes et plus structurées dans le sens que nous avons une description non structurée, une classification d'industrie, d'une classification de secteur d'activité et une capitalisation pour chaque entreprise.

Parmi les recherches avec des algorithmes de similarité, un article présente des méthodes

d'analyse de similarité dans un espace vectoriel (matrice termes-documents) avec transformation TF-IDF sur les rapports annuels des entreprises (Hoberg and Phillips, 2016). En effet, ils présentent une méthode d'analyse des textes avec TF-IDF avec la mesure de similarité cosinus avec les noms et noms propres dans les rapports annuels «10-k» sur plusieurs années (de 1997 à 2008, total de 50,673 rapports annuels) à des fins de classification d'industrie, recherche de compétiteurs et de produits similaires. Ils regroupent les rapports annuels des entreprises en 1997 selon leur classification d'industrie des entreprises pour déterminer les valeurs cibles des classifications. Par exemple si le rapport annuel des compagnies «MMM » et «GE» en 1997 est dans le secteur d'activité «Industrial Goods» alors pour les années suivantes, les rapports annuels similaires à ce regroupement seront classés dans «Industrial Goods» et comparés à leur vraie classification de secteur d'activité. Nous utiliserons aussi la transformation TF-IDF et filtrerons les noms et noms propres parmi nos modèles de filtre de description des entreprises. Cependant, il ne nous est pas possible de comparer nos résultats avec leur recherche, car les mesures et les données utilisées pour la validation rendent les résultats non comparables.

Lee, Charles MC et al. (Lee et al., 2015), étudient l'historique des recherches des compagnies sur le site financier pour la recherche de compagnies similaires «Peers». Ils étudient la relation entre les compagnies cherchées par les analystes et la chronologie de recherche par les utilisateurs du site gouvernemental des données des entreprises publiques «EDGAR». Leurs résultats de recherche montrent que par rapport à la chronologie de recherche des entreprises dans le site «EDGAR» par un utilisateur, les entreprises recherchées présentent des similarités entre elles. Par exemple, dans leur recherche ils arrivent à retrouver qu'entre les compagnies cherchées dans «EDGAR» par des analystes 49% ont les mêmes 6 premiers chiffres sur les 8 chiffres de classification du GICS.

Raman et al. (Raman et al., 2019), se concentrent sur l'utilisation des réseaux de neurones avec des mécanismes d'attention «neural attention mechanism» pour construire un graphique de relations entre les entreprises à partir de certaines informations telles que les revenues des entreprises, leur EBITDA, l'industrie dans laquelle la compagnie opère et le score de crédit de l'entreprise. Ils appliquent leurs modèles en entraînant des données des articles de nouvelles des entreprises. Ils étudient principalement les relations entre les entreprises — tels que compétiteurs ou partenaire.

De manière générale, plusieurs recherches ont été effectuées pour automatiser la classification dans les systèmes de classifications existantes ou pour faciliter la recherche de compagnie similaire. Ces techniques avancées utilisent soit des textes, soit des chiffres. Nous pensons qu'un modèle d'analyse qui inclut les classifications d'industrie, les descriptions et la capitalisation

permettrait de retrouver des comparables. Nous utilisons plusieurs algorithmes de traitement de la langue sur la description des entreprises et comparons nos modèles avec une base de données libellée provenant par un expert du domaine pour des entreprises comparables pour la valorisation des entreprises.

CHAPITRE 5 MODÈLES POUR LA RECHERCHE DE COMPAGNIES PROPOSÉS

Ce chapitre présente les modèles que nous avons développés. Ce qui nous conduira au chapitre suivant, dans lequel seront exposés nos résultats.

Les méthodes proposées dans présent mémoire seront des combinaisons provenant de donnée des catégories d'industries, de la description et de la finance avec différente manipulation sémantique sur les données des descriptions d'entreprises. La première méthode que nous présenterons concerne uniquement la catégorie d'industrie et la finance. Nous considérerons ce modèle comme référence, modèle de base auquel le rappel global sera comparé les autres modèles. Quant aux quatre autres modèles, ils incluent la description, les noms dans la description, les verbes dans la description et avec l'utilisation du graphique de connaissance «WordNet».

Dans notre travail, nous avons expérimenté avec la similarité cosinus et la corrélation de Pearson, mais nous avons présenté uniquement les résultats selon la corrélation de Pearson. Nous avons utilisé la corrélation de Pearson, car notre mesure de performance est le rappel des entreprises et que les rappels selon la mesure de corrélation de Pearson et de similarité cosinus sont très similaires.

Voici une liste des modèles et leur acronyme respectif :

1. Modèle MCF de base des recommandations avec catégorie d'industrie et capitalisation
2. Modèle MCFD avec catégorie, capitalisation et description avec tout les mots
3. Modèle MCFD_N avec catégorie, capitalisation et description avec mots étiquetés (Noms)
4. Modèle MCFD_V avec catégorie, capitalisation et description avec mots étiquetés (Verbes)
5. Modèle MCFD_{GV} avec catégorie, capitalisation et description avec groupes verbaux (Verbe et complément)
6. Modèles MCFDWN avec catégorie, capitalisation, description et graphique de connaissances «WordNet»
7. Modèles D2V avec vecteur de descriptions «Doc2Vec»

5.1 Modèle MCF de base des recommandations avec catégorie d'industrie et capitalisation

Le premier modèle repose sur un algorithme de base. Ce modèle s'appuie essentiellement sur la capitalisation ainsi que la catégorisation (secteur d'activité et groupe d'industrie). Nous présentons uniquement les méthodes selon la mesure de corrélation de Pearson, car nous avons observé un taux de rappel similaire qu'avec l'utilisation de la similarité cosinus.

Le score de recommandation prend ainsi la formule suivante :

$$R_1 = R_{cat} + R_{cap} \tag{5.1}$$

où :

R_1 : Score de recommandation catégorie d'industrie et capitalisation. $R_1 \in [-1, 2]$
R_{cat} : Score de recommandation selon TF-IDF des catégories d'industries
$R_{cat} \in [-1, 1]$
R_{cap} : Score de recommandation selon capitalisation des entreprises. $R_{cap} \in [0, 1]$

Pour R_{cat}, les mots décrivant la catégorie d'industrie et le secteur d'activité d'une entreprise sont regroupés dans la catégorie d'industrie pour y appliquer une transformation TF-IDF et calculer la similarité selon la corrélation de Pearson détaillé dans la section 3.1.4. Tout d'abord, nous créons un espace vectoriel avec les mots du secteur d'activité et de la catégorie d'industrie. Par exemple, pour 3D Systems Corporation présentées à la section 6.1.1 , son secteur d'activité est «Technology» et le groupe d'industrie est «Computer peripherals». Nous regroupons les mots du secteur d'activité avec le groupe d'industrie, tel que «technology computer peripherals», et créons l'espace vectoriel pour ensuite y appliquer une transformation TF-IDF et calculons la similarité selon la corrélation de Pearson entre les compagnies. Dans les autres modèles avec la description de la compagnie, nous créons un espace vectoriel séparé pour les descriptions avec une transformation TF-IDF pour calculer la similarité entre les descriptions selon la corrélation de Pearson.

Pour R_{cap}, la normalisation logarithmique des capitalisations des entreprises est faite selon l'équation suivante pour retrouver des capitalisations avec la plus petite distance entre elles :

$$R_{capAX} = 1 - D_{capAX} = 1 - |\frac{log(F_X) - log(F_A)}{log(F_{MAX})}| \tag{5.2}$$

où :

D_{capAX} : Distance de capitalisation entre Compagnie A et Compagnie X. $R_{capAX} \in [0, 1]$
R_{capAX} : Similarité de capitalisation entre Compagnie A et Compagnie X. $R_{capAX} \in [0, 1]$
F_X : Capitalisation associée à compagnie X
F_A : Capitalisation associée à compagnie A
F_{MAX} : Capitalisation maximale parmi toutes les compagnies

Quant à la combinaison, nous attribuons la même pondération pour toutes les sources de données, mais nous allons présenter une section sur nos résultats lors de l'optimisation des poids.

5.2 Modèle MCFD avec Catégorie, Capitalisation et Description

Le second modèle ajoute la description. Le score de recommandation des compagnies comparables prend la somme de la similarité de la catégorisation (groupe d'industrie, secteur d'activité), la similarité de la capitalisation et la similarité des descriptions brutes, comme l'expose l'équation suivante :

$$R_2 = R_1 + R_{des} \tag{5.3}$$

où :

R_2 : Score de recommandation selon catégorie d'industrie, capitalisation et description. $R_2 \in [-2, 3]$
R_1 : Score de recommandation catégorie d'industrie et la capitalisation. $R_1 \in [-1, 2]$
R_{des} : Score de recommandation descriptions TF-IFD. $R_{des} \in [-1, 1]$

Pour R_{des}, le score de recommandation des descriptions est calculé selon la transformation TF-IDF et la corrélation de Pearson dans un espace vectoriel comme pour la catégorisation, la transformation TF-IDF est expliquée à la section 3.2 et la corrélation de Pearson dans la section 3.1.4. Les manipulations ont été effectuées avec Python et la librairie NLTK (Chen, 2009). Notons par ailleurs qu'aucune réduction de dimension n'a eu lieu, nous gardons les mots tels qu'ils sont écrits dans ce premier modèle afin de garder toutes nuances possibles entre les mots.

5.3 Modèle MCFD_N avec calcul TF-IDF avec mots étiquetés (Noms)

Le 3ème modèle est identique à celui de la section précédente à l'exception que nous y utilisons uniquement les noms dans les descriptions avec un filtrage de mots, exécuté au moyen d'une méthode d'étiquetage lexicale.

$$R_3 = R_1 + R_{nom} \tag{5.4}$$

où :

R_3 : Score de recommandation selon catégorie d'industrie, capitalisation et description. $R_3 \in [-2, 3]$
R_1 : Score de recommandation catégorie d'industrie et la capitalisation. $R_1 \in [-1, 2]$
R_{nom} : Score de recommandation descriptions TF-IDF NOMS. $R_{nom} \in [-1, 1]$

Pour R_{nom}, le score de recommandation des noms des descriptions est calculé de la même manière que pour R_{des} qui est expliqué à la section 5.1. La réduction de dimension par la sélection des mots de type noms a été effectuée en vertu de l'étiquetage POS «Part-Of-Speech» dans la librairie NLTK (Chen, 2009) de Python. Celui-ci différencie les classes grammaticales des termes. Nous avons retenu les mots avec les étiquettes «NN» pour les noms singuliers, «NNS» pour les noms pluriels, «NNP» pour les noms propres et enfin, «NNPS» pour les noms propres et pluriels.

5.4 Modèle MCFD_V avec calcul TF-IDF avec mots étiquetés (Verbes)

Le modèle MCFD_V, similaire à celui de la section précédente, utilise uniquement les verbes dans les descriptions avec un filtrage des mots, selon une méthode d'étiquetage lexicale.

$$R_4 = R_1 + R_{ver} \tag{5.5}$$

où :

R_4 : Score de recommandation selon catégorie d'industrie, capitalisation et description. $R_4 \in [-2, 3]$
R_1 : Score de recommandation catégorie d'industrie et capitalisation. $R_1 \in [-1, 2]$
R_{ver} : Score de recommandation descriptions TF-IDF verbes. $R_{ver} \in [-1, 1]$

Pour R_{ver}, la mesure de recommandation des noms des descriptions est effectuée de la même

manière que pour R_{des} qui est expliqué à la section 5.1.La sélection des mots de type verbes a été effectuée en vertu de l'étiquetage POS «Part-Of-Speech» dans la librairie NLTK (Chen, 2009) de Python. Avec la librairie, nous avons retenu les étiquettes de verbes «VB» pour les verbes de forme basique, «VBD» pour les verbes au passé, «VBG» pour les verbes au participe présent, «VBN» pour les verbes au participe passé, «VBP» pour les verbes au présent singuliers et enfin, «VBZ» pour les verbes à la fois au présent et au pluriel.

5.5 Modèle MCFD_{GV} avec calcul TF-IDF avec groupes verbaux (Verbe et Complément)

Le modèle MCFD_{GV}, similaire à la section précédente, utilise uniquement les groupes verbaux et les groupes complément des descriptions. Le calcul est le suivant :

$$R_4 = R_1 + R_{vc} \tag{5.6}$$

où :

R_4 : Score de recommandation selon catégorie d'industrie, capitalisation et description. $R_4 \in [-2, 3]$
R_1 : Score de recommandation catégorie d'industrie et capitalisation. $R_1 \in [-1, 2]$
R_{vc} : Score de recommandation descriptions TF-IDF groupe verbe et complément. $R_{vc} \in [-1, 1]$

Pour la sélection des verbes et compléments, nous utilisons le «Stanford typed dependencies representation» (De Marneffe and Manning, 2008). Ce dernier, retrace les relations grammaticales présentes dans les phrases avec lequel on retient les verbes et les compléments, selon les étiquettes de «ACOMP» pour complément d'adjectif, «CCOMP» pour complément clausal, «PCOMP» pour complément de préposition, «XCOMP» pour complément clausal ouvert, «DOBJ» pour objet direct, «IOBJ» pour objet indirect, «POBJ» pour objet de préposition, «COP» complément de verbe «copula» - lesquels représentent les différents types de compléments dans les arbres de liaisons dans les phrases des descriptions.

5.6 Modèle MCFDWN, avec calcul TF-IDF avec descriptions, descriptions désambiguées et calcul de similarité avec graphique de connaissance WordNet

Le modèle MCFDWN utilise le graphique de connaissances «WordNet» pour créer un sac de mots désambigués des noms et des verbes des descriptions brutes. Cela est effectué en utilisant la fonction de désambiguïsation de chaque mot avec le contexte de la phrase comme expliquée à la section 3.3 avec le mot «bank». Le modèle de cette section sera composé d'une combinaison du modèle 1ial avec catégories et capitalisation ainsi que le score de recommandation selon la description brute avec transformation TF-IDF avec corrélation de Pearson, le score de recommandation selon les noms et verbes des descriptions désambiguées avec transformation TF-IDF avec corrélation de Pearson et la similarité selon la méthode de similarité selon le plus court chemin dans le graphique de connaissance WordNet dont un exemple est présenté à la section 3.3. La formulation du modèle est le suivant :

$$R_5 = R_1 + R_{des} + R_{Wdis} + R_{Wp} \tag{5.7}$$

où :

- R_5 : Score de recommandation selon catégorie d'industrie, capitalisation descriptions brutes , noms verbes des descriptions désambiguées et similarité WordNet. $R_5 \in [-3, 5]$
- R_1 : Score de recommandation catégorie d'industrie et capitalisation. $R_1 \in [-1, 2]$
- R_{des} : Score de recommandation descriptions TF-IDF. $R_{Wdes} \in [-1, 1]$
- R_{Wdis} : Score de recommandation noms verbes désambiguées TF-IDF. $R_{Wdis} \in [-1, 1]$
- R_{Wp} : Score de recommandation de similarité entre les noms et verbes désambiguées des descriptions selon le graphique de connaissance WordNet. $R_{Wp} \in [0, 1]$

Pour R_{Wp}, l'utilisation de la fonction path_similarity() de WordNet permet d'utiliser la distance dans le graphique de connaissance partant d'un mot à un autre. Nous utiliserons cette fonction sur tous les mots entre deux descriptions permettant de déterminer la similarité. Voici un exemple permettant de mieux comprendre l'application dans notre contexte.

Prenons comme exemple 2 phrases en assumant que le poids de chaque mot est égal à 1 :
«I have a dog.»
«I have a cat.»

Pour chaque mot de la première phrase, nous allons chercher les noms et verbes les plus proches selon leur «synset». La première étape est de trouver le «Synset» accessible dans le

Wordnet avec la fonction de désambiguation comme avec l'exemple de «bank» :
Phrase 1 : Synset('have.v.01') Synset('dog.n.01')
Phrase 2 : Synset('have.v.01') Synset('cat.n.01')

Pour chaque Synset dans la phrase 1 nous allons chercher le Synset dans la phrase 2 le plus similaire. Synset('have.v.01') de la phrase 1 avec Synset('have.v.01') de la phrase 2 donne une similarité de 1. Synset('have.v.01') de la phrase 1 avec Synset('cat.n.01') de la phrase 2 donne une similarité de 0 car «have» est un verbe et «cat» est un nom. Pour le Synset «have» de la phrase 1 on prend le Synset «have» de la phrase 2 car c'est le mot avec la plus grande similarité (1). On refait la même manipulation pour Synset('dog.n.01'). Synset('dog.n.01') de la phrase 1 avec Synset('have.v.01') de la phrase 2 donne une similarité de 0. Synset('dog.n.01') de la phrase 1 avec Synset('cat.n.01') de la phrase 2 donne une similarité de 0.2 comme vue à fin de la section 3.3 donc on retient 0.2.

Nous avons 2 mots avec un poids que nous avons assumé égal à 1 dans notre exemple dont on fait la moyenne de similarité entre les 2 phrases avec 1 pour «have» et 0.2 pour «dog» ce qui nous donne 0.6 de similarité (((1*1)+(1*0.2)) /2) entre les phrases.

5.7 Modèle D2V avec plongements lexicaux des descriptions «Doc2Vec»

Pour le modèle D2V, nous présentons nos modèles avec plongement lexical sur les descriptions «Word2Vec» dont nous avons entraîné pour chacun. L'entrainement sur nos données permet d'avoir un plongement lexical adapté à notre corpus permettant d'être plus spécialisé aux descriptions des entreprises. Cet algorithme est la version pour un document de «Word2Vec» qui est un algorithme qui construit une distribution des représentations sémantique des mots. Nous explorons les plongements lexicaux dans notre recherche tout en connaissant les limites d'un tel modèle pour notre situation. En effet, l'entraînement d'un modèle de plongement lexical est un long travail qui nécessite du temps pour l'optimisation des paramètres d'entraînement, plusieurs longs essais d'entraînement et un travail qui doit être répété lorsque le corpus change. Pour plus de détail sur «Word2Vec» et son fonctionnement, veillez-vous référer à l'article de Mikolov (Mikolov et al., 2013b) sur le sujet.

Paramètres d'entraînement			
Taille de vecteur	epoch	alpha	window
20	10	0,025	5
20	50	0,025	5
30	10	0,025	5
30	50	0,025	5
60	50	0,025	3
60	50	0,025	2
60	80	0,025	2

Tableau 5.1 Modèles Doc2Vec avec taille du vecteur et hyperparamètres sans augmentation de données.

Dans le tableau 5.1, nous présentons des configurations du modèle «Doc2Vec» avec leur taille de vecteur et leurs hyperparamètres. La recommandation d'entreprises, selon la méthode «Doc2Vec» (Le and Mikolov, 2014), est effectuée sur Python avec la librairie Gensim (Řehůřek and Sojka, 2010) et la fonction d'entraînement de donnée. Selon la vectorisation de document, connue sous le nom de «Doc2Vec» plusieurs hyperparamètres ont été appliqués afin d'avoir une diversité de modèles de différentes tailles de vecteur. À partir de ces modèles, nous comparons les recommandations à l'oracle des compagnies comparables, selon la méthode du rappel. Nous appliquons ce procédé aux compagnies avec un rappel global sur les dix compagnies les plus similaires et sur les cent compagnies les plus similaires. Nous avons utilisé la librairie Gensim (Řehůřek and Sojka, 2010) qui est une librairie de Python contenant les fonctions permettant d'effectuer un entraînement selon le modèle «Doc2Vec» et d'optimiser les hyperparamètres.

5.7.1 Augmentation des descriptions

Pour les prochains modèles, nous avons augmenté les descriptions des entreprises à partir de plongements lexicaux des mots dans les descriptions. Dans le tableau 5.2, nous présentons les modèles «Doc2Vec» avec leur taille de vecteur et leurs hyperparamètres pour les modèles avec des descriptions augmentées avec «Word2Vec». En effet, nous avons augmenté les descriptions des données en ajoutant du vocabulaire, en nous basant sur un plongement de mots de toutes les descriptions.

Paramètres d'entrainement			
Taille de vecteur	epoch	alpha	window
20	10	0,025	5
20	50	0,025	5

Tableau 5.2 Modèles Doc2Vec avec taille du vecteur et hyperparamètres avec augmentation de données.

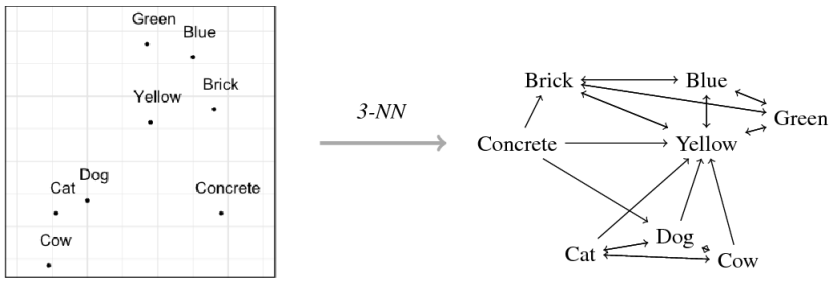

Figure 5.1 Exemple de Plongement lexical et mots voisins (Newman-Griffis and Fosler-Lussier, 1970)

Notre intuition sur l'augmentation des données des descriptions nous permet de croire que nous pourrions augmenter la similarité entre les descriptions si les mots qui y sont présents sont les mêmes. Cela voulant dire qu'une description avec le mot «Device» et une autre avec le mot «Hardware» pourraient être rapproché en les augmentant avec les qui sont proches dans l'espace vectoriel.

Nous avons créer un plongement des mots avec la fonction «gensim.models.Word2Vec()» de la librairie Gensim (Řehůřek and Sojka, 2010) sur le corpus de toutes les descriptions des entreprises avec la taille des vecteurs à 100 et le «window» à 5 avec un minimum d'au moins une fois le mot. Une fois le modèle entraîne, nous avons utilisé l'espace vectoriel des mots pour augmenter les descriptions. Dans la figure 5.1 nous pouvons observer que le mot «Dog» et «Cat» sont proches. Notre logique est la suivante. Si dans le plongement lexical

des descriptions le mot «Hardware» et «Device» ont une similarité cosinus arbitraire de au moins de 0.5, on ajoute le mot dans la description. Nous avons considéré 0.5 comme minimum de similarité cosinus, nous pouvons considérer ces mots comme des mots proches et les rajouter dans les descriptions respectives. D'où la description de «Device» deviendrait «Device Hardware» et la description «Hardware» deviendrait «Hardware Device» ce qui permettrait de retrouver des descriptions similaires qui au départ n'avaient aucun mot similaire. Cette tâche requière un temps de calcul considérable et dont on à fait quelques explorations et décidé de retenir 0.5 comme similarité cosinus minimum entre les mots pour les ajouter. Nous appliquons cette méthode sur tous les noms et les verbes des descriptions. Par la suite on ajoute dans les descriptions respectives les mots considérés similaires et entraînons les modèles «Doc2Vec» du tableau 5.2.

CHAPITRE 6 METHODOLOGIE

Cette section présente notre méthode de validation des modèles présentés au chapitre 5 en utilisant trois exemples de compagnies et de leur donnée ainsi que leurs valeurs cibles. Afin de valider nos modèles, nous les comparerons à un modèle de base et à un corpus de validation. Le total des entreprises de notre algorithme ainsi que de notre corpus de compagnies comparables cibles est composé de 2865 compagnies. Nous utiliserons le rappel afin de comparer nos résultats.

6.1 Exemples de compagnies comparables cibles

Nous allons illustrer nos méthodes et les résultats des modèles dans le chapitre 7 à partir de trois exemples de compagnies avec leur recherche de comparable. Précisons que leurs description, capitalisation, groupe d'industrie et secteur d'activité proviennent de la source suivante : *Intrinio US Fundamentals and Stock Prices*. Intrinio est une compagnie américaine qui offre plusieurs bases de données financières privées pour les analystes financiers. Nous utilisons les données de 2991 compagnies provenant de la base de données d'Intrinio.

Compagnie	Symbole	SIC Code	NAICS Code
3D Systems Corporation	NYSE :DDD	7372	333244
3M Company	NYSE :MMM	3841	339332
Aramark	NYSE :ARMK	5812	561990

Tableau 6.1 Exemples d'information des compagnies

Dans les trois prochaines sections nous présenterons les informations générales portant sur 3D Systems Corporation, une entreprise de fabrication d'imprimante 3D, 3M Company un conglomérat et Aramark une compagnie d'alimentation.

6.1.1 3D Systems Corporation

Description : *3D Systems Corp. is a holding company, which engages in the provision of three dimensional printing centric designs. It offers 3D printers, Quickparts solutions, 3D authoring tools and scanners, Bespoke Modeling, and TeamPlatform. It operates its business in America, Germany and Asia Pacific. The company was founded by Charles W. Hull in 1986 and is headquartered in Rock Hill, SC.*

Secteur d'activité : Technology
Groupe Industrie : Computer Peripherals
Capitalisation : 960,165,000 $ US

6.1.2 3M Company

Description : *3M Co. is a diversified technology company, which manufactures industrial, safety and consumer products. The company operates its business through the following segments : Industrial, Safety Graphics, Health Care, Electronics Energy, and Consumer. The Industrial segment provides products, including tapes, abrasives, adhesives, specialty materials and filtration systems to diverse markets from purification to aerospace. The Safety Graphics segment offers personal protective equipment, traffic safety security products, commercial graphics systems, commercial cleaning protection products, floor matting, roofing granules for asphalt shingles, and fall protection products. The Health Care segment supplies medical and surgical equipment, skin health infection prevention products, drug delivery systems, dental orthodontic products, health information systems and food safety products. The Electronics Energy segment offers optical films solutions for electronic displays, packaging and interconnection devices ; insulating and splicing solutions ; touch screens and touch monitors ; renewable energy component solutions ; and infrastructure protection products. The Consumer segment provides sponges, scouring pads, high-performance cloths, consumer and office tapes, repositionable notes, indexing systems, home improvement products, home care products, protective material products, and consumer office tapes, as well as adhesives. The company was founded by Henry S. Bryan, Hermon W. Cable, John Dwan, William A. McGonagle and J. Danley Budd in 1902 and is headquartered in St. Paul, MN.*
Secteur d'activité : Industrial Goods
Groupe Industrie : Diversified Machinery
Capitalisation : 97,964,645,000 $ US

6.1.3 Aramark

Description : *Aramark engages in the provision of food, facilities, and uniform services to education, healthcare, business, sports, leisure, and corrections clients. It operates through the following segments : Food and Support Services North America (""FSS North America"") ; Food and Support Services International (""FSS International"") ; and Uniform. The FSS North America segment provides food and facilities services to colleges, universities, schools, hospitals, nursing homes, offices, concert venues, and correctional facilities in North America. The FSS International segment offers dining, catering, and facilities management services*

to schools, hospitals, nursing homes, office parks and buildings, arenas, stadiums, and correctional centers in the United Kingdom, Germany, Chile, China, and Ireland. The Uniform segment designs, manufactures, and delivers uniforms and work clothes. The company was founded in 1959 and is headquartered in Philadelphia, PA.
Secteur d'activité : Services
Groupe Industrie : Specialty Eateries
Capitalisation : 11,106,080,700 $ US

6.2 Corpus de validation et méthode d'analyse des résultats

Afin d'évaluer nos modèles, nous avons calculé le rappel par rapport à un ensemble de validation. La préparation de l'ensemble de validation a été effectuée par le regroupement des données fournies par un expert du domaine de la finance en investissement pour les compagnies de notre base de données (2991 compagnies) et par consultation des données en lignes des comparables et des pairs de compagnies similaires «Peers» des sites financiers. Enfin, les entreprises qui ne sont pas dans la base de données «Intrinio US Fundamentals and Stock Prices» ont été retirées de l'ensemble de validation de compagnies comparables pour avoir un total de 2991 entreprises présent dans les compagnies cibles.

Nous utiliserons le pourcentage de rappel comme mesure de performance sur toutes les entreprises possédant des comparables dans nos données de validations. La mesure de rappel sera considérée pour des tranches de comparables de nos modèles allant de dix à cent compagnies, par tranches de dix compagnies (10,20,30,40,50,60,70,80,90,100). La moyenne du pourcentage de rappel sera ainsi calculée pour toutes les tranches de rappel.

La distribution du nombre d'entreprises comparables cibles par entreprises dans notre corpus est présentée dans la figure 6.1 :

Notre meilleur modèle aura montré son efficacité par son taux de rappel. Par ailleurs, il est important de noter que ces recommandations de compagnies n'ont aucun ordre de similarité dans notre corpus de validation. La méthode de validation est celle du taux de rappel sur le corpus de validation et la comparaison au modèle de base.

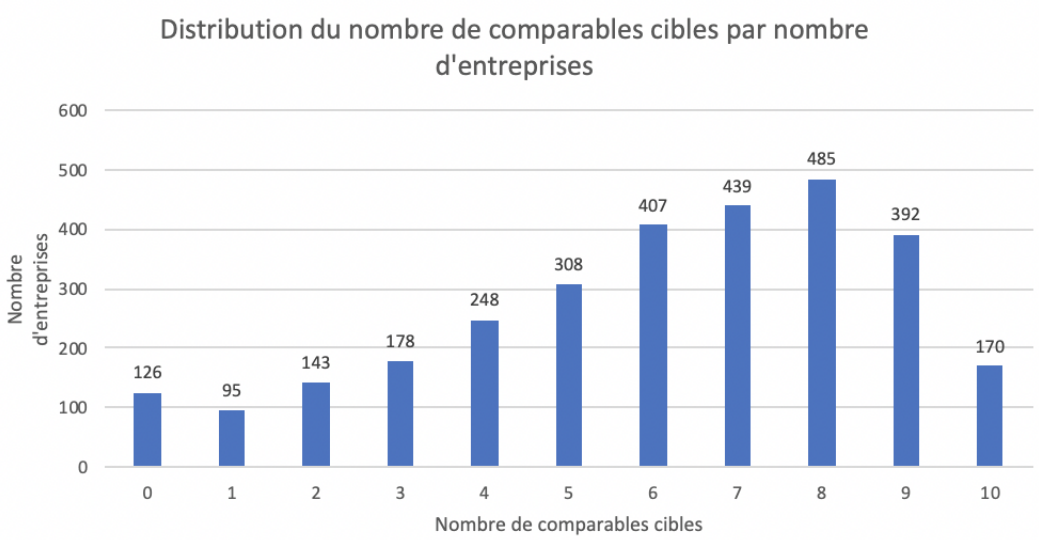

Figure 6.1 Distribution du nombre de comparables par entreprises (total de 2865 avec au moins un comparable)

Les tableaux ci-dessous présentent les comparables cibles définis pour les trois entreprises auxquelles nous nous sommes intéressés au tout début du présent chapitre avec leur code SIC et NAICS qui est expliqué dans la section 2.4 :

Comparables DDD	Symbole	SIC Code	NAICS Code
Eastman Kodak Co	NYSE :KODK	3861	812921
HP Inc	NYSE :HPQ	3570	443142

Tableau 6.2 Comparables de 3D Systems Corporation

Comparables MMM	Symbole	SIC Code	NAICS Code
Honeywell International Inc	NYSE :HON	3714	541330
General Electric Co	NYSE :GE	3600	423830
Rockwell Automation Inc	NYSE :ROK	3829	335314
Roper Technologies Inc	NYSE :ROP	3823	551112

Tableau 6.3 Comparables de 3M Company

À partir des tableaux 6.2, 6.3 et 6.4, nous pouvons remarquer que les codes de classification (SIC et NAICS) ne sont pas représentatifs et fidèles à la réalité des activités des compagnies. En effet, s'ils ne représentent pas toutes les activités des compagnies, c'est que les codes SIC et NAICS sont quelque peu limités comme expliqué dans la section 2.4.

Comparables ARMK	Symbole	SIC Code	NAICS Code
Starbucks Corp	NYSE :SBUX	5810	311811
Cintas Corp	NYSE :CTAS	2320	423830
Darden Restaurants Inc	NYSE :DRI	5812	722511
Servicemaster Global Holdings Inc	NYSE :SERV	8741	551112

Tableau 6.4 Comparables de Aramark

6.2.1 Limites du corpus de validation

Notre recherche présente certaines limites dans sa liste de comparables cibles. En effet, notre corpus de validation a été regroupé par un expert du domaine et présente des limites, car il est possible que des compagnies présent dans le corpus de validation soient erronées et dans l'autre sens qu'il en manque certaines. Cette limite crée la possibilité de faux négatifs qui seraient retournés par l'algorithme et qui ne seraient pas présents dans les compagnies cibles. Comme nous comparons nos modèles entre eux par rapport au modèle de base, cette limite des compagnies cibles est à prendre en considération lors de l'analyse de nos résultats et nous présenterons des investigations spécifiques pour les trois exemples (DDD, MMM, ARMK). De plus, nous sommes limités par rapport aux données dont nous avons accès par la base de données de Intrinio et dont les comparables sont présent dans le corpus de validation.

6.3 Distribution des données des entreprises

Dans cette section nous présentons dans les 3 figures suivantes les distributions des groupes d'industrie, des secteurs d'activité et des capitalisations des entreprises des données d'Intrinio.

Dans la figure 6.2 nous présentons dans le carré les 205 groupes ou l'aire représente le nombre de compagnies dans chaque groupe d'industrie, cette liste est présentée sous forme de tableau dans l'annexe B.

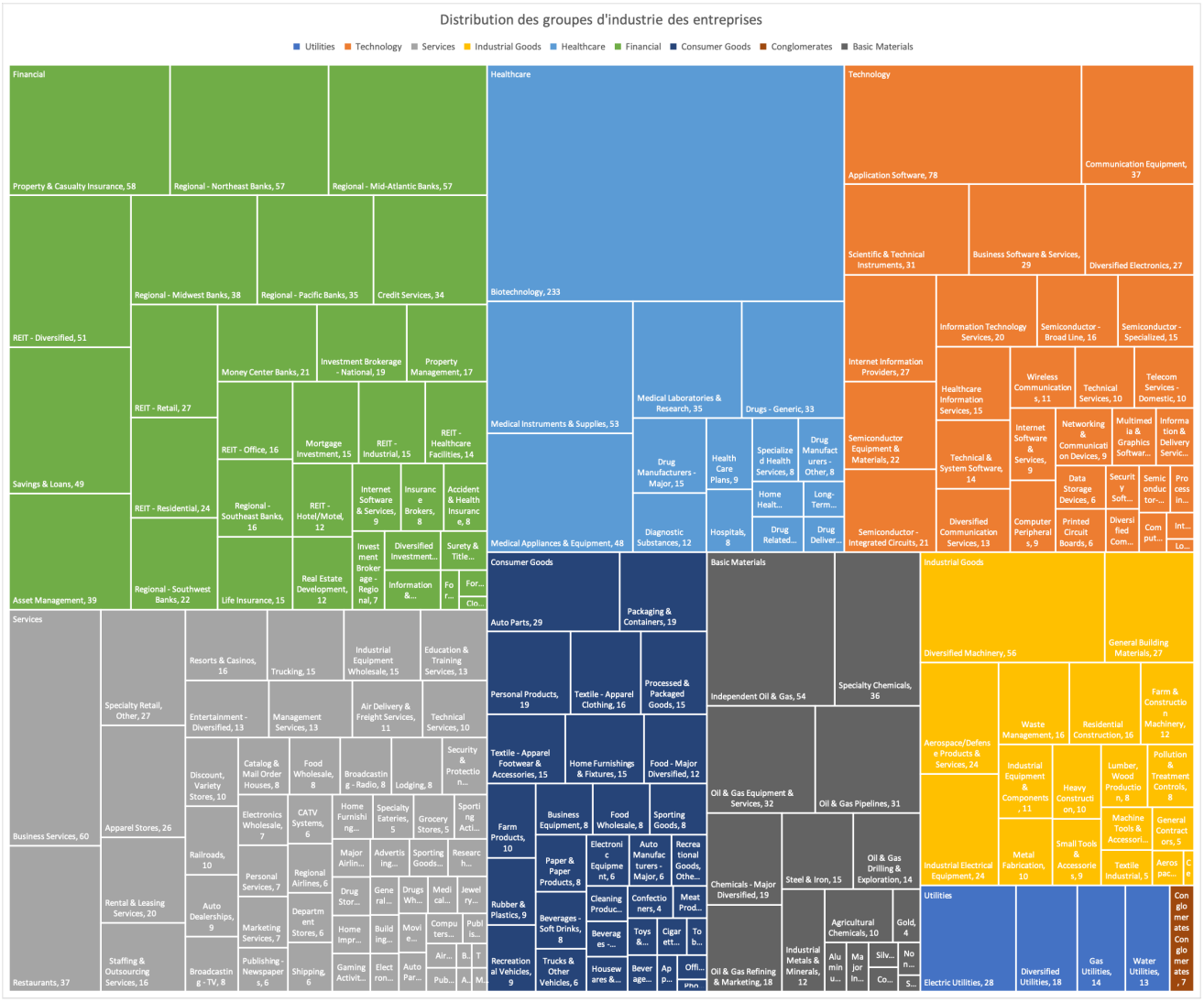

Figure 6.2 Distribution des groupes d'industrie des entreprises

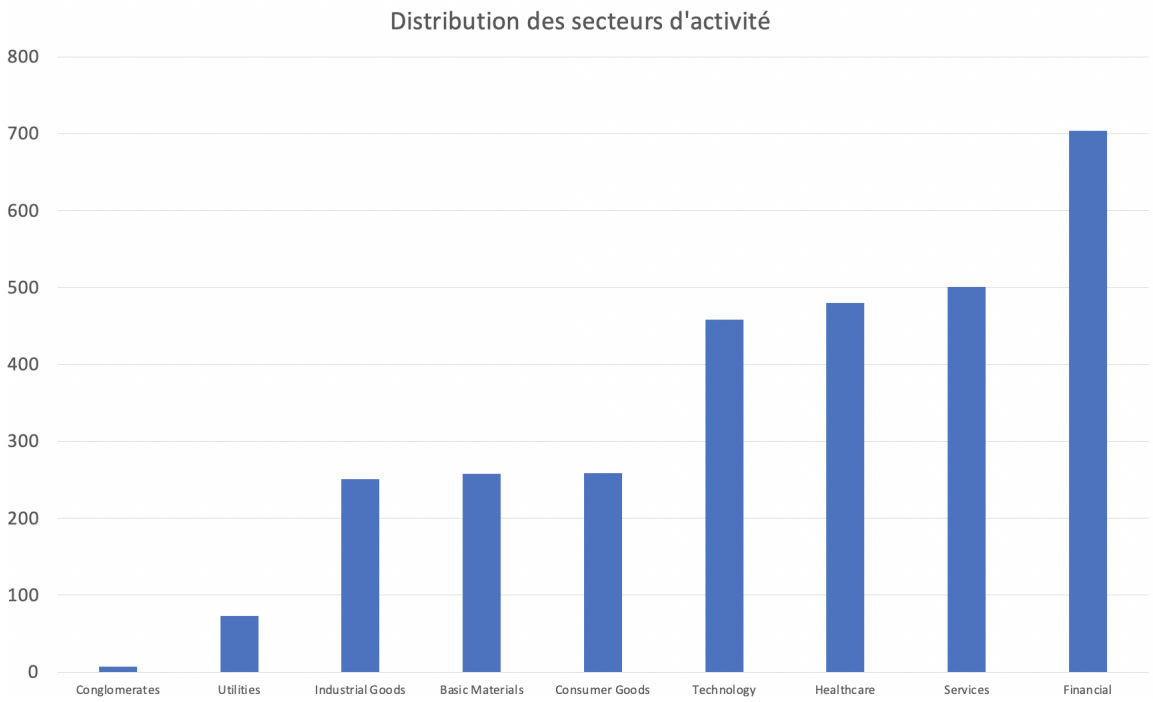

Figure 6.3 Distribution des secteurs d'activités des entreprises

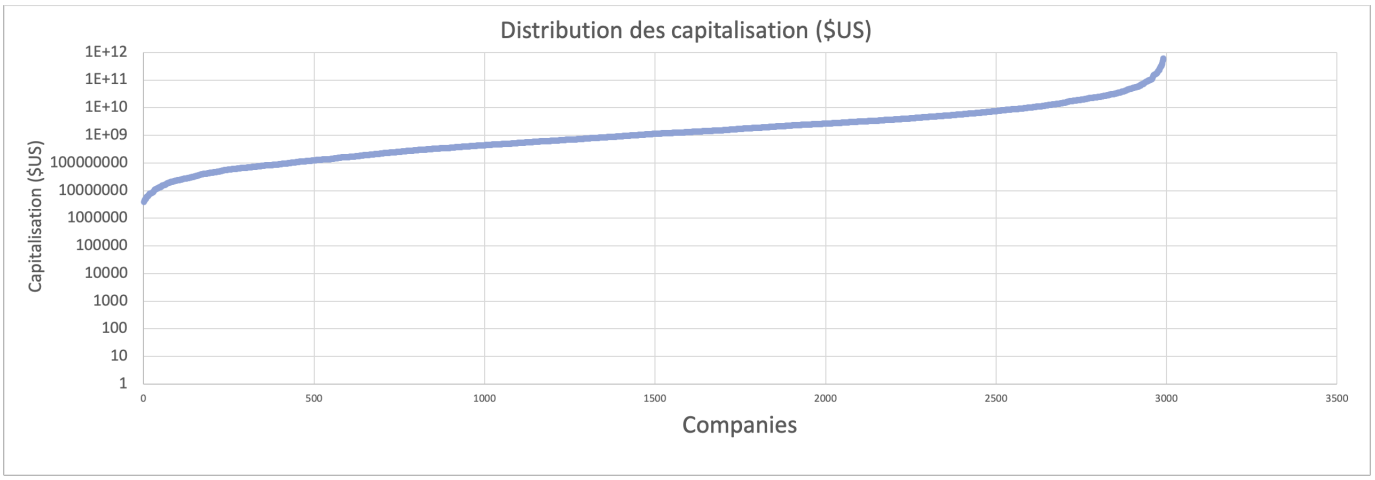

Figure 6.4 Distribution des capitalisations des entreprises

Dans le prochain chapitre, nous présenterons les résultats sur les trois exemples présentés dans ce chapitre et les résultats des taux de rappel globaux de nos différents modèles.

CHAPITRE 7 RÉSULTATS ET ÉVALUATION

Nous avons fait un survol des concepts financiers derrière l'analyse des compagnies publiques et présenté les algorithmes de recherche de compagnies comparables. Dans ce chapitre, nous présenterons nos résultats en nous basant sur les méthodes présentées aux chapitres précédents. Avant de conclure notre étude, ce dernier chapitre présentera les résultats globaux d'une part et ceux spécifiques aux trois compagnies présentées comme exemple «3D Systems», «3M Company» et «Aramark» du modèle sans description, et avec description seulement. Rappelons que ces compagnies ont été présentées dans la méthodologie, et que, pour chacune, nous avons appliqué d'abord le modèle de base, à savoir le modèle sans analyse de descriptions, et par la suite, avec description brute. Les résultats de rappel globaux seront ensuite présentés, englobant tous les modèles décrits dans le chapitre 5.

Dans le tableau 7.1 un résumé des différents modèles d'algorithmes est présenté.

MODÈLE	Catégorie industrie	Capitalisation	Noms Verbes WordNet	Similarité chemin dans WordNet	Descriptions	Descriptions Augmentées	FILTRE DESCRIPTIONS			
							Aucun	Noms	Verbes	Autres
MCF	✓	✓								
MCFD	✓	✓			✓		X			
MCFDn	✓	✓			✓			X		
MCFDv	✓	✓			✓				X	
MCFDgv	✓	✓			✓					X
MCFDWN	✓	✓	✓	✓	✓		X			
D2V					✓		X			
D2Va						✓	X			

Figure 7.1 Récapitulation des modèles et du contenu considéré

7.1 Modèle de base sans analyse des descriptions

Nous présenterons dans ce chapitre les modèles de base. Ici, nous avons uniquement considéré les valeurs financières, ainsi que les catégories.

7.1.1 Résultats 3D Systems Corporation modèle de base

Dans cette section nous présentons les résultats de 3D Systems Corporation selon le modèle de base avec la capitalisation et les classifications d'industries et les secteurs d'activités.

SYMBOLE	NOM DE L'ENTREPRISE	Groupe Industrie	Capitalisation (M$)	R_{cap}	R_{cat}	Rappel
0.DDD	3D Systems Corporation	Computer Peripherals	1493	1	1	0
1.MRCY	Mercury Systems, Inc.	Computer Peripherals	1239	0.9690	1	0
2.LOGI	Logitech International S.A.	Computer Peripherals	4114	0.8317	1	0
3.ALOT	AstroNova, Inc.	Computer Peripherals	106	0.5617	1	0
4.DBD	Diebold Nixdorf, Incorporated	Diversified Computer Systems	1889	0.9609	0.571	0
5.KTCC	Key Tronic Corporation	Computer Peripherals	84	0.5225	1	0
6.CNDT	Conduent Incorporated	Computer Based Systems	3022	0.8829	0.571	0
7.SYKE	Sykes Enterprises, Incorporated	Information Technology Services	1224	0.9670	0.465	0
8.TACT	TransAct Technologies Incorporated	Computer Peripherals	48	0.4319	1	0
9.DAKT	Daktronics Inc.	Computer Based Systems	471	0.8085	0.571	0
10.INVE	Identiv, Inc.	Computer Peripherals	35	0.3782	1	0

Tableau 7.1 10 premiers comparables de 3D Systems Corporation selon le secteur d'activité le groupe d'industrie et la capitalisation

Résultat du rappel sur les 10 premières recommandations : 0 rappel sur 3 cibles. Total de 0 % de rappel sur les 10 premières recommandations.

En prenant comme référence le tableau 6.2 des compagnies comparables cibles à 3D Systems, nous remarquons d'abord que les entreprises recommandées selon la capitalisation et la catégorie d'industrie du tableau 7.1 ne coïncident pas avec les compagnies de notre corpus de validation. En effet, un nombre important d'entreprises possèdent une capitalisation comparable et des catégories comparables à l'industrie de 3D Systems. En raison de quoi, avant que les recommandations coïncident avec notre corpus de validation, un grand nombre de compagnies sont retournées. De ces résultats, nous pouvons retenir que la seule utilisation de la catégorie d'industrie et de la capitalisation est insuffisante pour trouver de bons comparables — précisons, du moins, dans le cas de 3D Systems.

Le tableau nous fait relève que la compagnie la plus comparable, selon cette méthode, est Mercury Systems Inc : une entreprise privée d'équipements informatiques pour l'armée des États-Unis. La clientèle de Mercury Systems présente une grande différence avec 3D Systems, une entreprise de fabrication d'imprimante 3D.

7.1.2 Résultats 3M Company modèle de base

Les résultats du modèle de base pour 3M Company sont rapportés dans le tableau 7.2.

SYMBOLE	NOM DE L'ENTREPRISE	Groupe Industrie	Capitalisation (M$)	R_{cap}	R_{cat}	Rappel
0.MMM	3M Company	Diversified Machinery	107403	1	1	0
1.HON	Honeywell International Inc.	Diversified Machinery	88292	0.9808	1	0
2.DHR	Danaher Corporation	Diversified Machinery	53842	0.9324	1	1
3.ITW	Illinois Tool Works Inc.	Diversified Machinery	42984	0.9103	1	0
4.GE	General Electric Company	Diversified Machinery	279545	0.9063	1	1
5.ETN	Eaton Corporation plc	Diversified Machinery	30304	0.8761	1	0
6.CMI	Cummins Inc.	Diversified Machinery	22903	0.8487	1	0
7.IR	Ingersoll-Rand Plc	Diversified Machinery	19385	0.8324	1	0
8.ROP	Roper Technologies, Inc.	Diversified Machinery	18570	0.8282	1	1
9.ROK	Rockwell Automation Inc.	Diversified Machinery	17299	0.8212	1	1
10.DOV	Dover Corporation	Diversified Machinery	11640	0.7824	1	0

Tableau 7.2 10 premiers comparables de 3M Company selon la catégories d'industrie et la capitalisation

Résultat du rappel sur les 10 premières recommandations : 4 rappels sur 4 cibles. Total de 100 % de rappel sur les 10 premières recommandations.

Utilisons le tableau 6.3 des entreprises comparables cibles à 3M Company (un conglomérat d'entreprises) en tant que référence. Suivant cette démarche, nous remarquons, à l'inverse des résultats de la section précédente, que les entreprises recommandées selon la capitalisation et la catégorie d'industrie du tableau 7.2 conduit en effet aux compagnies de notre corpus de validation pour 3M Company, c'est-à-dire «Honeywell Internatioanl Inc», «Roper Technologies Inc», «General Electric Company» et «Rockwell Automation Inc.». Cette coïncidence des compagnies avec les compagnies cibles procède du fait qu'elles sont chacune, non pas une simple compagnie, mais plutôt un conglomérat d'entreprises. Ce qui signifie essentiellement qu'elles possèdent des capitalisations bien plus importantes que les capitalisations que nous retrouvons généralement dans l'industrie — concrètement, plus de 10 milliards de dollars. Au surplus, le tableau affiche, en tant que résultat le plus comparable, dans le cadre de cette méthode, la compagnie «Honeywell International Inc», un conglomérat d'entreprises également basé aux États-Unis. Tandis que «Honeywell International Inc» manufacture des produits dans le domaine de l'aérospatiale, notamment des moteurs, «3M Company» est aussi un conglomérat, mais plus spécifiquement un agglomérat relatif au domaine de la construction. C'est pourquoi le principal trait commun entre ces deux entreprises repose uniquement sur leur taille et leur organisation, dès lors qu'elles se spécialisent chacune dans son propre secteur. Deux secteurs différents, donc, imposent deux clientèles différentes. Par rapport à notre corpus de validation, l'exemple de 3M Company met en évidence l'importance de tenir

compte de la taille d'une entreprise pour les entreprises comparable, autant que la catégorie d'industrie, cependant on remarque que cette catégorie d'industrie représente des conglomérats qui sont généralement de très grades compagnies.

7.1.3 Résultats Aramark modèle de base

Dans cette section nous présentons les résultats de Aramark selon le modèle de base avec uniquement les valeurs financières et la classification d'industrie et de secteur d'activité.

SYMBOLE	NOM DE L'ENTREPRISE	Groupe Industrie	Capitalisation (M$)	R_{cap}	R_{cat}	Rappel
0.ARMK	Aramark	Specialty Eateries	8781	1	1	0
1.PZZA	Papa John's International, Inc.	Specialty Eateries	3157	0.8673	1	0
2.SBUX*	Starbucks Corporation	Specialty Eateries	80803	0.7122	1	1
3.SHAK	Shake Shack Inc.	Specialty Eateries	891	0.7034	1	0
4.PBPB	Potbelly Corporation	Specialty Eateries	324	0.5725	1	0
5.TSCO	Tractor Supply Company	Specialty Retail, Other	9950	0.9838	0.5717	0
6.KAR	KAR Auction Services, Inc.	Specialty Retail, Other	5888	0.9481	0.5717	0
7.GPC	Genuine Parts Company	Specialty Retail, Other	14210	0.9375	0.5717	0
8.IAC	IAC/InterActiveCorp	Specialty Retail, Other	5137	0.9304	0.5717	0
9.ULTA	Ulta Beauty, Inc.	Specialty Retail, Other	1586	0.9233	0.5717	0
10.ROL	Rollins, Inc.	Business Services	7358	0.9770	0.5122	0

Tableau 7.3 10 premiers comparables de Aramark selon la catégorie d'industrie et la capitalisation

Résultat du rappel sur les 10 premières recommandations : 1 rappel sur 4 cibles. Total de 25 % de rappel sur les 10 premières recommandations.

Utilisons le tableau 6.4 des entreprises comparables cibles à Aramark (une entreprise de nourriture et de services d'uniformes pour les écoles et les hôpitaux et les institutions) en tant que référence. Nous remarquons alors que les entreprises recommandées, selon la capitalisation et la catégorie d'industrie du tableau 7.3, ne coïncident qu'avec une seule des compagnies de notre corpus de validation «Starbucks Corporation». Cette entreprise est englobée par la même catégorie d'industrie, mais elle possède une capitalisation très grande. Néanmoins, Starbucks est une entreprise de café qui opère entièrement des chaînes. Un comportement global que nous pouvons observer par rapport au groupe d'industrie est que la présence d'un mot similaire entre les mots des catégories peut souvent causer des erreurs comme dans l'exemple d'Aramark ou le mot «Speciality» dans le groupe d'industrie et «Service» dans le secteur d'activité sont des mots qui peuvent être présent dans des catégories qui ne sont pas comparables comme dans cet exemple la compagnie «IAC», une compagnie de

média contenant le mot «Speciality» et lui donnant une similarité de 0.5717 avec la catégorie d'Aramark qui est une compagnie de restauration.

7.1.4 Résultats sur l'ensemble des données du modèle de base

Établissons maintenant la performance du modèle de base pour l'ensemble du corpus de validation. Ce rappel se base sur différents nombres de recommandations, allant de dix à cent, par tranches de dix recommandations supplémentaires.

Nombre de recom.	Rappel Global : Catégorie Capital.
10	26.4%
20	40.3%
30	49.4%
40	54.0%
50	57.7%
60	61.0%
70	64.0%
80	66.2%
90	68.6%
100	69.9%

Tableau 7.4 Rappel sur les résultats globaux pour le modèle de base avec la catégorie d'industrie et la capitalisation par nombre de recommandations.

Le tableau 7.1.4 représente le rappel global des recommandations par tranches de dix, allant de dix à cent, pour les recommandations en fonction de la catégorie d'industrie et de la capitalisation. En retenant uniquement les dix premières recommandations pour chaque entreprise, nous obtenons un rappel moyen de 26.4%. Parallèlement, sur les cent premières recommandations d'entreprises, un rappel moyen de 69.9 % est obtenu. Les résultats de cette méthode seront alors utilisés comme référence base.

7.2 Recommandation avec analyse des descriptions

Dans les sections qui suivent, nous allons présenter dans le même format que pour la section 7.1 les résultats avec les trois exemples (3D systems Corporation, 3M company et Aramark) ainsi que les résultats globaux avec analyse de descriptions.

7.2.1 Résultats 3D Systems Corporation avec analyse des descriptions (MCFD)

Les résultats de 3D Systems Corporation selon le modèle MCFD avec analyse des descriptions, les capitalisations et les groupes d'industrie et de secteur d'activité sont rapportés au tableau 7.5.

SYMBOLE	NOM DE L'ENTREPRISE	R_{cap}	R_{cat}	R_{des}	Rappel
0.DDD	3D Systems Corporation	1	1	1	0
1.MRCY	Mercury Systems, Inc.	0.9690	1	0.2600	0
2.LOGI	Logitech International S.A.	0.8317	1	0.2747	0
3.ALOT	AstroNova, Inc.	0.5617	1	0.3258	0
4.DBD	Diebold Nixdorf, Incorporated	0.9609	0.5717	0.3289	0
5.TACT	TransAct Technologies Incorporated	0.4319	1	0.3965	0
6.KTCC	Key Tronic Corporation	0.5225	1	0.2996	0
7.CACI	CACI International Inc	0.8825	0.4656	0.3979	0
8.EPAM	EPAM Systems, Inc.	0.8692	0.4656	0.3979	0
9.SYKE	Sykes Enterprises, Incorporated	0.9670	0.4656	0.2799	0
10.VRTU	Virtusa Corporation	0.8862	0.4656	0.3424	0

Tableau 7.5 10 premiers comparables de 3D Systems Corporation selon la catégorie d'industrie, la capitalisation et la description brute

Résultat du rappel sur les 10 premières recommandations : 0 rappel sur 3 cibles. Total de 0 % de rappel sur les 10 premières recommandations.

Le tableau 7.5 présente les dix premières compagnies comparables de «3D Systems», ainsi que les résultats des corrélations, en ayant inclus la description. Par rapport aux compagnies comparables cibles du tableau 6.3, nous ne retrouvons aucun comparable. Remarquons, par ailleurs, parmi les entreprises comparables recommandées plusieurs sont présent dans les résultats du tableau 7.1.

7.2.2 Résultats 3M Company avec analyse des descriptions

Dans cette section nous présentons les résultats des compagnies comparables de 3M Company selon le modèle MCFD avec analyse des descriptions, les valeurs financières et les classifications d'industrie et de secteur d'activité.

SYMBOLE	NOM DE L'ENTREPRISE	R_{cap}	R_{cat}	R_{des}	Rappel
0.MMM	3M Company	1	1	1	0
1.DHR	Danaher Corporation	0.9324	1	0.5327	0
2.HON	Honeywell International Inc.	0.9808	1	0.4351	1
3.ITW	Illinois Tool Works Inc.	0.9103	1	0.4627	0
4.GE	General Electric Company	0.9063	1	0.4576	1
5.ETN	Eaton Corporation plc	0.8761	1	0.4351	0
6.ROP	Roper Technologies, Inc.	0.8282	1	0.4457	1
7.ROK	Rockwell Automation Inc.	0.8212	1	0.4279	1
8.DOV	Dover Corporation	0.7824	1	0.4445	0
9.CMI	Cummins Inc.	0.8487	1	0.3650	0
10.AME	AMETEK, Inc.	0.7791	1	0.4301	0

Tableau 7.6 10 premiers comparables de 3M Company selon la catégories d'industrie, la capitalisation et la description

Résultat du rappel sur les 10 premières recommandations : 4 rappels sur 4 cibles
Total de 100 % de rappel sur les 10 premières recommandations

Dans le tableau 7.6 ci-dessus, on présente les dix premières compagnies comparables de 3M Company avec le modèle MCFD, ainsi que les résultats des corrélations. Remarquons, en premier lieu, que selon les résultats du tableau 7.2, nous obtenons le même rappel avec la catégorie d'industrie et la capitalisation. Par rapport à nos entreprises cibles, nous avons les quatre entreprises comparables à celles des compagnies cibles «Honeywell Internatioanl Inc», «Roper Technologies Inc», «General Electric Company» et «Rockwell Automation Inc.» ce qui nous montre que la catégorie d'industrie avec la finance peut avoir un poids déterminant.

7.2.3 Résultats Aramark avec analyse des descriptions

Le tableau 7.7 rapporte les résultats de Aramark selon le modèle MCFD avec analyse des descriptions, les valeurs financières et les classifications d'industrie et de secteur d'activité.

SYMBOLE	NOM DE L'ENTREPRISE	R_{cap}	R_{cat}	R_{des}	Rappel
0.ARMK	Aramark	1	1	1	0
1.PZZA	Papa John's International, Inc.	0.8673	1	0.2304	0
2.GJEC	Jacobs Engineering Group Inc.	0.9684	0.5122	0.5938	0
3.SBUX	Starbucks Corporation	0.7122	1	0.2684	1
4.G	Genpact Limited	0.9239	0.5122	0.5038	0
5.CTAS	Cintas Corporation	0.9580	0.5122	0.4459	1
6.ROL	Rollins, Inc.	0.9770	0.5122	0.4158	0
7.FISV	Fiserv, Inc.	0.8747	0.5122	0.5000	0
8.ACM	AECOM	0.9416	0.5122	0.4307	0
9.ABM	ABM Industries Incorporated	0.8245	0.5122	0.5460	0
10.TSCO	Tractor Supply Company	0.9838	0.5717	0.2990	0

Tableau 7.7 10 premiers comparables de Aramark selon la catégorie d'industrie, la capitalisation et la description

Résultat du rappel sur les 10 premières recommandations : 2 rappels sur 4 cibles
Total de 50 % de rappel sur les 10 premières recommandations

Comme on le remarque au tableau 7.7, deux compagnies recommandées y figurent «SBUX et CTAS». Nous remarquons que par rapport au modèle de base sans la description, nous arrivons à obtenir un comparable en plus «CTAS» , la compagnie «Cintas Corporation».

7.3 Résultats par modèles avec analyse des descriptions sur l'ensemble des entreprises

Dans cette section, nous présentons la moyenne du rappel de toutes les entreprises sur les différents modèles de traitement des descriptions, avec différents nombres de recommandations, allant de dix à cent par tranche de dix recommandations supplémentaires.

7.3.1 Modèle (MCFD) sur l'ensemble des données avec analyse des descriptions

Dans cette section, on présente les résultats de nos expérimentations du modèle de la section 5.2. La métrique que nous présentons est la moyenne du rappel de toutes les entreprises, en fonction du modèle avec analyse des descriptions complète, les données des capitalisations, les groupes d'industries et de secteur d'activité, avec différents nombres de recommandations, allant de dix à cent par tranche de dix recommandations supplémentaires.

Nombre de recom.	Rappel Global : Catégorie Cap. Description
10	28.9%
20	43.8%
30	52.5%
40	57.6%
50	62.8%
60	66.4%
70	68.7%
80	71.0%
90	73.4%
100	75.4%

Tableau 7.8 Rappel sur les résultats globaux pour le modèle avec la catégorie d'industrie, la capitalisation et la description

À partir des résultats présentés jusqu'à maintenant, nous pouvons arriver au constat suivant : l'utilisation de description, en plus des catégories industrielles et financières, permet en effet de parvenir d'augmenter le rappel. Notons que le rappel avec les dix premières compagnies comparables passe de 26.4%, sans description, à 28.9%, avec description cette fois. De la même manière, quant au rappel avec les cent premières compagnies comparables, nous passons de 69.9%, sans description, à 75.4%, avec description. Notons que nous utilisons des données de rappel afin de comparer les modèles entre eux et que si nous avons des taux de rappel

faibles cela ne veut pas dire que les entreprises recommandées sont fausses, car ne sont pas dans notre corpus de validation on assume la possibilité de faux négatifs et de faux positifs comme présentés à travers la compagnie «DDD 3D Systems».

7.3.2 Modèle ($MCFD_n$) avec les noms dans la description avec transformation TF-IDF

On présente dans le tableau 7.9 les résultats de nos expérimentations du modèle de la section 5.3. La métrique que nous présentons est la moyenne du rappel de toutes les entreprises, en fonction du modèle avec analyse des descriptions filtré aux noms, les données financières, les classifications d'industrie et de secteur d'activité, avec différents nombres de recommandations, allant de dix à cent par tranche de dix recommandations supplémentaires.

Nombre de recom.	Rappel Global : Catégorie Cap. Description (Noms)
10	29.5%
20	44.6%
30	53.2%
40	58.2%
50	63.8%
60	67.0%
70	69.7%
80	72.1%
90	74.3%
100	76.4%

Tableau 7.9 Rappel sur les résultats globaux pour le modèle avec la catégorie d'industrie, la capitalisation et les noms dans la description

Lorsque nous comparons notre modèle de base avec le modèle qui inclut les descriptions et qui filtre les noms dans les descriptions. Nous remarquons que le modèle qui filtre les noms dans les descriptions permet d'avoir un meilleur rappel que le modèle qui ne prend pas compte de la description. Dans les 10 premiers comparable on passe de 26.4% sans description à 29.5% de taux de rappel global moyen et dans les 100 premiers comparables on passe de 69.9% à 76.4%. Cette augmentation nous permet d'identifier le modèle avec les noms dans la description, la catégorie d'industrie, le secteur d'activité et la capitalisation comme étant parmi les meilleurs modèles de recommandation de compagnies comparables selon notre recherche et nos données.

7.3.3 Modèle ($MCFD_v$) avec les verbes dans la description avec transformation TF-IDF

Les résultats de nos expérimentations du modèle de la section 5.4 sont présentés dans le tableau 7.10. La métrique que nous présentons est la moyenne du rappel de toutes les entreprises, en fonction du modèle avec analyse des descriptions filtré aux verbes, les données financières, les classifications d'industrie et de secteur d'activité, avec différents nombres de recommandations.

Nombre de recom.	Rappel Global : Catégorie Cap. Description (Verbes)
10	24.3%
20	39.3%
30	47.9%
40	54.4%
50	58.6%
60	62.4%
70	65.1%
80	67.7%
90	70.3%
100	72.8%

Tableau 7.10 Rappel sur les résultats globaux pour le modèle avec la catégorie d'industrie, la capitalisation et les verbes dans la description

Dans nos résultats avec filtrage des verbes dans les descriptions, on remarque que le modèle qui filtre les verbes dans les descriptions donne un moins bon rappel sur les 10 premières recommandations que les modèles sans description. Dans les 10 premières compagnies recommandées, on passe de 26.4% sans description à 24.3% sur le taux de rappel moyen global avec les verbes dans les descriptions et dans les 100 premières compagnies comparables on passe de 69.9% à 72.8% ce qui présente une petite augmentation. Nous remarquons que les verbes des descriptions ne permettent pas aussi bien que les noms dans les descriptions de recommander les comparables dans notre corpus de validation. Nous pouvons assumer qu'en effet, les verbes sont des mots qui varient moins que les noms dans chaque description et qui permettent de moins différencier les activités des entreprises que les noms ce qui résulte à obtenir des recommandations de compagnies comparables qui sont moins bonnes.

7.3.4 Modèle ($\mathrm{MCFD_{gv}}$) avec les groupes verbe et groupes complément dans la description avec transformation TF-IDF

Dans le tableau 7.11 on présente les résultats de nos expérimentations du modèle de la section 5.5, TF-IDF avec les groupes verbes et groupes compléments dans la description. La métrique que nous présentons est la moyenne du rappel de toutes les entreprises, en fonction du modèle avec analyse des descriptions filtrées au groupe verbe et groupe complément, la capitalisation, les groupes d'industrie et de secteur d'activité.

Nombre de recom.	Rappel Global : Catégorie Cap. Description (Groupe Verbaux)
10	26.6%
20	42.1%
30	51.5%
40	57.8%
50	62.1%
60	66.1%
70	68.3%
80	70.3%
90	72.1%
100	73.9%

Tableau 7.11 Rappel sur les résultats globaux pour le modèle avec la catégorie d'industrie, la capitalisation et les groupes verbes et complément dans la description

On remarque que l'utilisation des groupes verbes et groupes compléments dans les descriptions permet d'avoir un meilleur rappel que sans la description, mais pas de l'utilisation des noms dans la description. Dans les 10 premières compagnies comparables, on passe de 26.4% sans description à 26.6% et dans les 100 premières compagnies comparables on passe de 69.9% à 73.9% avec l'utilisation des groupes verbes et groupes compléments. Lors de notre expérimentation, nous avons aussi éprouvé des difficultés avec ce modèle dues au fait que la segmentation de phrases selon les groupes verbaux n'est pas parfaite par la librairie que nous avons utilisée. En effet, les phrases qui sont très longues ou qui sont mal structurées ne peuvent être parfaitement segmentées et génèrent des erreurs et ne sont pas considérées dans l'algorithme. Cette source d'erreur et ces faibles résultats nous forcent de ne pas retenir de modèle parmi nos meilleurs.

7.3.5 Modèle (MCFDWN) avec graphique de connaissance (WORDNET) et transformation TF-IDF

Dans cette section, on présente les résultats du modèle de la section 5.6 avec le groupe d'industrie, le secteur d'activité, la capitalisation, la similarité selon les parcours dans le graphique de connaissance «WORDNET», la transformation TF-IDF avec les définitions des noms et des verbes désambiguïés selon les méthodes présentées à la section 3.2.3, la transformation TF-IDF avec la description brute. La métrique que nous présentons est identique aux sections précédentes.

Nombre de recom.	Rappel Global : Catégorie Cap. Description (WORDNET)
10	29.9%
20	47.0%
30	56.2%
40	63.6%
50	68.2%
60	71.1%
70	74.4%
80	76.1%
90	77.7%
100	79.5%

Tableau 7.12 Rappel sur les résultats globaux pour le modèle avec la catégorie d'industrie, la capitalisation, la description et le graphique de connaissance pour la similarité et les descriptions désambiguïsées.

Notre méthode qui est la plus optimale dans l'implémentions avec nos résultats et selon le contexte financier et des besoins de l'industrie (reproductibilité, adaptable avec de nouvelles données sans besoin de temps d'entraînement de paramètres étendus tels que les réseaux de neurones) et qui donne les meilleures recommandations de compagnies comparables selon notre étude est le modèle MCFDWN. Ce modèle utilise les descriptions brutes, la capitalisation, le secteur d'activité, le groupe d'industrie, le graphique de connaissance «WordNet» pour la désambiguïsation et pour la similarité selon la distance des chemins parcourus entre des descriptions.

7.4 Modèles (D2V) exploratoires avec vecteur des descriptions Doc2Vec et Augmentation des données

Dans cette section, nous présentons notre exploration avec la méthode «Doc2Vec» (vectorisation des documents) et l'augmentation des descriptions avec plongement des mots. Le détail de la méthode et des librairies utilisées est expliqué à la section 5.6. La motivation de cette exploration vient des nouvelles méthodes de l'intelligence artificielle dans le domaine des réseaux de neurones et du traitement des textes afin d'évaluer la performance d'utiliser ces méthodes plus tôt que les modèles dont nous avons présenté. Notons que ces méthodes nécessitent un long travail d'entraînement et d'optimisation d'hyperparamètres qui doit être optimisé lorsque chaque nouvelle compagnie apparaît ou les données des descriptions changent. Ce modèle nécessitant d'adapter les plongements lexicaux en fonction des changements et d'entraîner le modèle pour des résultats plus faibles ce qui est la raison principale qui nous pousse à ne pas étendre notre recherche sur ces types d'algorithmes.

Résultats					
Taille de vecteur	epoch	alpha	window	Rappel Global 10	Rappel Global 100
20	10	0,025	5	18.40%	61.30%
20	50	0,025	5	15.90%	60.70%
30	10	0,025	5	15.60%	58.60%
30	50	0,025	5	16.50%	62.80%
60	50	0,025	3	14.00%	56.20%
60	50	0,025	2	13.70%	58.20%
60	80	0,025	2	14.60%	58.80%

Tableau 7.13 Rappel sur les résultats globaux pour les 10 et 100 premières recommandations sur les modèles DOC2VEC avec taille du vecteur et hyperparamètres sans augmentation de données.

En appliquant la méthode commune de «Doc2Vec» on remarque avec le tableau 7.12 que nous arrivons à des rappels globaux sur les 100 premières recommandations qui varient de 58.8% a 61.3% en fonction des hyperparamètres, et que sur les 10 premiers varient de 14.6% à 18.4% de rappel global. Notre hypothèse est que cela est dû à l'entraînement et à l'expérimentation d'hyperparamètres qui changent constamment les recommandations dont nous estimons ne pas être la meilleure approche pour créer un modèle de recommandation de compagnie comparable qui est fiable pour un investisseur. Nous détaillerons par après un exemple concret afin d'explorer le comportement de nos meilleurs modèles de cette section.

Dans le tableau 7.17, nous présentons nos résultats des modèles exploratoires avec augmentation des données selon un plongement des mots des descriptions «word-embeddings» afin

Résultats					
Taille de vecteur	epoch	alpha	window	Rappel Global 10	Rappel Global 100
20	10	0,025	5	15.96%	57.94%
20	50	0,025	5	16.00%	52.38%

Tableau 7.14 Rappel sur les résultats globaux pour 10 et 100 premières recommandations sur le modèle avec plongement des descriptions augmentées.

de pouvoir augmenter les descriptions selon les mots les plus proches qui ne seraient pas présents directement dans les descriptions. Nous remarquons que les résultats ne sont pas concluants et que ce modèle ne donne pas de meilleurs résultats que les modèles exploratoires sans augmentations. En effet, nous obtenons 18.4% de rappel global moyen dans le modèle exploratoire avec les mêmes hyperparamètres et taille de vecteur de 20 sans augmentation des données et 15.96% avec l'augmentation avec les 10 premières recommandations et respectivement 61.3% et 57.94% pour les 100 premières recommandations sans augmentation et avec augmentation des descriptions. De ces résultats nous en concluons que l'augmentation des données ne permet pas d'avoir un meilleur taux de rappel global moyen.

7.4.1 Investigation Doc2Vec

Afin d'investiguer davantage les résultats des modèles avec «Doc2Vec», on a pris le meilleur modèle D2V sur les 100 premières recommandations (vecteur 30, epoch50, alpha 0,025, window 5) et regardé les comparables de 3D systèmes pour mieux comprendre le comportement du modèle «Doc2Vec» par rapport aux modèles avec transformations TD-IDF dont la transformation TF-IDF et ses avantages sont expliqués à la section 3.2.

SYMBOLE	Similarité
MATW	0.760
ARCW	0.752
XONE	0.724
SXI	0.713
COHR	0.691
PRLB	0.689
EFII	0.6841
AMAT	0.6817

Tableau 7.15 Liste des 9 compagnies comparables a 3D Systems selon le modèle avec vecteur de taille 30 (50 epoch, alpa 0,025, Window 5)

Le modèle «Doc2Vec» peut recommander des compagnies qui n'ont aucune pertinence dans la recherche de comparables dont nous avons observé manuellement plusieurs cas, mais n'en présenterons qu'un sur «DDD» ou on l'observera qu'il n'y a aucune coïncidence avec les compagnies de notre corpus de validation. La première compagnie comparable, selon le modèle de «Doc2Vec», est Matthews International Corporation. Voici la description abrégée de cette compagnie, tel qu'elle y figure dans la base de données :

> *«Matthews International Corp. engages in the provision of brand solutions, memorialization products, and industrial products. It operates through the following segments : SGK Brand Solutions, Memorialization, Industrial, and Others. The SGK Brand Solutions segment involves in graphics imaging business, including Schawk, and the merchandising solutions operations. The Memorialization segment offers cemetery products, funeral home products, and cremation operations. The Industrial segment offers includes company's marking and automation products and fulfillment systems. The company was founded by John Dixon Matthews in 1850 and is headquartered in Pittsburgh, PA.»*

En bref, cette entreprise est constituée de trois compagnies, toutes différentes de 3D Systems. La première entreprise est dans l'impression de produits publicitaires, la seconde, dans les

produits d'obsèques et enfin, la troisième se consacre aux produits d'automatisation d'entrepôts.

Voici, pour notre seconde entreprise, ARC Group Worldwide, sa description abrégée telle qu'elle y figure dans notre base de données :

> *«ARC Group Worldwide, Inc. engages in the development and provision of wireless network components and solutions. It operates through the following segments : Precision Components ; 3DMT ; Flanges and Fittings ; and Wireless. The Precision Components segment produces fabricated metal components through metal injection molding, precision metal stamping, and hermetic sealing. The 3DMT segment offers rapid prototyping, short-run production, and rapid tooling with its three-dimensional printing and additive manufacturing operations solutions. The Flanges and Fittings segment consists of custom machining services and special flange facings. The Wireless segment designs and manufactures antennas, radios, and accessories used in broadband networks. The company was founded on September 30, 1987 and is headquartered in Deland, FL.»*

Cette entreprise, quant à elle, se spécialise dans l'impression de métaux à partir de moules qui présente une forme l'impression qui est plus industrielle et non technologique comme 3D Systems. Cette différence expose à un risque d'erreur avec l'utilisation du modèle «Doc2Vec» sur les dix premiers comparables, en plus que ce modèle ne s'adapte pas facilement. Par conséquent, l'utilisation d'une méthode TF-IDF avec filtrage des descriptions nous aura semblé être plus adaptée, à l'inverse d'un modèle à entraîner. C'est qu'en effet, ce dernier rallonge le temps, en raison de son paramétrage et de son entraînement et nécessite de répéter cela lorsque les données des entreprises changent. D'où un travail plus long pour des erreurs flagrantes n'est pas justifié par rapport aux méthodes avec transformation TF-IDF qui donnent d'aussi bons résultats. Aussi, ayant généralement de petits corpus pour l'entraînement (2990 compagnies) et uniquement une description par compagnies, une méthode avec une transformation et une analyse plus contrôlée telle que TF-IDF est plus appropriée.

7.5 Modèle combiné et régression

Dans tous nos modèles présentés dans le chapitre 5, nous n'avons considéré aucun poids pour chaque source de données. Nous avons cependant exploré l'optimisation des paramètres afin de connaître l'effet sur les résultats de rappel selon des régressions. Pour cet exercice, nous avons pris notre meilleur modèle, balancé les données et effectué une régression logistique afin de retrouver les poids optimaux sur chaque paramètres du modèle d'algorithme à partir la librairie Scikit-learn (Pedregosa et al., 2011) avec la fonction «sklearn.linear_model.LogisticRegression ». La moitié des données est avec les comparables cibles avec les colonnes des paramètres et la dernière colonne si la compagnie est comparable (1 ou 0). L'autre moitié des données est sélectionnée au hasard au même ratio que le nombre de comparable par entreprise avec des compagnies qui ne sont pas comparables. Par exemple pour la compagnie «DDD» qui à deux comparables, on va sélectionner les comparable et au hasard deux compagnies qui ne sont pas comparables afin d'estimer les paramètres de régression.

Notre modèle le plus performant est le modèle avec graphique de connaissances «WordNet» (MCFDWN). Ce modèle est formulé par le calcul suivant lorsqu'on ajoute les poids aux paramètres :

$$R_6 = W_{cat}R_{Cat} + W_{cap}R_{cap} + W_{des}R_{des} + W_{Wdis}R_{Wdis} + W_{Wp}R_{Wp} \tag{7.1}$$

où :

- R_6 : Score de similarité selon catégorie d'industrie, capitalisation descriptions brutes , noms verbes des descriptions désambiguïsées et similarité WordNet.
- R_{cat} : Score de similarité catégorie d'industrie. $R_{cat} \in [-1, 1]$
- R_{cap} : Score de similarité Capitalisation. $R_{cap} \in [0, 1]$
- R_{des} : Score de similarité descriptions TF-IDF. $R_{des} \in [-1, 1]$
- R_{Wdis} : Score de similarité noms verbes desambiguées TF-IDF. $R_{Wdis} \in [-1, 1]$
- R_{Wp} : Score de similarité entre les noms et verbes desambiguées des descriptions selon le graphique de connaissance WordNet. $R_{Wp} \in [0, 1]$

Dans le tableau 7.16 nous présentons les résultats des dix régressions logistiques avec chacune des données balancées.

Résultats régression logistique						
Essai	W_{cat}	W_{fin}	W_{des}	W_{Wdis}	W_{Wp}	R^2
1	2.351	1.674	7.390	13.001	4.148	0.629
2	2.390	1.661	7.381	12.236	4.217	0.617
3	2.389	1.686	7.446	13.053	4.215	0.628
4	2.344	1.589	7.409	12.291	4.310	0.627
5	2.469	1.730	7.671	11.978	4.409	0.633
6	2.402	1.664	7.766	12.644	3.829	0.629
7	2.405	1.719	7.718	12.245	4.533	0.638
8	2.332	1.638	7.653	12.885	4.116	0.632
9	2.377	1.704	7.138	12.730	4.252	0.624
10	2.370	1.698	7.229	12.423	4.221	0.627
Moyenne	2.383	1.676	7.480	12.549	4.225	0.628

Tableau 7.16 Résultats des essais de régression logistique

Pour chaque paramètre nous avons cherché par régression logistique les poids W d'où nous obtenons à partir de 10 essais les paramètres optimaux moyens suivant :

W_{cat} : catégories $W_{cat} = 2.383$

W_{cap} : capitalisation $W_{cap} = 1.676$

W_{des} : descriptions $W_{des} = 7.480$

W_{Wdis} : noms et verbes desambiguées $W_{Wdis} = 12.549$

W_{Wp} : «PathSimilarity» de similarité avec WordNet. $W_{Wp} = 4.225$

Lorsque nous analysons la précision de la régression sur nos données balancées, nous obtenons 0.895 de précision. Dans le tableau 7.17 nous présentons l'analyse des coefficients d'où nous pouvons remarquer que les variables sont toutes significatives.

Coefficients				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.12115	0.14729	-41.559	< 2e-16
R_{Cat}	2.35684	0.04994	47.197	< 2e-16
R_{Cap}	1.73055	0.11675	14.822	< 2e-16
R_{des}	6.74758	0.30941	21.808	< 2e-16
R_{Wdis}	19.11058	0.87284	21.895	< 2e-16
R_{Wp}	2.02116	0.54053	3.739	0.000185

Tableau 7.17 Analyse des Variables

Résultats de rappel global avec régression logistique

Dans cette section nous présentons les résultats de rappels sur nos modèles avec les variables optimales de la régression en utilisant toutes les données.

Dans le tableau 7.18, on remarque qu'avec la régression logistique sur notre meilleur modèle, on n'arrive pas à obtenir de meilleurs résultats de rappel pour les 10 premières recommandations cependant selon les 100 premières recommandations le rappel augmente de 79.5% à 81.09%.

Nombre de recom.	Rappel Global : Catégorie Cap. Description (WORDNET) Régression Logistique
10	28.09%
20	43.81%
30	54.62%
40	61.90%
50	67.51%
60	71.21%
70	74.97%
80	77.80%
90	79.42%
100	81.09%

Tableau 7.18 Rappel pour le modèle avec la catégorie d'industrie, la capitalisation, la description et le graphique de connaissance pour la similarité et les descriptions désambiguïsée avec régression logistique

Dans le tableau 7.19 nous présentons la matrice de corrélation entre tous les paramètres utilisés dans les différents modèles.

Variables	R_{cat}	R_{cap}	R_{des}	R_{dis}	R_{pat}	R_{nom}	R_{ver}	R_{vc}
R_{cat}	1.000	0.020	0.256	0.302	0.240	0.323	0.043	0.043
R_{cap}	0.020	1.000	0.015	0.048	0.009	0.039	0.002	0.019
R_{des}	0.256	0.015	1.000	0.639	0.641	0.804	0.251	0.101
R_{dis}	0.302	0.048	0.639	1.000	0.789	0.786	0.191	0.108
R_{pat}	0.240	0.009	0.641	0.789	1.000	0.662	0.189	0.086
R_{nom}	0.323	0.039	0.804	0.786	0.662	1.000	0.166	0.119
R_{ver}	0.043	0.002	0.251	0.191	0.189	0.166	1.000	0.020
R_{vc}	0.043	0.019	0.101	0.108	0.086	0.119	0.020	1.000

Tableau 7.19 Matrice de corrélation entre les variables

7.6 Récapitulation des résultats

Dans cette section nous présentons un résumé de tous les résultats obtenus pour les différents modèles.

Nombre de recom.	Catégories Cap. (MCF)	Catégories Cap. Description (MCFD)	Catégories Cap. Description (Noms) (MCFDn)	Catégories Cap. Description (Verbe) (MCFDv)	Catégories Cap. Description (Groupe Verbe) (MCFDgv)	Catégories Cap. Description (Wornet) (MCFDWN)	Moyenne
10	26.4%	28.9%	29.5%	24.3%	26.6%	**29.9%**	27.60%
20	40.3%	43.8%	44.6%	39.3%	42.1%	**47.0%**	42.85%
30	49.3%	52.5%	53.2%	47.9%	51.5%	**56.2%**	51.76%
40	54.0%	57.6%	58.2%	54.4%	57.8%	**63.6%**	57.60%
50	57.7%	62.8%	63.8%	58.6%	62.1%	**68.2%**	62.20%
60	61.0%	66.4%	67.0%	62.4%	66.1%	**71.1%**	65.66%
70	64.0%	68.7%	69.7%	65.1%	68.3%	**74.4%**	68.36%
80	66.2%	71.0%	72.1%	67.7%	70.3%	**76.1%**	70.56%
90	68.6%	73.4%	74.3%	70.3%	72.1%	**77.7%**	72.73%
100	69.9%	75.4%	76.4%	72.8%	73.9%	**79.5%**	74.65%

Tableau 7.20 Rappel sur les résultats globaux pour les principaux modèles

Dans le tableau récapitulatif, nous présentons nos données des différents modèles et les comparons au modèle de base (MCF). De manière générale, tous les modèles contenant la description permettent d'obtenir de meilleur rappel que le modèle de base. Les meilleurs rappels proviennent du modèle (MCFDWN) avec le graphique de connaissance «WordNet» avec un rappel global atteignant 79.5% sur les 100 premières recommandations. Selon notre étude, le meilleur modèle est le modèle MCFDWN avec le calcul TF-IDF avec la capitalisation, le secteur d'activité, le groupe d'industrie, la description, la description désambiguïsée et le calcul de similarité avec le graphique de connaissance «WordNet».

CHAPITRE 8 CONCLUSION

Pour conclure, nous avons été en mesure de créer des modèles de recommandation de compagnies comparables avec des données des entreprises publiques en mesurant le rappel sur une liste d'entreprises comparables cibles. Nous avons été en mesure de déterminer que nos modèles avec la description des entreprises ont permis d'obtenir de meilleurs résultats que notre modèle de base. Le modèle de recommandation avec le meilleur rappel moyen est le modèle MCFDWN présenté à la section 5.6 du mémoire avec la transformation TF-IDF sur les descriptions, la capitalisation, le secteur d'activité, le groupe d'industrie, le graphique de connaissance «WordNet» pour la désambiguïsation et pour la similarité selon la distance des chemins parcourus dans les graphiques de connaissance.

8.1 Synthèse des travaux

À partir de ce projet, plusieurs domaines de recherche ont été étudiés et regroupés. Le domaine de la finance et les compagnies comparables pour un analyste investisseur ont été étudiés. Le domaine de la recherche d'information, le domaine de traitement du langage naturel pour l'analyse et la réduction de texte des descriptions des entreprises ont été étudiés. Le domaine des graphiques de connaissance a aussi été utilisé pour la recherche de similarité des descriptions selon le «Wordnet».

Nous avons inclus des concepts des outils courants de recherche de compagnies comparables telle que l'utilisation des systèmes de code des entreprises (code NAICS). Nous avons aussi inclus des techniques de traitement des langues et du filtrage de l'information telles que : l'étiquetage sémantique, le plongement des mots, la fréquence inverse des mots, la troncature des mots à la racine et les graphiques de connaissance «Wordnet».

Nous avons par la suite présenté les différentes étapes de création de nos modèles de recommandations de compagnies comparables selon les sources de données des entreprises. Par après, nous avons utilisé la méthode du rappel pour comparer nos modèles d'algorithmes de recommandations entre eux avec notre corpus de validation de compagnies cibles.

8.2 Limitations de la solution proposée

Bien que nos approches ont permis d'apporter de nouvelles méthodes permettant de créer des recommandations d'entreprises comparables, certaines limitations sont présentes.

Une des limites de notre recherche est au niveau du corpus de validation qui nous limite dans notre recherche de modèle optimal. En effet, nous retrouvons dans notre corpus de validation des compagnies qui varient beaucoup par leurs catégories, capitalisation et par leurs descriptions. Cette grande variation ne permettent pas nos modèles de retrouver les compagnies cibles parmi nos 100 premières recommandations ce qui limite notre taux de rappel global. Nous pouvons nous retrouver dans des situations où plus de 100 entreprises sont recommandées avant de retrouver les comparables cibles du corpus de validation ce qui réduit nos taux de rappel.

Nous étions en mesure d'obtenir des descriptions des entreprises à partir de la base de données d'Intrinio US Fundamentals and Stock Prices, mais plus d'information sur les entreprises, telle qu'une source d'information supplémentaire qui décrit plus en détail les opérations de chaque entreprise, aurait hypothétiquement permis d'avoir des algorithmes plus performants.

L'analyse de compagnies comparables, après avoir sélectionné les compagnies, est un processus qui nécessite plusieurs informations comptables fondamentales — telles que les profits — afin qu'il soit possible de comparer les entreprises et de valoriser la compagnie qu'on analyse. Nos modèles sont limités aux entreprises publiques dans les bourses américaines dont les informations fondamentales sont aussi accessibles, car les entreprises privées ne sont pas obligées de partager publiquement ces informations. Par exemple, il n'est pas nécessaire à une entreprise privée de communiquer publiquement de ses activités ou de présenter ses livres comptables selon un standard. Plus les entreprises sont soumises à des législations qui diffèrent, plus elles sont difficilement comparables financièrement, car elles nécessiteraient une compréhension des différences législatives et comptables avant de parvenir à des informations comparables des deux entreprises.

Notre modèle est aussi limité aux entreprises publiques présent dans les marchés américains, il est difficile de valoriser une compagnie dans le marché canadienne à une compagnie dans le marché chinoise sans connaître les différences entre les deux pays. D'où les entreprises enregistrées dans un même marché public ou opérant dans la même région sont l'option la moins contraignante. Cela facilite l'accès à leurs informations, car elles sont publiques, les standards de présentation sont les mêmes pour toutes les entreprises dans un marché.

8.3 Améliorations futures

Plusieurs améliorations futures permettraient d'avoir des modèles d'algorithme de recommandation de comparables financiers plus performant. Dans un premier temps, l'optimisation de l'utilisation des graphiques de connaissances en comparant les différentes possibilités de re-

cherche de similarité entre les descriptions, par exemple l'utilisation des autres méthodes de mesure de similarité de chemin présenté à la section 3.4 comme la méthode de similarité de «Wup» (Wu and Palmer, 1994).

Une autre amélioration possible est d'adapter les modèles pour un domaine spécifique d'entreprise ou une caractéristique spécifique financière telle que la croissance des revenus ou le taux de dette.

D'autres améliorations possibles sont dans le choix des sources d'information à utiliser dans les modèles d'algorithmes de recommandations et des comparables cibles. Ces informations pourraient être financières fondamentales telles qu'une nouvelle métrique sur les entreprises ou linguistique telles que les textes des rapports annuels des entreprises.

RÉFÉRENCES

Albuquerque, A. M., De Franco, G., and Verdi, R. S. (2013). Peer choice in ceo compensation. *Journal of Financial Economics*, 108(1) :160–181.

Alford, A. W. (1992). The effect of the set of comparable firms on the accuracy of the price-earnings valuation method. *Journal of Accounting Research*, 30(1) :94–108.

Analytics, C. (2015). Anaconda software distribution. computer software. vers. 2-2.4. 0.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Bradshaw, M. T., Miller, G. S., Serafeim, G., et al. (2009). Accounting method heterogeneity and analysts' forecasts. *Unpublished paper, University of Chicago, University of Michigan, and Harvard University*.

Chen, W.-K., editor (2009). *Analog and VLSI Circuits : The Circuits and Filters Handbook*. CRC Press, Boca Raton, FL, third edition.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

De Franco, G., Hope, O.-K., and Larocque, S. (2015). Analysts' choice of peer companies. *Review of Accounting Studies*, 20(1) :82–109.

De Franco, G., Kothari, S. P., and Verdi, R. S. (2011). The benefits of financial statement comparability. *Journal of Accounting Research*, 49(4) :895–931.

De Marneffe, M.-C. and Manning, C. D. (2008). The stanford typed dependencies representation. In *Coling 2008 : proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Deev, O. (2011). Methods of bank valuation : a critical overview. *Financial Assets and Investing*, 3(3) :33–44.

Dittmann, I. and Weiner, C. (2005). Selecting comparables for the valuation of european firms. *Available at SSRN 644101*.

Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48.

Fellbaum, C. (2012). Wordnet. *The encyclopedia of applied linguistics*.

Fris, P. and Gonnet, S. (2010). The state of the art in comparability for transfer pricing. *International Tranfer pricing Journal*, pages 99–106.

Gilson, S. C., Hotchkiss, E. S., and Ruback, R. S. (2000). Valuation of bankrupt firms. *The Review of Financial Studies*, 13(1) :43–74.

Gkotsis, P., Pugliese, E., and Vezzani, A. (2018). A technology-based classification of firms : Can we learn something looking beyond industry classifications ? *Entropy*, 20(11) :887.

Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13) :13–18.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5) :1423–1465.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J. E. L., Kirrane, S., Neumaier, S., Polleres, A., et al. (2020). Knowledge graphs. *arXiv preprint arXiv :2003.02320*.

Hoitash, R., Kogan, A., and Vasarhelyi, M. A. (2006). Peer-based approach for analytical procedures. *Auditing : A Journal of Practice & Theory*, 25(2) :53–84.

Kaplan, S. N. and Ruback, R. S. (1995). The valuation of cash flow forecasts : An empirical analysis. *The journal of Finance*, 50(4) :1059–1093.

Kee, T. (2019). Peer firm identification using word embeddings. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5536–5543. IEEE.

Kee, T. et al. (2018). A text-based approach to industry classification.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Leacock, C., Chodorow, M., and Miller, G. A. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1) :147–165.

Lee, C. M., Ma, P., and Wang, C. C. (2015). Search-based peer firms : Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, 116(2) :410–431.

Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems : State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics : system demonstrations*, pages 55–60.

Meitner, M. (2006). *The market approach to comparable company valuation*, volume 35. Springer Science & Business Media.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

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

Mitkov, R. (2003). *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)*. Oxford University Press, Inc., New York, NY, USA.

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing : an introduction. *Journal of the American Medical Informatics Association*, 18(5) :544–551.

NAICS Association (2019). Naics codes drilldown table.

Newman-Griffis, D. and Fosler-Lussier, E. (1970). [pdf] second-order word embeddings from nearest neighbor topological features : Semantic scholar.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

Princeton, U. (2007). Wordnet. *The Trustees of Princeton University*. `https://www.cs.princeton.edu/courses/archive/spring07/cos226/assignments/wordnet.html`.

Raman, N., Bang, G., and Nematzadeh, A. (2019). Multigraph attention network for analyzing company relations. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 426–433.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 69(1) :99–118.

Singhal, A. (2012). Introducing the knowledge graph : things, not strings. *Official google blog*, 5.

Sun, Z., Guo, Q., Yang, J., Fang, H., Guo, G., Zhang, J., and Burke, R. (2019). Research commentary on recommendations with side information : A survey and research directions. *Electronic Commerce Research and Applications*, 37 :100879.

US Bureau of labor statistic (2019). The north american industry classification system in the current employment statistics program.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244.

Weerasooriya, T., Perera, N., and Liyanage, S. (2017). Keyxtract twitter model-an essential keywords extraction model for twitter designed using nlp tools. *arXiv preprint arXiv :1708.02912*.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Xing, W., Yuan, X., Li, L., Hu, L., and Peng, J. (2018). Phenotype extraction based on word embedding to sentence embedding cascaded approach. *IEEE Transactions on NanoBioscience*.

ANNEXE A LOGICIEL DE CALCUL ET D'ANALYSE

Python

Python est un langage de programmation interprété qui permet beaucoup de puissance grâce à sa gestion automatique de la mémoire, des erreurs et de tapage dynamique. Ce langage de programmation permet un orienté objet simple, l'utilisation de librairie puissante pour le traitement des données, le traitement des langues naturelles et les systèmes de recommandations.
`https://www.python.org/`

Natural Language Toolkit

NLTK est une librairie entière pour le traitement des langues naturelles avec python. Il offre plus de 50 dictionnaire lexicaux et d'outils d'analyse tel que la séparation des mots 'tokenization', la classification, le l'étiquetage des mots par leur classe grammaticale.

`https://www.nltk.org/`

Gensim

Gensim est une librairie de python d'analyse sémantique pour retrouver la structure d'un document, pour faire des analyses statistiques de sémantique et pour retrouver des textes similaires incluant multiples fonctions diverses.

`https://radimrehurek.com/gensim/`

Anaconda

Anaconda est un logiciel qui permet de faciliter la gestion des librairies , des environnements et des distributions de Python avec plus de 1500 librairies à accès libre. (Analytics, 2015)

`https://www.anaconda.com/`

ANNEXE B INFORMATIONS DES ENTREPRISES RECOMMANDÉES

Dans le tableau B.1 on présente les différents groupes d'industrie dans nos données avec leur fréquence.

Groupe d'industries

Tableau B.1 Distribution des groupes d'industrie

Groupe d'industrie	**Fréquence**
Biotechnology	250
Application Software	78
Business Services	60
Property & Casualty Insurance	58
Regional - Mid-Atlantic Banks	57
Regional - Northeast Banks	57
Diversified Machinery	56
Independent Oil & Gas	54
Medical Instruments & Supplies	53
REIT - Diversified	51
Savings & Loans	49
Medical Appliances & Equipment	48
Asset Management	39
Regional - Midwest Banks	38
Communication Equipment	37
Restaurants	37
Specialty Chemicals	36
Medical Laboratories & Research	35
Regional - Pacific Banks	35
Credit Services	34
Drugs - Generic	33
Oil & Gas Equipment & Services	32
Oil & Gas Pipelines	31

Suite dans la prochaine page

Tableau B.1 – *Suite de la page précédente*

Catégories Distribution	**Fréquence**
Scientific & Technical Instruments	31
Auto Parts	29
Business Software & Services	29
Electric Utilities	28
Diversified Electronics	27
General Building Materials	27
Internet Information Providers	27
REIT - Retail	27
Specialty Retail, Other	27
Apparel Stores	26
Aerospace/Defense Products & Services	24
Industrial Electrical Equipment	24
REIT - Residential	24
Regional - Southwest Banks	22
Semiconductor Equipment & Materials	22
Money Center Banks	21
Semiconductor - Integrated Circuits	21
Information Technology Services	20
Rental & Leasing Services	20
Chemicals - Major Diversified	19
Investment Brokerage - National	19
Packaging & Containers	19
Personal Products	19
Diversified Utilities	18
Oil & Gas Refining & Marketing	18
Property Management	17
Regional - Southeast Banks	16
REIT - Office	16
Residential Construction	16
Resorts & Casinos	16
Semiconductor - Broad Line	16
Staffing & Outsourcing Services	16

Suite dans la prochaine page

Tableau B.1 – *Suite de la page précédente*

Catégories Distribution	**Fréquence**
Textile - Apparel Clothing	16
Waste Management	16
Drug Manufacturers - Major	15
Healthcare Information Services	15
Home Furnishings & Fixtures	15
Industrial Equipment Wholesale	15
Life Insurance	15
Mortgage Investment	15
Processed & Packaged Goods	15
REIT - Industrial	15
Semiconductor - Specialized	15
Steel & Iron	15
Textile - Apparel Footwear & Accessories	15
Trucking	15
Gas Utilities	14
Oil & Gas Drilling & Exploration	14
REIT - Healthcare Facilities	14
Technical & System Software	14
Diversified Communication Services	13
Education & Training Services	13
Entertainment - Diversified	13
Management Services	13
Water Utilities	13
Diagnostic Substances	12
Farm & Construction Machinery	12
Food - Major Diversified	12
Industrial Metals & Minerals	12
Real Estate Development	12
REIT - Hotel/Motel	12
Air Delivery & Freight Services	11
Industrial Equipment & Components	11
Wireless Communications	11

Suite dans la prochaine page

Tableau B.1 – *Suite de la page précédente*

Catégories Distribution	**Fréquence**
Agricultural Chemicals	10
Discount, Variety Stores	10
Farm Products	10
Heavy Construction	10
Metal Fabrication	10
Railroads	10
Technical Services	10
Telecom Services - Domestic	10
Auto Dealerships	9
Computer Peripherals	9
Health Care Plans	9
Internet Software & Services	9
Networking & Communication Devices	9
Recreational Vehicles	9
Rubber & Plastics	9
Small Tools & Accessories	9
Accident & Health Insurance	8
Beverages - Soft Drinks	8
Broadcasting - Radio	8
Broadcasting - TV	8
Business Equipment	8
Catalog & Mail Order Houses	8
Drug Manufacturers - Other	8
Food Wholesale	8
Hospitals	8
Insurance Brokers	8
Lodging	8
Lumber, Wood Production	8
Paper & Paper Products	8
Pollution & Treatment Controls	8
Specialized Health Services	8
Sporting Goods	8

Suite dans la prochaine page

Tableau B.1 – *Suite de la page précédente*

Catégories Distribution	**Fréquence**
Conglomerates	7
Electronics Wholesale	7
Investment Brokerage - Regional	7
Marketing Services	7
Multimedia & Graphics Software	7
Personal Services	7
Security & Protection Services	7
Auto Manufacturers - Major	6
CATV Systems	6
Data Storage Devices	6
Department Stores	6
Diversified Investments	6
Electronic Equipment	6
Information & Delivery Services	6
Machine Tools & Accessories	6
Printed Circuit Boards	6
Publishing - Newspapers	6
Regional Airlines	6
Shipping	6
Trucks & Other Vehicles	6
Drug Related Products	5
General Contractors	5
Grocery Stores	5
Home Furnishing Stores	5
Home Health Care	5
Recreational Goods, Other	5
Specialty Eateries	5
Surety & Title Insurance	5
Textile Industrial	5
Advertising Agencies	4
Beverages - Brewers	4
Cleaning Products	4

Suite dans la prochaine page

Tableau B.1 – *Suite de la page précédente*

Catégories Distribution	**Fréquence**
Confectioners	4
Diversified Computer Systems	4
Drug Delivery	4
Drug Stores	4
Gaming Activities	4
Gold	4
Home Improvement Stores	4
Housewares & Accessories	4
Long-Term Care Facilities	4
Major Airlines	4
Research Services	4
Security Software & Services	4
Semiconductor- Memory Chips	4
Sporting Activities	4
Sporting Goods Stores	4
Aerospace/Defense - Major Diversified	3
Aluminum	3
Auto Parts Stores	3
Beverages - Wineries & Distillers	3
Building Materials Wholesale	3
Cigarettes	3
Computer Based Systems	3
Computers Wholesale	3
Drugs Wholesale	3
Electronics Stores	3
General Entertainment	3
Jewelry Stores	3
Major Integrated Oil & Gas	3
Meat Products	3
Medical Equipment Wholesale	3
Movie Production, Theaters	3
Processing Systems & Products	3

Suite dans la prochaine page

Tableau B.1 – *Suite de la page précédente*

Catégories Distribution	**Fréquence**
Toys & Games	3
Air Services, Other	2
Appliances	2
Copper	2
Foreign Money Center Banks	2
Foreign Regional Banks	2
Internet Service Providers	2
Nonmetallic Mineral Mining	2
Office Supplies	2
Publishing - Books	2
Publishing - Periodicals	2
Silver	2
Tobacco Products, Other	2
Auto Parts Wholesale	1
Basic Materials Wholesale	1
Cement	1
Closed-End Fund - Equity	1
Long Distance Carriers	1
Music & Video Stores	1
Photographic Equipment & Supplies	1
Synthetics	1
Toy & Hobby Stores	1

Informations des entreprises

La liste ci-dessous contient les informations des entreprises dont le mémoire fait mention provenant de la source de donnée de «Intrinio US Fundamentals and Stock Prices» en ordre alphabétique.

Compagnie : ABM
Nom : ABM Industries Incorporated
Secteur d'activité : Services
Groupe d'industrie : Business Services
Capitalisation (Million) : 2268.89556

Description :

« ABM Industries Incorporated provides integrated facility solutions in the United States and internationally. It offers carpet cleaning and dusting, floor cleaning and finishing, window washing, and other building cleaning services for commercial office buildings, data centers, educational institutions, government buildings, health facilities, industrial buildings, retail stores, sport event facilities, and transportation hubs. The company also provides onsite mechanical engineering and technical services and solutions relating to a range of facilities and infrastructure systems ; and parking and transportation services for clients at various locations, including commercial office buildings, educational institutions, health facilities, hotels, sport event facilities, and transportation hubs. In addition, it offers custom energy solutions, HVAC, electrical, lighting, and other general maintenance and repair services comprising bundled energy solutions, energy efficiency upgrades, installations, preventative maintenance, retro-commissioning, and retrofits for clients in the private and public sectors ; construction management, energy efficiency upgrades, healthcare support, leadership development, military base operations, and other mission support to the U.S. government entities ; and facility management and environmental, food and nutrition, healthcare technology management, and patient and guest services to healthcare systems and hospitals. Further, the company franchises engineering services under the Linc Service and TEGG brands through individual and area franchises ; and provides facility solutions to airlines and airports related to access control, aircraft cabin cleaning, shuttle bus operations, and passenger assistance. The company was founded in 1909 and is headquartered in New York, New York. »

Compagnie : ACM
Nom : AECOM
Secteur d'activité : Services
Groupe d'industrie : Technical Services
Capitalisation (Million) : 5597.53546
Description :

« AECOM, together with its subsidiaries, engages in designing, building, financing, and operating infrastructure assets worldwide. The company operates through three segments : Design and Consulting Services (DCS), Construction Services (CS), and Management Services (MS). The DCS segment provides planning, consulting, architectural and engineering design, program management, and construction management services for industrial, commercial, institutional, and govern-

ment clients, such as transportation, facilities, environmental, and energy/power markets. The CS segment offers building construction and energy, as well as infrastructure and industrial construction services. The MS segment provides program and facilities management and maintenance, training, logistics, consulting, technical assistance, and systems integration and information technology services primarily for agencies of the U.S. government and other national governments. The company was formerly known as AECOM Technology Corporation and changed its name to AECOM in January 2015. AECOM was founded in 1980 and is headquartered in Los Angeles, California. »

Compagnie : ALOT
Nom : AstroNova, Inc.
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals
Capitalisation (Million) : 106.48847
Description :

« AstroNova, Inc. designs, develops, manufactures, and distributes specialty printers, and data acquisition and analysis systems in the United States, Canada, and Western Europe. The company operates in two segments, Product Identification and Test Measurement. The Product Identification segment offers digital color label printers and specialty OEM printing systems ; and consumables, such as labels, tags, inks, toner, and thermal transfer ribbons, as well as software used to design and print labels under the QuickLabel brand. It serves the chemicals, cosmetics, food and beverage, medical products, and pharmaceuticals, and other industries. The Test Measurement segment provides Daxus portable data acquisition systems ; TMX high-speed data acquisition systems ; DDX100 SmartCorder portable data acquisition systems ; EVX multi-channel chart recording systems ; ToughWriter, Miltope, and RITEC branded airborne printers ; and ToughSwitch ruggedized Ethernet switches. The company's AstroNova airborne printers are used in flight decks, as well as military, commercial, and business aircraft cabins ; ToughSwitch Ethernet switches are used in military aircraft and vehicles ; ToughWriter airborne printers are used in aircraft made by Airbus, Boeing, Embraer, Bombardier, Lockheed, Gulfstream, and others ; and portable data recorders are used in research and development, and maintenance applications in aerospace and defense, energy discovery and production, rail, automotive, and transportation and other industrial applications. Its TMX data acquisition systems are used for data capture in long-term testing ; and Daxus and DDX 100 SmartCorder instruments

are used for portability and ease of use in facilities maintenance, field work, test cells, and transportation applications. The company was formerly known as Astro-Med, Inc. and changed its name to AstroNova, Inc. in May 2016. AstroNova, Inc. was founded in 1969 and is headquartered in West Warwick, Rhode Island. »

Compagnie : AMAT
Nom : Applied Materials, Inc.
Secteur d'activité : Technology
Groupe d'industrie : Semiconductor Equipment Materials
Capitalisation (Million) : 34740.91822
Description :

« Applied Materials, Inc. provides manufacturing equipment, services, and software to the semiconductor, display, and related industries worldwide. It operates through three segments : Semiconductor Systems, Applied Global Services, and Display and Adjacent Markets. The Semiconductor Systems segment develops, manufactures, and sells a range of manufacturing equipment used to fabricate semiconductor chips or integrated circuits. It offers products and technologies for transistor and interconnect fabrication, including epitaxy, ion implantation, oxidation and nitridation, rapid thermal processing, chemical vapor deposition, physical vapor deposition, chemical mechanical planarization, and electrochemical deposition ; patterning, selective removal, and packaging products and systems that enable the transfer of patterns onto device structures ; and metrology, inspection, and review systems for front- and back-end-of-line applications. The Applied Global Services segment provides integrated solutions to optimize equipment and fab performance and productivity, including spares, upgrades, services, remanufactured earlier generation equipment, and factory automation software for semiconductor, display, and other products. The Display and Adjacent Markets segment offers products for manufacturing liquid crystal displays, organic light-emitting diodes, and other display technologies for TVs, personal computers, tablets, smart phones, and other consumer-oriented devices, as well as equipment for flexible substrates. The company serves manufacturers of semiconductor wafers and chips, liquid crystal and other displays, and other electronic devices. Applied Materials, Inc. was founded in 1967 and is headquartered in Santa Clara, California. »

Compagnie : AME
Nom : AMETEK, Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery

Capitalisation (Million) : 11247.59403
Description :

« AMETEK, Inc. manufactures electronic instruments and electromechanical devices worldwide. Its Electronic Instruments Group segment offers advanced instruments for the process, power and industrial, and aerospace markets ; process and analytical instruments for the oil, gas, petrochemical, pharmaceutical, semiconductor, and automation markets ; instruments for the laboratory equipment, ultraprecision manufacturing, medical, and test and measurement markets ; and vision systems to inspect surfaces. This segment also provides aircraft and engine sensors, monitoring systems, power instruments, data acquisition units, and fuel and fluid measurement systems for the aerospace industry ; power quality monitoring and metering devices, industrial battery chargers, uninterruptible power supplies, programmable power and electrical test equipment, and gas turbine sensors ; and dashboard instruments for heavy trucks and other vehicles, as well as timing controls and cooking computers for the food service industry. Its Electromechanical Group segment offers thermal management systems, specialty metals, and electrical interconnects ; precision motion control products for data storage, medical devices, business equipment, automation, and other applications ; engineered electrical connectors and packaging products to protect sensitive electronic devices ; floor care and specialty motors ; and metal tubing products. This segment also provides high-purity metals, metal strips, shaped wires, and advanced composites for various industrial applications ; and motors used in commercial appliances, fitness equipment, food and beverage machines, hydraulic pumps, industrial blowers, and vacuum cleaners, as well as operates a network of aviation maintenance, repair, and overhaul facilities. In addition, the company offers clinical and education communication solutions for hospitals, health systems, and educational facilities. AMETEK, Inc. was founded in 1930 and is headquartered in Berwyn, Pennsylvania. »

Compagnie : ARMK
Nom : Aramark
Secteur d'activité : Services
Groupe d'industrie : Specialty Eateries
Capitalisation (Million) : 8781.68524
Description :

« Aramark provides food, facilities, and uniform services to education, healthcare, business and industry, sports, leisure, and corrections clients in North America

and internationally. It offers managed services include dining, catering, food service management, convenience-oriented retail operations, grounds and facilities maintenance, custodial, energy and construction management, and capital project management. The company also provides non-clinical support services, such as patient food and nutrition, and retail food services ; and facilities services comprising clinical equipment maintenance, environmental, laundry and linen distribution, plant operations, strategic/technical, energy and supply chain management, purchasing, and central transportation. In addition, it offers on-site restaurants, catering, convenience stores, and executive dining services ; beverage and vending services ; and facility management services comprising housekeeping, plant operations and maintenance, energy management, laundry and linen, grounds keeping, landscaping, transportation, capital program management and commissioning, and other facility consulting services. Further, the company provides facility and business support services for mining and oil operations ; and concessions, banquet and catering, retail and merchandise sales, recreational and lodging, and facility management services for sports, entertainment, and recreational facilities. Additionally, it offers correctional food, and food and facilities management services for parks ; and operates commissaries, laundry facilities, and property rooms. It also rents, sells, cleans, maintains, and delivers uniform and career apparel, and other textile items ; and provides other garments and work clothes, as well as ancillary items. The company was formerly known as ARAMARK Holdings Corporation and changed its name to Aramark in May 2014. Aramark was founded in 1959 and is based in Philadelphia, Pennsylvania. »

Compagnie : CACI
Nom : CACI International Inc
Secteur d'activité : Technology
Groupe d'industrie : Information Technology Services
Capitalisation (Million) : 3030.20243
Description :

« CACI International Inc, together with its subsidiaries, provides information solutions and services in North America and internationally. The company offers business systems solutions in the areas of financial, human capital, asset and materials, and administrative management ; develops, integrates, and operates command and control solutions ; and develops and integrates solutions that deliver multi-level unified communications from the enterprise directly to and from the tactical edge. It also provides cyber security solutions, as well as supports

cyber operations of intelligence community and Department of Defense. In addition, the company provides enterprise-wide information solutions and services for the design, development, integration, deployment, operations and management, sustainment, and security of its customers' IT solutions ; and supports various initiatives that improves healthcare delivery systems, integrates electronic health records, improves health outcomes for communities, and enhances emergency responsiveness. Further, it provides intelligence services, such as cyber analytics, counterintelligence, and other services that help in the disruption of terrorist activities and counter the proliferation of weapons of mass destruction ; and designs, develops, integrates, deploys, and prototypes hardware-and software-enabled tools and applications, as well as offers instrumentation signals intelligence systems. Additionally, the company provides investigation and litigation support services ; logistics and material readiness solutions, and professional services ; and data and software products, as well as integrates surveillance and reconnaissance technologies into platforms. It primarily serves the U.S. government, as well as other customers comprising state and local governments, commercial enterprises, and agencies of foreign governments. CACI International Inc was founded in 1962 and is headquartered in Arlington, Virginia. »

Compagnie : CMI
Nom : Cummins Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 22903.1586
Description :

« Cummins Inc., together with its subsidiaries, designs, manufactures, distributes, and services diesel and natural gas engines, and engine-related component products worldwide. It operates through Engine, Distribution, Components, and Power Systems segments. The Engine segment manufactures and markets a range of diesel and natural gas powered engines under the Cummins and other customer brand names for the heavy- and medium-duty truck, bus, recreational vehicle, light-duty automotive, construction, mining, marine, rail, oil and gas, defense, and agricultural markets. This segment also offers new parts and services, as well as remanufactured parts and engines. The Distribution segment distributes parts, engines, and power generation products ; and provides service solutions, such as maintenance contracts, engineering services, and integrated products. The Components segment offers emission solutions, including custom engineering systems

and integrated controls, oxidation catalysts, particulate filters, selective catalytic reduction systems, and engineered components ; and turbochargers for light-duty, mid-range, heavy-duty, and high-horsepower diesel markets. This segment also provides air and fuel filters, fuel water separators, lube and hydraulic filters, coolants, fuel additives, and other filtration systems ; and fuel systems for heavy-duty on-highway diesel engine applications, as well as remanufactures fuel systems. The Power Generation segment offers components that back-up and prime power generators, controls, paralleling systems, and transfer switches, as well as A/C generator/alternator products under the Stamford, AVK, and Markon brands. Cummins Inc. sells its products to original equipment manufacturers, dealers/distributors, and end users. The company was formerly known as Cummins Engine Company and changed its name to Cummins Inc. in 2001. Cummins Inc. was founded in 1919 and is headquartered in Columbus, Indiana. »

Compagnie : CNDT
Nom : Conduent Incorporated
Secteur d'activité : Technology
Groupe d'industrie : Computer Based Systems
Capitalisation (Million) : 3022.83893
Description :

« Conduent Incorporated provides business process services with capabilities in transaction-intensive processing, analytics, and automation in the United States, Europe, and internationally. It operates through three segments : Commercial Industries, Government Services, and Transportation. The Commercial Industries segment offers business process services and customized solutions to clients in various industries. The Government Services segment provides government-centric business process services to the United States federal, state, local, and foreign governments for public assistance, program administration, transaction processing, and payment services. The Transportation segment offers systems and support comprising of mission-critical mobility and payment solutions to government clients. The company also provides end-user customer experience, transaction processing, commercial healthcare, human resource, and learning services ; government healthcare, payment solutions, child support and labor workforce, and federal services ; and tolling, transit, photo and parking, and computer-aided dispatch/automatic vehicle location solutions. Conduent Incorporated is headquartered in Florham Park, New Jersey. »

Compagnie : COHR

Nom : Coherent, Inc.
Secteur d'activité : Technology
Groupe d'industrie : Scientific Technical Instruments
Capitalisation (Million) : 3372.80175
Description :

> *« Coherent, Inc. provides lasers and laser-based technology in a range of scientific, commercial, and industrial applications worldwide. It operates through two segments, Specialty Lasers and Systems, and Commercial Lasers and Components. The company designs, manufactures, services, and markets lasers, laser tools, precision optics, and related accessories. Its products are used in markets, such as microelectronics, materials processing, original equipment manufacturer components and instrumentation, and scientific research and government programs. The company markets its products through a direct sales force in the United States, as well as through direct sales personnel and independent representatives internationally. Coherent Inc. was founded in 1966 and is headquartered in Santa Clara, California. »*

Compagnie : CTAS
Nom : Cintas Corporation
Secteur d'activité : Services
Groupe d'industrie : Business Services
Capitalisation (Million) : 12134.92579
Description :

> *« Cintas Corporation provides corporate identity uniforms and related business services primarily in North America, Latin America, Europe, and Asia. Its Rental Uniforms and Ancillary Products segment rents and services uniforms and other garments, including flame resistant clothing, mats, mops and shop towels, and other ancillary items ; and provides restroom cleaning services and supplies, and carpet and tile cleaning services. The company's Uniform Direct Sales segment is involved in the direct sale of uniforms and related items. Its First Aid, Safety, and Fire Protection Services segment offers first aid, safety, and fire protection products and services. The company offers its products and services through its distribution network and local delivery routes, or local representatives to small service and manufacturing companies, as well as corporations. Cintas Corporation was founded in 1968 and is based in Cincinnati, Ohio. »*

Compagnie : DAKT
Nom : Daktronics Inc.

Secteur d'activité : Technology
Groupe d'industrie : Computer Based Systems
Capitalisation (Million) : 471.11532
Description :

« Daktronics, Inc., together with its subsidiaries, designs, manufactures, and sells a range of electronic display systems and related products worldwide. It operates through five segments : Commercial, Live Events, High School Park and Recreation, Transportation, and International. The company offers video display systems, such as displays to show various levels of video, graphics, and animation, as well as controllers ; LED ribbon board displays ; mobile and modular display systems ; freeform LED displays, which include architectural lighting and display products ; indoor and outdoor scoreboards for various sports, digit displays, scoring and timing controllers, statistics software, and other related products ; and timing systems for sports events primarily aquatics and track competitions, as well as swimming touchpads, race start systems, and relay take-off platforms. It also provides message displays ; ITS dynamic message signs, including LED displays for road management, mass transit, and aviation applications ; digit and directional displays for use in parking facilities ; and audio systems for outdoor sports venues. In addition, the company offers static and digital billboards used to display static images, which change at regular intervals for the out-of-home (OOH) advertising industry ; Visiconn system, a software application for controlling content and playback loops for digital billboard applications ; and street furniture comprising advertising light boxes for static, scrolling, and digital OOH campaigns. Further, it provides digit and price displays, such as outdoor time and temperature displays, as well as Fuelight digit displays for the petroleum industry ; and dynamic messaging systems for retailers, convenience stores, and other businesses, as well as maintenance and professional services related to its products. The company sells its products through direct sales and resellers. Daktronics, Inc. was founded in 1968 and is based in Brookings, South Dakota. »

Compagnie : DBD
Nom : Diebold Nixdorf, Incorporated
Secteur d'activité : Technology
Groupe d'industrie : Diversified Computer Systems
Capitalisation (Million) : 1889.77636
Description :

« Diebold Nixdorf, Incorporated provides connected commerce services, software,

and technology for financial, commercial, and industrial customers. The company operates in four segments : North America ; Asia Pacific ; Europe, Middle East and Africa ; and Latin America. It offers financial self-service solutions and technologies, including automated teller machine (ATM) outsourcing, ATM security, deposit automation, recycling and payment terminals, and software. The company also provides financial self-service support and maintenance services comprising installation and ongoing maintenance of products, availability management, branch automation, and distribution channel consulting ; outsourced and managed services, such as remote monitoring, troubleshooting, transaction processing, currency management, maintenance, and online communication services ; and strategic analysis and planning for new systems, systems integration, architectural engineering, consulting, and project management services, as well as multivendor services. In addition, it offers electronic security services and products ; security monitoring solutions comprising remote monitoring and diagnostics, fire detection, intrusion protection, managed access control, energy management, remote video management and storage, logical security, and Web-based solutions ; and physical security and facility products. Further, the company provides development, training, support, and maintenance of elections and lottery equipment, networking, tabulation, and diagnostic software ; and IT solutions and services to retail banks and the retail industry. The company was formerly known as Diebold, Incorporated and changed its name to Diebold Nixdorf, Incorporated in December 2016. Diebold Nixdorf, Incorporated was founded in 1859 and is headquartered in North Canton, Ohio. »

Compagnie : DDD
Nom : 3D Systems Corporation
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals
Capitalisation (Million) : 1493.26032
Description :

« 3D Systems Corporation, through its subsidiaries, provides 3D printing products and services worldwide. The company's 3D printers transform data input generated by 3D design software, CAD software, or other 3D design tools into printed parts using a range of print materials, including plastic, nylon, metal, composite, elastomeric, wax, polymeric dental materials, and Class IV bio-compatible materials. It offers various 3D printing technologies, such as stereolithography, selective laser sintering, direct metal printing, multijet printing, and colorjet printing. The

company also develops, blends, and markets various print materials, such as plastic, nylon, metal, composite, elastomeric, wax, polymeric dental materials, and Class IV bio-compatible materials. It offers its printers under the Accura, DuraForm, LaserForm, CastForm, and VisiJet brand names. In addition, the company provides digital design tools, including software, scanners, and haptic devices, as well as products for product design, mold and die design, 3D scan-to-print, reverse engineering, and production machining and inspection. Further, it offers proprietary software and drivers that provide part preparation, part placement, support placement, build platform management, and print queue management ; and 3D virtual reality simulators and simulator modules for medical applications, as well as digitizing scanners for medical and mechanical applications. Additionally, the company provides warranty, maintenance, and training services ; on-demand solutions ; and software and healthcare services. It primarily serves companies and small and midsize businesses in a range of industries, including healthcare, automotive, aerospace, government, defense, technology, electronics, education, consumer goods, and energy. The company sells its products and services through direct sales force, partner channels, and distributors. 3D Systems Corporation was founded in 1986 and is headquartered in Rock Hill, South Carolina. »

Compagnie : DHR
Nom : Danaher Corporation
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 53842.09287
Description :

« Danaher Corporation designs, manufactures, and markets professional, medical, industrial, and commercial products and services worldwide. The company's Life Sciences segment provides laser scanning, compound, and surgical and other stereo microscopes. This segment also offers filtration, separation, and purification technologies to the biopharmaceutical, food and beverage, medical, aerospace, microelectronics, and general industrial sectors. Its Diagnostics segment provides chemistry, immunoassay, microbiology, and automation systems, as well as hematology and flow cytometry products. The company offers analytical instruments, reagents, consumables, software, and services for hospitals, physicians' offices, reference laboratories, and other critical care settings. Its Dental segment provides consumables, equipment, and services to diagnose, treat, and prevent disease and ailments of the teeth, gums, and supporting bone. The company's products com-

prise implant systems, dental prosthetics, and associated treatment planning software ; orthodontic bracket systems and lab products ; endodontic systems and related consumables ; restorative materials and instruments ; infection prevention products ; digital imaging systems and software ; air and electric powered handpieces, and consumables ; and treatment units. Its Environmental Applied Solutions segment offers instrumentation, services, and disinfection systems to analyze, treat, and manage water in residential, commercial, industrial, and natural resource applications. This segment also provides equipment, consumables, software, and services for various printing, marking, coding, traceability, packaging, design, and color management applications on consumer, pharmaceutical, and industrial products. The company was formerly known as Diversified Mortgage Investors, Inc. and changed its name to Danaher Corporation in 1984. Danaher Corporation was founded in 1969 and is headquartered in Washington, the District of Columbia. »

Compagnie : DOV
Nom : Dover Corporation
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 11640.13827
Description :

« Dover Corporation manufactures and sells a range of equipment and components, specialty systems, software and digital solutions, and support services worldwide. The company operates in four segments : Energy, Engineered Systems, Fluids, and Refrigeration Food Equipment. The Energy segment provides solutions and services for the production and processing of fuels to drilling and production, bearings and compression, and automation end markets. The Engineered Systems segment offers precision marking and coding, digital textile printing, soldering and dispensing equipment, and related consumables and services ; and automation components, including manual clamps, power clamps, rotary and linear mechanical indexers, conveyors, pick and place units, glove ports, and manipulators, as well as end-of-arm robotic grippers, slides, and end effectors for fast-moving consumer goods, digital textile printing, vehicle service, environmental solutions, and industrials end markets. The Fluids segment focuses on the safe handling of critical fluids across the retail fueling, chemical, hygienic, oil and gas, and industrial markets. This segment also manufactures connectors for use in various bio-processing applications ; and pumps and compressors that are used to transfer liquid and bulk products in various markets, including refined fuels, LPG,

food/sanitary, transportation, and chemical process industries. The Refrigeration Food Equipment segment manufactures refrigeration systems, refrigeration display cases, specialty glass, commercial glass refrigerator and freezer doors, and brazed heat exchangers ; and electrical distribution products and engineering services, commercial food service equipment, cook-chill production systems, custom food storage and preparation products, kitchen ventilation systems, conveyer systems, and beverage can-making machinery. The company was founded in 1947 and is headquartered in Downers Grove, Illinois. »

Compagnie : DRI
Nom : Darden Restaurants, Inc.
Secteur d'activité : Services
Groupe d'industrie : Restaurants
Capitalisation (Million) : 8950.43629
Description :

« Darden Restaurants, Inc., through its subsidiaries, owns and operates full-service restaurants in the United States and Canada. As of May 29, 2016, it owned and operated 1,536 restaurants, which included 843 Olive Garden, 481 LongHorn Steakhouse, 54 The Capital Grille, 65 Yard House, 40 Seasons 52, 37 Bahama Breeze, and 16 Eddie V's restaurants. The company was founded in 1968 and is headquartered in Orlando, Florida. »

Compagnie : EPAM
Nom : EPAM Systems, Inc.
Secteur d'activité : Technology
Groupe d'industrie : Information Technology Services
Capitalisation (Million) : 3282.12059
Description :

« EPAM Systems, Inc. provides product development and software engineering solutions worldwide. The company offers software product development services, including product research, customer experience design and prototyping, program management, component design and integration, lifecycle software testing, product deployment and end-user customization, performance tuning, product support and maintenance, and managed services, as well as porting and cross-platform migration. It also provides custom application development services, such as business and technical requirement analysis, user experience design, solution architecture creation and validation, development, quality assurance and testing, legacy applications re-engineering/refactoring, porting, and cross-platform migration and

documentation. In addition, the company offers software application testing services, including test automation tools and frameworks ; testing for enterprise IT, such as test management, automation, functional and non-functional testing, and defect management ; and consulting services. Further, it provides enterprise application platform services comprising requirements analysis and platform selection, customization, cross-platform migration, implementation, integration, and support and maintenance. Additionally, the company offers application maintenance and support services, such as incident management, fault investigation diagnosis, work-around provision, application bug fixes, release management, enhancements, and third-party maintenance ; and infrastructure management services, including application, database, network, server, storage, and systems operations management, as well as incident notification and resolutions. It serves software and technology companies in financial service, travel and consumer, software and hi-tech, media and entertainment, life sciences, and healthcare industries. The company was founded in 1993 and is headquartered in Newtown, Pennsylvania. »

Compagnie : ETN
Nom : Eaton Corporation plc
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 30304.553
Description :

« Eaton Corporation plc operates as a power management company worldwide. Its Electrical Products segment offers electrical and industrial components, residential products, single phase power quality products, emergency lighting and fire detection products, wiring devices, structural support systems, and circuit protection and lighting products. The company's Electrical Systems and Services segment provides power distribution and assemblies, three phase power quality products, hazardous duty electrical equipment, explosion-proof instrumentation, utility power distribution, power reliability equipment, and services. Its Hydraulics segment offers various power products, controls and sensing products, fluid conveyance products, filtration systems solutions, industrial drum and disc brakes, and golf grips. The company's Aerospace segment provides hydraulic power generation systems, controls and sensing products, fluid conveyance products, and fuel systems for commercial and military use. Its Vehicle segment designs, manufactures, markets, and supplies drivetrain, powertrain systems, and critical components, including transmissions, clutches, hybrid power systems, superchargers, engine valves

and valve actuation systems, cylinder heads, locking and limited slip differentials, transmission controls, fuel vapor components, fluid connectors, and conveyance products. The company serves industrial, institutional, governmental, utility, commercial, residential, information technology, renewable energy, marine, agriculture, oil and gas, construction, mining, forestry, material handling, truck and bus, machine tools, molding, primary metals, and power generation markets, as well as original equipment manufacturers and aftermarket customers of heavy, medium, and light-duty trucks, SUVs, CUVs, passenger cars, and agricultural equipment. Eaton Corporation plc was founded in 1916 and is based in Dublin, Ireland. »

Compagnie : FISV
Nom : Fiserv, Inc.
Secteur d'activité : Services
Groupe d'industrie : Business Services
Capitalisation (Million) : 23068.85261
Description :

« Fiserv, Inc., together with its subsidiaries, provides financial services technology worldwide. The company's Payments and Industry Products segment provides debit and credit card processing and services ; electronic bill payment and presentment services ; Internet and mobile banking software and services ; person-to-person payment services ; and other electronic payments software and services. This segment also offers card and print personalization services ; investment account processing services for separately managed accounts ; and fraud and risk management products and services. Its Financial Institution Services segment provides account processing services, item processing and source capture services, loan origination and servicing products, cash management and consulting services, and other products and services that support various types of financial transactions. This segment also offers a range of services, such as customization, business process outsourcing, education, consulting, and implementation services ; and ACH, treasury management, source capture optimization, and enterprise cash and content management solutions, as well as case management and resolution services to the financial services industry. The company also provides document and payment card production and distribution, check processing and imaging, source capture systems, and lending and risk management products and services. Fiserv, Inc. serves banks, thrifts, credit unions, investment management firms, leasing and finance companies, retailers, merchants, mutual savings banks, and building societies. The company was founded in 1984 and is headquartered in Brookfield,

Wisconsin. »

Compagnie : G
Nom : Genpact Limited
Secteur d'activité : Services
Groupe d'industrie : Business Services
Capitalisation (Million) : 4886.41581
Description :

« Genpact Limited provides business process outsourcing and information technology (IT) management services worldwide. The company offers finance and accounting services, including accounts payable comprising document management, invoice processing, approval, resolution management, and TE processing ; order to cash services, such as customer master data management, credit and contract management, fulfillment, billing, collections, and dispute management services ; record to report services consisting of accounting, closing and reporting, treasury, tax, and product cost accounting services ; enterprise performance management, including budgeting, forecasting, business performance reporting, and analytics ; and enterprise risk and compliance services comprising SOX advisory, enterprise risk management, internal audit, FCPA, and IT risk management services. It also provides analytics and research services ; core industry operation services ; business and enterprise risk consulting services ; transformation services ; and supply chain and procurement services, including direct and indirect strategic sourcing, category management, spend analytics, procurement operations, master data management, and other procurement and supply chain advisory services. In addition, the company offers enterprise application services comprising business intelligence and data services, enterprise resource planning, quality assurance, and technology integration ; IT infrastructure management services, including end user computing, infrastructure management, application production support, and database management services ; and collections and customer services in the areas of consumer banking, business-to-business finance, and mortgage servicing. It serves banking and financial services, capital markets, consumer product goods, healthcare, high tech, infrastructure, manufacturing and services, insurance, and life sciences industries. Genpact Limited was founded in 1997 and is based in Hamilton, Bermuda. »

Compagnie : GE
Nom : General Electric Company
Secteur d'activité : Industrial Goods

Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 279545.924
Description :

« General Electric Company operates as an infrastructure and technology company worldwide. Its Power segment offers gas and steam power systems ; maintenance, service, and upgrade solutions ; distributed power gas engines ; water treatment, wastewater treatment, and process system solutions ; and nuclear reactors, fuels, and support services. The company's Renewable Energy segment provides wind turbine platforms, and hardware and software ; onshore and offshore wind turbines ; and solutions, products, and services to hydropower industry. Its Oil Gas segment offers surface and subsea drilling and production systems, and equipment for floating production platforms ; and compressors, turbines, turboexpanders, high pressure reactors, industrial power generation, and auxiliary equipment. The company's Aviation segment provides designs and produces commercial and military aircraft engines, integrated digital components, and electric power and mechanical aircraft systems ; and aftermarket services. Its Healthcare segment offers diagnostic imaging and clinical systems ; products for drug discovery, biopharmaceutical manufacturing, and cellular technologies ; and medical technologies, software, analytics, cloud solutions, and implementation services. The company's Transportation segment provides freight and passenger locomotives, rail, and support advisory services ; and parts, integrated software solutions and data analytics, software-enabled solutions, mining equipment and services, and marine diesel and stationary power diesel engines and motors, as well as overhaul, repair and upgrade, and wreck repair services. Its Energy Connections Lighting segment offers industrial, grid, power conversion, automation and control, lighting, and current solutions. The company's Capital segment provides energy financial services ; and commercial aircraft leasing, financing, and consulting services. General Electric Company was founded in 1892 and is headquartered in Boston, Massachusetts. »

Compagnie : GPC
Nom : Genuine Parts Company
Secteur d'activité : Services
Groupe d'industrie : Specialty Retail, Other
Capitalisation (Million) : 14210.35601
Description :

« Genuine Parts Company distributes automotive replacement parts, industrial replacement parts, office products, and electrical/electronic materials in the United

States, Canada, Australia, New Zealand, Mexico, and Puerto Rico. It distributes automotive replacement parts for imported vehicles, trucks, SUVs, buses, motorcycles, recreational vehicles, farm vehicles, small engines, farm equipment, and heavy duty equipment ; and accessory items used in the automotive aftermarket, including repair shops, service stations, fleet operators, automobile and truck dealers, leasing companies, bus and truck lines, mass merchandisers, farms, industrial concerns, and individuals through 57 NAPA automotive parts distribution centers and 1,100 NAPA AUTO PARTS stores. The company also distributes industrial replacement parts and related supplies, such as bearings, mechanical and electrical power transmission products, industrial automation products, hoses, hydraulic and pneumatic components, industrial supplies, and material handling products primarily for food and beverage, forest products, primary metal, pulp and paper, mining, automotive, oil and gas, petrochemical, and pharmaceutical industries through 483 branches, 13 distribution centers, and 43 service centers. In addition, it distributes office furniture, technology products, general office and school supplies, cleaning, janitorial and breakroom supplies, safety and security items, healthcare products, and disposable food service products to resellers through 56 distribution centers. Further, the company distributes wires and cables, connectivity solutions, insulating and conductive materials, assembly tools, test equipment, custom fabricated parts, and specialty coated materials to original equipment manufacturers, motor repair shops, specialty wire and cable users, and various industrial assembly markets. Genuine Parts Company was founded in 1928 and is based in Atlanta, Georgia. »

Compagnie : HON
Nom : Honeywell International Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 88292.14754
Description :

« Honeywell International Inc. operates as a diversified technology and manufacturing company worldwide. It operates through four segments : Aerospace ; Home and Building Technologies ; Performance Materials and Technologies ; and Safety and Productivity Solutions. The Aerospace segment supplies products, software, and services for aircraft and vehicles that it sells to original equipment manufacturers and other customers in various markets, including air transport, regional, business and general aviation aircraft, airlines, aircraft operators, defense and

space contractors, and automotive and truck manufacturers. The Home and Building Technologies segment provides products, software, solutions, and technologies that help homes owners, commercial building owners, and occupants, as well as electricity, gas, and water providers. The Performance Materials and Technologies segment develops and manufactures advanced materials, process technologies, and automation solutions. The Safety and Productivity Solutions segment provides products, software, and connected solutions to customers that enhance productivity, workplace safety, and asset performance. The company was founded in 1920 and is based in Morris Plains, New Jersey. »

Compagnie : HPQ
Nom : HP Inc.
Secteur d'activité : Technology
Groupe d'industrie : Diversified Computer Systems
Capitalisation (Million) : 25308.89346
Description :

« HP Inc. provides products, technologies, software, solutions, and services to individual consumers, small- and medium-sized businesses, and large enterprises, including customers in the government, health, and education sectors worldwide. It operates through Personal Systems and Printing segments. The Personal Systems segment offers commercial personal computers (PCs), consumer PCs, workstations, thin clients, commercial tablets and mobility devices, retail point-of-sale systems, displays and other related accessories, software, support, and services for the commercial and consumer markets. The Printing segment provides consumer and commercial printer hardware, supplies, media, solutions, and services, as well as scanning devices ; and laserJet and enterprise, inkjet and printing, graphics, and 3D printing solutions. The company was formerly known as Hewlett-Packard Company and changed its name to HP Inc. in October 2015. HP Inc. was founded in 1939 and is headquartered in Palo Alto, California. »

Compagnie : IAC
Nom : IAC/InterActiveCorp
Secteur d'activité : Services
Groupe d'industrie : Specialty Retail, Other
Capitalisation (Million) : 5137.16664
Description :

« IAC/InterActiveCorp operates as a media and Internet company in the United States and internationally. It operates through four segments : The Match

Group, Search Applications, Media, and eCommerce. The Match Group segment provides subscription-based and ad-supported online personals services through its Websites and applications. This segment also operates The Princeton Review that offers college and graduate school admissions test preparation and college readiness services ; Tutor.com, which offers various live, one-on-one, and on-demand tutoring services ; and Daily Burn, a health and fitness property that provides streaming fitness and workout videos in various platforms. The Search Applications segment operates various Websites to offer search services, and content and other services comprising Ask.com, About.com, CityGrid, Dictionary.com, Investopedia, PriceRunner, and Ask.fm ; and develops, markets, and distributes various downloadable applications, which provide users the ability to access search services. The Media segment offers Vimeo, a video sharing platform that provides video creators tools to share, distribute, and monetize content online, as well as offers viewers a clutter-free environment to watch content in various Internet-enabled devices ; and The Daily Beast, a Website dedicated to news, commentary, culture, and entertainment. This segment also provides Electus, an integrated multimedia entertainment studio to produce video content for distribution, as well as operates Electus Digital, which consists of various Websites and properties. The eCommerce segment offers HomeAdvisor, an online marketplace for matching consumers with home services professionals ; and operates Shoebuy, an Internet retailer of footwear and related apparel and accessories. The company, formerly known as InterActiveCorp, was founded in 1986 and is headquartered in New York, New York. »

Compagnie : INVE
Nom : Identiv, Inc.
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals
Capitalisation (Million) : 35.25396
Description :

« Identiv, Inc. operates as a security technology company that secures and manages access to physical places, things, and information worldwide. It operates through four segments : Premises (PACS), Identity, Credentials, and All Other. The PACS segment offers modular Hirsch MX controllers that allow customers to start with a small system and expand over time ; Hirsch Velocity software platform for centralized management of access and security operations across an organization ; Federal Identity, Credential and Access Management architecture, an access

control system ; and TouchSecure door readers that provide various features to support security standards. The Identity segment provides smart card readers, which include various contact, contactless, portable, and mobile smart card readers, as well as tokens and terminals to enable logical access, and security and identification applications, such as national ID, payment, e-health, and e-government. It also offers access cards and other devices related to its reader products. The Credentials segment provides NFC and radio frequency identification products, including inlays and inlay-based, and other cards ; and labels, tags, and stickers, as well as other radio frequency and integrated circuits components for use in various applications, such as virtual reality, games, loyalty cards, mobile payment systems, transit and event ticketing, and brand authenticity from pharmaceuticals to consumer goods, hospital resource management, cold-chain management, and others. The All Other segment offers chip drives and digital media readers. Identiv, Inc. markets and sells its products through original equipment manufacturers, dealers, systems integrators, value added resellers, resellers, and Internet, as well as directly to end users. The company was formerly known as Identive Group, Inc. and changed its name to Identiv, Inc. in May 2014. Identiv, Inc. was founded in 1990 and is headquartered in Fremont, California. »

Compagnie : IR
Nom : Ingersoll-Rand Plc
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 19385.14668
Description :

« Ingersoll-Rand plc designs, manufactures, sells, and services industrial and commercial products. It operates through Climate and Industrial segments. The Climate segment offers building management, bus and rail HVAC, control, container and cryogenic refrigeration, diesel-powered refrigeration, ductless, geothermal, package heating and cooling, rail and self-powered truck refrigeration, temporary heating and cooling, trailer refrigeration, unitary, and vehicle-powered truck refrigeration systems. This segment also provides aftermarket and OEM parts and supplies, air conditioners, air exchangers and handlers, airside and terminal devices, auxiliary power units, chillers, coils and condensers, gensets, furnaces, heat pumps, home automation, humidifiers, hybrid and non-diesel transport refrigeration solutions, indoor air quality, industrial refrigeration, motor replacements, performance contracting, refrigerant reclamation, thermostats/controls, transport

heater products, and water source heat pumps. In addition, this segment offers energy and facility management, installation contracting, rental, and repair and maintenance services ; and service agreements. The Industrial segment provides air treatment and separation, engine starting, ergonomic material handling, fluid handling, precision fastening, and mobile golf information systems ; and compressors, airends, blowers, dryers, filters, golf vehicles, hoists, fluid power components, power tools, pumps, rough terrain vehicles, utility and low-speed vehicles, and winches, as well as aftermarket controls, parts, accessories, and consumables. The company markets and sells its products under the American Standard, ARO, Club Car, Nexia, Thermo King, and Trane brand names through sales offices, distributors, and dealers in the United States ; and through subsidiary sales and service companies with a supporting chain of distributors worldwide. Ingersoll-Rand plc was founded in 1872 and is headquartered in Swords, Ireland. »

Compagnie : ITW
Nom : Illinois Tool Works Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 42984.04095
Description :

« Illinois Tool Works Inc. manufactures and sells industrial products and equipment worldwide. It operates through seven segments : Automotive OEM ; Test Measurement and Electronics ; Food Equipment ; Polymers Fluids ; Welding ; Construction Products ; and Specialty Products. The Automotive OEM segment produces plastic and metal components, fasteners, and assemblies for automotive-related applications. The Test Measurement and Electronics segment provides equipment, consumables, and related software for testing and measuring of materials and structures. This segment also offers equipment and consumables used in the production of electronic subassemblies and microelectronics. The Food Equipment segment provides commercial food processing, warewashing, cooking, and refrigeration equipment ; and kitchen exhaust, ventilation, and pollution control systems, as well as related services. The Polymers Fluids segment produces adhesives, sealants, lubrication and cutting fluids, and fluids and polymers for auto aftermarket maintenance and appearance. The Welding segment produces arc welding equipment, consumables, and accessories ; and metal jacketing and other insulation products for various industrial and commercial applications. The Construction Products segment produces engineered fastening systems and solu-

tions. The Specialty Products segment provides beverage packaging equipment and consumables, product coding and marking equipment and consumables, and appliance components and fasteners. The company distributes its products directly to industrial manufacturers, as well as through independent distributors. Illinois Tool Works Inc. was founded in 1912 and is headquartered in Glenview, Illinois. »

Compagnie : KAR
Nom : KAR Auction Services, Inc.
Secteur d'activité : Services
Groupe d'industrie : Specialty Retail, Other
Capitalisation (Million) : 5888.47437
Description :

« KAR Auction Services, Inc., together with its subsidiaries, provides vehicle auction services in the United States, Canada, Mexico, and the United Kingdom. It operates in three segments : ADESA Auctions, IAA, and AFC. The ADESA Auctions segment offers whole car auctions and related services to the vehicle remarketing industry through online auctions and auction facilities. It also provides value-added services, such as auction related, transportation, reconditioning, inspection, title and repossession administration and remarketing, vehicle research, and analytical services. This segment sells its products and services through vehicle manufacturers, fleet companies, rental car companies, finance companies, and others. The IAA segment offers various loss solutions and salvage vehicle auction services that facilitate the remarketing of vehicles for a range of sellers, including insurance companies, dealerships, rental car companies, fleet lease companies, and charitable organizations. This segment also provides catastrophe, vehicle inspection center, and transportation and towing services. The AFC segment offers floorplan financing, a short-term inventory-secured financing, to independent used vehicle dealers. As of December 31, 2016, the company had a network of 77 whole car auction locations and 172 salvage auction sites. The company was formerly known as KAR Holdings, Inc. and changed its name to KAR Auction Services, Inc. in November 2009. KAR Auction Services, Inc. was founded in 2006 and is headquartered in Carmel, Indiana. »

Compagnie : KODK
Nom : Eastman Kodak Company
Secteur d'activité : Consumer Goods
Groupe d'industrie : Electronic Equipment

Capitalisation (Million) : 656.58
Description :

« Eastman Kodak Company provides hardware, software, consumables, and services to customers in various markets worldwide. The company operates through seven segments : Print Systems ; Micro 3D Printing and Packaging ; Software and Solutions ; Consumer and Film ; Enterprise Inkjet Systems ; Intellectual Property Solutions ; and Eastman Business Park. It offers digital offset plate and computer-to-plate imaging solutions, and electro photographic printing solutions to a range of commercial industries, including commercial print, direct mail, book publishing, newspapers and magazines, and packaging. The company also provides flexographic printing equipment and plates, and related consumables and services, as well as printed functional materials and components ; suite of software solutions for print production workflow, as well as print and managed media services ; motion picture and industrial films, chemicals, and inks ; and publishing, transactional, commercial print, and direct mail systems, as well as licenses Kodak brands to third parties, and consumer products. In addition, it offers intellectual property solutions ; and leases technology center and industrial complex. The company sells its products and services through third party resellers and distributors, as well as directly and indirectly to enterprise accounts and customers. Eastman Kodak Company was founded in 1880 and is headquartered in Rochester, New York. »

Compagnie : KTCC
Nom : Key Tronic Corporation
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals
Capitalisation (Million) : 84.12942
Description :

« Key Tronic Corporation, doing business as KeyTronicEMS Co., provides electronic manufacturing services (EMS) to original equipment manufacturers in the United States and internationally. The company offers integrated electronic and mechanical engineering, assembly, sourcing and procurement, logistics, and new product testing services. Its services include product design ; surface mount technologies and pin through hole capability for printed circuit board assembly ; tool making ; precision plastic molding ; sheet metal fabrication ; liquid injection molding ; complex assembly ; automated tape winding ; prototype design ; and full product assembly services. The company also manufactures keyboards and other input devices. It markets its products and services primarily through field sales people

and distributors. The company was founded in 1969 and is headquartered in Spokane Valley, Washington. »

Compagnie : LOGI
Nom : Logitech International S.A.
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals
Capitalisation (Million) : 4114.34331
Description :

« Logitech International S.A. engages in design, manufacture, and marketing of personal peripherals for personal computers and other digital platforms in Europe, the Middle East, Africa, the Americas, and the Asia Pacific. It offers mice, trackballs, keyboards and desktops, gaming controllers, multimedia speakers, headsets, Webcams, 3D control devices, speakers, headphones, earphones, and custom in-ear monitors. The company also provides wireless music solutions for home, remote controls for home entertainment systems, and PC-based video security systems for home or small business. It offers its products for PC navigation, Internet communications, digital music, home-entertainment control, video security, interactive gaming, and wireless devices. The company sells its products to a network of distributors and resellers, including wholesale distributors, consumer electronics retailers, mass merchandisers, specialty electronics stores, computer and telecommunications stores, value-added resellers, and online merchants ; and original equipment manufacturers. Logitech International S.A. was founded in 1981 and is based in Apples, Switzerland. »

Compagnie : MATW
Nom : Matthews International Corporation
Secteur d'activité : Services
Groupe d'industrie : Personal Services
Capitalisation (Million) : 2469.959
Description :

« Matthews International Corporation provides brand solutions, memorialization products, and industrial products in the United States, Central and South America, Canada, Europe, Australia, and Asia. It operates through three segments : SGK Brand Solutions, Memorialization, and Industrial Technologies. The SGK Brand Solutions segment provides brand development, deployment, delivery, brand management, pre-media graphics services, printing plates, gravure cylinders, steel bases, embossing tools, special purpose machinery, engineering assistance, print process

assistance, print production management, digital asset management, content management, and package design services to brand owners and packaging industry converters ; and creative digital graphics services, as well as designs, engineers, manufactures, and executes merchandising and display systems. The Memorialization segment manufactures and markets a range of memorialization products used primarily in cemeteries, funeral homes, and crematories. Its products include cast bronze memorials, flush bronze and granite memorials, upright granite memorials and monuments, cremation memorialization products, granite benches, flower vases, crypt plates and letters, cremation urns, niche units, cemetery features and statues, caskets, community and family mausoleums, and other memorialization products, as well as architectural products used to identify or commemorate people, places, events, and accomplishments. The Industrial Technologies segment designs, manufactures, and distributes marking and coding equipment and consumables, industrial automation products, and order fulfillment systems for identifying, tracking, picking, and conveying consumer and industrial products. It serves manufacturers, suppliers, and distributors of durable goods and building products ; consumer goods manufacturers ; and producers of pharmaceuticals. Matthews International Corporation was founded in 1850 and is based in Pittsburgh, Pennsylvania. »

Compagnie : MMM
Nom : 3M Company
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 107403.8552
Description :

« 3M Company operates as a diversified technology company worldwide. The company's Industrial segment offers tapes ; coated, non-woven, and bonded abrasives ; adhesives ; advanced ceramics ; sealants ; specialty materials ; separation and purification products ; closure systems for personal hygiene products ; acoustic systems products ; automotive components ; and abrasion-resistant films, and paint finishing and detailing products. Its Safety and Graphics Business segment provides personal protection products, traffic safety and security products, commercial graphics systems, commercial cleaning and protection products, floor matting, roofing granules for asphalt shingles, and fall protection products. The company's Health Care segment offers medical and surgical supplies, skin health and infection prevention products, inhalation and transdermal drug delivery systems, dental

and orthodontic products, health information systems, and food safety products. Its Electronics and Energy segment provides optical films ; packaging and interconnection devices ; insulating and splicing solutions ; touch screens and touch monitors ; renewable energy component solutions ; and infrastructure protection products. The company's Consumer segment offers sponges, scouring pads, high-performance cloths, repositionable notes, indexing systems, home improvement and care products, protective materials, and consumer and office tapes and adhesives. The company serves automotive, electronics and energy, appliance, paper and printing, packaging, food and beverage, construction, medical clinics and hospitals, pharmaceuticals, dental and orthodontic practitioners, health information systems, food manufacturing and testing, consumer and office retail, office business to business, home improvement, drug and pharmacy retail, and other markets directly, as well as through wholesalers, retailers, jobbers, distributors, and dealers. The company was founded in 1902 and is headquartered in St. Paul, Minnesota. »

Compagnie : MRCY
Nom : Mercury Systems, Inc.
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals
Capitalisation (Million) : 1239.47131
Description :

« Mercury Systems, Inc. provides secure processing subsystems for various critical defense and intelligence programs in the United States. The company's products and solutions are deployed in approximately 300 programs with 25 defense prime contractors. Its principal programs include Aegis, Patriot, Surface Electronic Warfare Improvement Program, Gorgon Stare, Predator, F-35, and Reaper. The company also designs, markets, and licenses software and middleware environments under the MultiCore Plus name to accelerate development and execution of signal and image processing applications on a range of heterogeneous and multi-computing platforms. In addition, it offers hardware products, such as signal and image processing, multi-computer, and sensor interfaces, including embedded processing boards, switch fabric boards, high speed input/output boards, digital receiver boards, high-density memory modules, secure solid-state drives, secure GPS receiver modules, and chassis-based systems ; radio frequency (RF) and microwave assemblies, such as tuners, converters, transceivers, and switch filters ; and RF and microwave components, which include power amplifiers and limiters,

switches, oscillators, and equalizers. The company was formerly known as Mercury Computer Systems, Inc. and changed its name to Mercury Systems, Inc. in November 2012. Mercury Systems, Inc. was founded in 1981 and is headquartered in Chelmsford, Massachusetts. »

Compagnie : PBPB
Nom : Potbelly Corporation
Secteur d'activité : Services
Groupe d'industrie : Specialty Eateries
Capitalisation (Million) : 324.94512
Description :

« Potbelly Corporation, through its subsidiaries, owns and operates Potbelly Sandwich Works sandwich shops in the United States. It also sells and administers franchises of Potbelly Sandwich Works sandwich shops. As of December 25, 2016, the company operated 441 shops in 29 states and the District of Columbia, including 411 company operated shops and 30 franchisees operated shops ; and 13 international franchised shops, including 11 shops in the Middle East, 1 shop in the United Kingdom, and 1 shop in Canada. The company was formerly known as Potbelly Sandwich Works, Inc. and changed its name to Potbelly Corporation in 2002. Potbelly Corporation was founded in 1977 and is headquartered in Chicago, Illinois. »

Compagnie : PRLB
Nom : Proto Labs, Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Machine Tools Accessories
Capitalisation (Million) : 1357.285
Description :

« Proto Labs, Inc., together with its subsidiaries, operates as an e-commerce enabled digital manufacturer of custom parts for prototyping and short-run production in the Unites States, Europe, and Japan. The company utilizes injection molding, computer numerical control machining, and three-dimensional (3D) printing to manufacture custom parts for developers and engineers who use 3D computer-aided design software to design products in the medical, aerospace, computer/electronics, consumer products, industrial machinery, and other markets. Proto Labs, Inc. was founded in 1999 and is headquartered in Maple Plain, Minnesota. »

Compagnie : PZZA

Nom : Papa John's International, Inc.
Secteur d'activité : Services
Groupe d'industrie : Specialty Eateries
Capitalisation (Million) : 3157.00127
Description :

> *« Papa John's International, Inc. operates and franchises pizza delivery and carryout restaurants under the Papa John's trademark in the United States and internationally. It operates through five segments : Domestic Company-Owned Restaurants, North America Commissaries, North America Franchising, International Operations, and All Others. The company also operates dine-in and delivery restaurants. As of December 25, 2016, it operated 5,097 Papa John's restaurants, including 744 company-owned and 4,353 franchised restaurants. The company was founded in 1984 and is headquartered in Louisville, Kentucky. »*

Compagnie : ROK
Nom : Rockwell Automation Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 17299.968
Description :

> *« Rockwell Automation Inc. provides industrial automation and information solutions worldwide. It operates in two segments, Architecture Software ; and Control Products Solutions. The Architecture Software segment provides control platforms, including controllers, electronic operator interface devices, electronic input/output devices, communication and networking products, and industrial computers that perform multiple control disciplines and monitoring of applications, such as discrete, batch and continuous process, drives control, motion control, and machine safety control. This segment also offers software products that include configuration and visualization software, which is used to operate and supervise control platforms, process control software, and manufacturing execution systems and information solution software to enhance manufacturing productivity and meet regulatory requirements ; and other products comprising sensors, machine safety components, and linear motion control products. The Control Products Solutions segment offers low and medium voltage electro-mechanical and electronic motor starters, motor and circuit protection devices, AC/DC variable frequency drives, push buttons, signaling devices, termination and protection devices, relays, and timers ; and various packaged solutions, such as configured drives and motor*

control centers to automation and information solutions. This segment also offers total life-cycle support services, including technical support and repair, asset management, training, maintenance, and safety and network consulting services. The company serves food and beverage, home and personal care, life sciences, automotive and tire, oil and gas, and mining and metal industries through independent distributors and direct sales force in the United States, Canada, Europe, the Middle East, Africa, the Asia Pacific, and Latin America. Rockwell Automation Inc. was founded in 1903 and is headquartered in Milwaukee, Wisconsin. »

Compagnie : ROL
Nom : Rollins, Inc.
Secteur d'activité : Services
Groupe d'industrie : Business Services
Capitalisation (Million) : 7358.31422
Description :

« Rollins, Inc., through its subsidiaries, provides pest and termite control services to residential and commercial customers. Its pest control services include protection against termite damage, rodents, and insects to homes and businesses, including hotels, food service establishments, food manufacturers, retailers, and transportation companies. The company also provides pest management and sanitation services and products to the food and commodity industries ; consulting services on border protection related to Australia's biosecurity program ; and bird control and specialist services, as well as offers specialized services to mining, and oil and gas sectors. It serves clients directly, as well as through franchises operations in North America, Australia, Europe, Central America, the Caribbean, the Middle East, Asia, the Mediterranean, Africa, Canada, Australia, and Mexico. Rollins, Inc. was founded in 1948 and is headquartered in Atlanta, Georgia. »

Compagnie : ROP
Nom : Roper Technologies, Inc.
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 18570.57352
Description :

« Roper Technologies, Inc. designs and develops software, and engineered products and solutions. It operates in four segments : Medical Scientific Imaging ; RF Technology ; Industrial Technology ; and Energy Systems Controls. The com-

pany offers diagnostic and laboratory software solutions ; patient positioning devices and related software, 3-D measurement technology, and diagnostic and therapeutic disposable products ; non-invasive instruments and video laryngoscopes ; and a cloud-based financial analytics and performance software platform, as well as electron filters, charged couple devices, and complementary metal oxide semiconductor cameras, detectors, and related software. It also offers radio frequency identification communication technology and software solutions that are used primarily in toll and traffic systems, security and access controls, campus card systems, card readers, software-as-a-service, and metering and remote monitoring applications, as well as management software for legal and construction firms. In addition, the company offers fluid handling pumps, materials analysis equipment and consumables, leak testing equipment, flow measurement and metering equipment, and water meter and automatic meter reading products and systems. Further, it provides control systems, fluid properties testing equipment, industrial valves and controls, vibration sensors and controls, and non-destructive inspection and measurement products and solutions. Additionally, the company provides enterprise software and information solutions for government contractors, professional services firms, and other project-based businesses. It serves healthcare, transportation, commercial construction, food, energy, water, education, and academic research markets in the United States and internationally. The company was formerly known as Roper Industries, Inc. and changed its name to Roper Technologies, Inc. in April 2015. The company was founded in 1981 and is based in Sarasota, Florida. »

Compagnie : SBUX
Nom : Starbucks Corporation
Secteur d'activité : Services
Groupe d'industrie : Specialty Eateries
Capitalisation (Million) : 80803.808
Description :

« Starbucks Corporation, together with its subsidiaries, operates as a roaster, marketer, and retailer of specialty coffee worldwide. The company operates in four segments : Americas ; China/Asia Pacific ; Europe, Middle East, and Africa ; and Channel Development. Its stores offer coffee and tea beverages, packaged roasted whole bean and ground coffees, single-serve and ready-to-drink coffee and tea products, juices, and bottled water ; an assortment of fresh food and snack offerings ; and various food products, such as pastries, breakfast sandwiches, and lunch items,

as well as beverage-making equipment and accessories. The company also licenses its trademarks through licensed stores, and grocery and national foodservice accounts. It offers its products under the Starbucks, Teavana, Tazo, Seattle's Best Coffee, Evolution Fresh, La Boulange, Ethos, Frappuccino, Starbucks Doubleshot, Starbucks Refreshers, and Starbucks VIA brand names. As of November 3, 2016, the company operated 25,085 stores. Starbucks Corporation was founded in 1971 and is based in Seattle, Washington. »

Compagnie : SERV
Nom : ServiceMaster Global Holdings, Inc.
Secteur d'activité : Services
Groupe d'industrie : Business Services
Capitalisation (Million) : 5076.33024
Description :

« ServiceMaster Global Holdings, Inc. provides residential and commercial services in the United States. It operates in three segments : Terminix, American Home Shield, and the Franchise Services Group. The Terminix segment offers termite and pest control services, including termite remediation, annual termite inspection and prevention treatments with damage claim guarantees, periodic pest control services, insulation services, mosquito control, crawlspace encapsulation, and wildlife exclusion. The American Home Shield segment provides home warranty plans that cover the repair or replacement of household systems and appliances, such as electrical, plumbing, central heating and air conditioning systems, water heaters, refrigerators, dishwashers, and ovens/cook tops. The Franchise Services Group segment offers disaster restoration, janitorial, residential cleaning, furniture repair, and home inspection services through franchise under the ServiceMaster Restore, ServiceMaster Clean, Merry Maids, Furniture Medic, and AmeriSpec brands. The company markets its services to homeowners and businesses through the Internet, direct mail, television and radio advertising, print advertisements, marketing partnerships, franchise network, branch operations, telemarketing, and various social media channels, as well as through various participants in the residential real estate marketplace, such as real estate brokerages, financial institutions, and insurance carriers. ServiceMaster Global Holdings, Inc. was founded in 1929 and is headquartered in Memphis, Tennessee. »

Compagnie : SHAK
Nom : Shake Shack Inc.
Secteur d'activité : Services

Groupe d'industrie : Specialty Eateries
Capitalisation (Million) : 891.95487
Description :

« Shake Shack Inc. owns, operates, and licenses Shake Shack restaurants (Shacks) in the United States and internationally. Shacks offers hamburgers, hot dogs, crispy chicken, crinkle-cut fries, shakes, frozen custard, beer, shakes, wine, and other products. As of December 28, 2016, it had 114 Shacks, including 64 domestic company-operated Shacks, 7 domestic licensed Shacks, and 43 international licensed Shacks. The company was founded in 2004 and is headquartered in New York, New York. »

Compagnie : SXI
Nom : Standex International Corporation
Secteur d'activité : Industrial Goods
Groupe d'industrie : Industrial Equipment Components
Capitalisation (Million) : 1119.49566
Description :

« Standex International Corporation manufactures and sells various products and services for commercial and industrial market segments in the United States and internationally. The company's Food Service Equipment segment offers refrigerated cabinets, cases, display units, coolers and freezers, ovens, griddles, char broilers, commercial ranges, toasters, warmers, roller grills, countertop merchandisers, cook and hold units, rotisseries, pressure fryers, deep fryers, baking equipment, pump systems, and display cases. Its Engraving segment provides mold texturizing, slush molding, and in-mold graining tools ; and roll engraving, hygiene product tooling, low observation vents for stealth aircraft, and process machineries, as well as project management and design services. It serves automotive, plastic, building product, synthetic material, converting, textile and paper, computer, houseware, hygiene product tooling, and aerospace industries. The company's Engineering Technologies segment offers customized solutions used in the manufacture of engineered components for the aviation, aerospace, defense, energy, industrial, medical, marine, oil and gas, and space markets. Its Electronics segment offers electronic components, including reed relays, fluid level sensors, and electronic assemblies ; and wound transformers and inductors, assemblies, and mechanical packaging and planar transformers for the transportation, smart-grid, energy, appliance, HVAC, security, military, medical, aerospace, test and measurement, power distribution, and general industrial applications. The company's Hydraulics

segment offers telescopic and piston rod hydraulic cylinders, and pneumatic cylinders for use in construction equipment, refuse, airline support, mining, oil and gas, and other material handling applications. Standex International Corporation sells its products through dealers, and industry representatives. The company was founded in 1955 and is headquartered in Salem, New Hampshire. »

Compagnie : SYKE
Nom : Sykes Enterprises, Incorporated
Secteur d'activité : Technology
Groupe d'industrie : Information Technology Services
Capitalisation (Million) : 1224.50094
Description :

« Sykes Enterprises, Incorporated, together with its subsidiaries, provides business process outsourcing solutions. Its customer care services include product information requests, describing product features, activating customer accounts, resolving complaints, cross-selling/up-selling, handling billing inquiries, changing addresses, claims handling, ordering/reservations, prequalification and warranty management, providing health information, and roadside assistance. The company's technical support services comprise handling inquiries regarding hardware, software, communications services, communications equipment, Internet access technology, and Internet portal usage ; and customer acquisition services focuses around digital marketing, demand generation, and in-bound sales conversion, as well as inbound and outbound up-selling its clients' products and services. It also provides technical staffing services and outsourced corporate help desk services ; and fulfillment services, such as order and payment processing, inventory control, product delivery, and product returns handling. The company offers its services through phone, email, social media, text messaging, chat, and digital self-service support. Sykes Enterprises, Incorporated provides its services to corporations, medium-sized businesses, and public institutions in the communications, financial services, technology/consumer, transportation and leisure, healthcare, and other industries. It operates in the United States, Canada, Latin America, Australia, the Asia Pacific Rim, Europe, and Africa. Sykes Enterprises, Incorporated was founded in 1977 and is headquartered in Tampa, Florida. »

Compagnie : TACT
Nom : TransAct Technologies Incorporated
Secteur d'activité : Technology
Groupe d'industrie : Computer Peripherals

Capitalisation (Million) : 48.7302
Description :

« TransAct Technologies Incorporated designs, develops, assembles, and markets transaction-based and specialty printers and terminals in the United States and internationally. It offers thermal, inkjet, and impact printers and terminals to generate food rotation date and nutritional labels, promotional coupons, and transaction records, such as receipts, tickets, register journals, and other documents, as well as for printed logging and plotting of oil field and drilling data. The company also provides consumable products, including inkjet cartridges, ribbons, receipt papers, color thermal papers, and other printing supplies, as well as replacement parts ; maintenance, repair, and testing services ; and refurbished printers. In addition, it offers EPICENTRALTM print system, a software solution that enables casino operators to create promotional coupons and marketing messages, and print them at the slot machine ; and technical support services. The company markets its products under the AccuDate, Epic, EPICENTRAL, Ithaca, Responder, and Printrex brand names for restaurant solutions, POS automation and banking, casino and gaming, lottery, mobile, and oil and gas. It sells its products to original equipment manufacturers, value-added resellers, and distributors, as well as directly and online to end-users. TransAct Technologies Incorporated was founded in 1996 and is headquartered in Hamden, Connecticut. »

Compagnie : TSCO
Nom : Tractor Supply Company
Secteur d'activité : Services
Groupe d'industrie : Specialty Retail, Other
Capitalisation (Million) : 9950.03877
Description :

« Tractor Supply Company operates rural lifestyle retail stores in the United States. The company offers a selection of merchandise, including equine, livestock, pet, and small animal products necessary for their health, care, growth, and containment ; hardware, truck, towing, and tool products ; seasonal products, such as heating products, lawn and garden items, power equipment, gifts, and toys ; work/recreational clothing and footwear ; and maintenance products for agricultural and rural use. As of January 26, 2017, it operated 1,600 retail stores in 49 states. The company operates its retail stores under the Tractor Supply Company, Del's Feed Farm Supply, and Petsense names. It also operates an e-commerce Website, TractorSupply.com. The company sells its products to recreational far-

mers, ranchers, and others, as well as tradesmen and small businesses. Tractor Supply Company was founded in 1938 and is headquartered in Brentwood, Tennessee. »

Compagnie : ULTA
Nom : Ulta Beauty, Inc.
Secteur d'activité : Services
Groupe d'industrie : Specialty Retail, Other
Capitalisation (Million) : 15865.62697
Description :

« Ulta Beauty, Inc. operates as a beauty retailer in the United States. The company's stores provide cosmetics, fragrance, skincare, haircare, bath and body products, and salon styling tools, as well as others, including nail products and accessories. It offers private label products consisting of Ulta Beauty Collection branded cosmetics, skincare, and bath products. As of March 9, 2017, the company operated 974 retail stores in 48 states and the District of Columbia. Its full-service salon offers hair, skin, and brow services ; and provides products through its Website, ulta.com. The company was formerly known as Ulta Salon, Cosmetics Fragrance, Inc. and changed its name to Ulta Beauty, Inc. in January 2017. Ulta Beauty, Inc. was founded in 1990 and is based in Bolingbrook, Illinois. »

Compagnie : VRTU
Nom : Virtusa Corporation
Secteur d'activité : Technology
Groupe d'industrie : Information Technology Services
Capitalisation (Million) : 752.65338
Description :

« Virtusa Corporation operates as an information technology (IT) services company. It offers business and IT consulting services comprising application inventory and portfolio assessment, business/technology alignment analysis, business process optimization, and quality assurance process consulting ; accelerated solution design, enterprise architecture analysis, technology roadmaps, product evaluation and selection, and business process analysis and design ; and program governance and change management, program management planning, and complex program management. The company also provides technology implementation services, such as application development, package implementation and integration, software product engineering, application maintenance and support, business process management, CRM and SAP implementation, customer experience and

content management, enterprise mobility, cloud computing, and social media solutions ; systems consolidation and rationalization, technology migration and porting, and legacy application Web-enablement ; data management, business intelligence, reporting and decision support, master data management, data integration, and big data analytics ; and software quality assurance and managed testing services. In addition, it offers application outsourcing services, such as the production support, application maintenance and enhancement, and ongoing software engineering ; systems and database administration, and monitoring ; outsourcing of quality assurance planning ; and preparation and execution of test cases, scripts, and data. Virtusa Corporation provides its services to communications and technology ; banking, financial services, and insurance ; and media and information industries worldwide. The company was formerly known as eRunway, Inc. and changed its name to Virtusa Corporation in April 2002. Virtusa Corporation was founded in 1996 and is headquartered in Westborough, Massachusetts. »

Compagnie : XONE
Nom : The ExOne Company
Secteur d'activité : Industrial Goods
Groupe d'industrie : Diversified Machinery
Capitalisation (Million) : 150.47941
Description :

« The ExOne Company develops, manufactures, and markets three dimensional (3D) printing machines, 3D printed and other products, materials, and services primarily in North America, Europe, and Asia. The company provides various machines that enable designers and engineers to design and produce industrial prototypes and production parts. Its machines include Exerial, S-Max/S-Max+, and S-Print, which are indirect printing machines ; M-Print, M-Flex, and Innovent, that are direct printing machines ; and MWT industrial grade microwaves. The company also supplies associated materials comprising consumables and replacement parts ; and other services, such as training and technical support services. It markets its products to industrial customers and other end-market users in the aerospace, automotive, heavy equipment, energy/oil/gas, and other industries under the ExOne brand name. The ExOne Company was founded in 2005 and is headquartered in North Huntingdon, Pennsylvania. »