

Comparaison de représentations de textes en vue d'une analyse exploratoire

Florian Barbaro*, Fabrice Rossi**

*Université Paris 1 Panthéon-Sorbonne - Laboratoire SAMM EA 4543
florian.barbaro@etu.univ-paris1.fr

**Université Paris Dauphine-PSL - Laboratoire CEREMADE UMR 7534
rossi@ceremade.dauphine.fr

1 Introduction

Dans cet article, nous étudions de façon qualitative l'intérêt de plusieurs représentations vectorielles de textes pour l'analyse exploratoire d'un corpus. Une représentation élémentaire par sac de mots (unigrammes) est comparée à celle obtenue à partir de la distance Sinkhorn entre les textes calculée sur une représentation vectorielle des mots. Puis, une classification des textes ainsi représentés est construite à l'aide de l'algorithme *high-dimensional data clustering* (HDDC). Les différences, entre les représentations, sont illustrées grâce à un nouveau corpus de textes constitués à partir des rapports 8-K de l'entreprise *Wells Fargo* (pour les années 2015 et 2016). Nous analysons la cohérence des classes ainsi obtenues et cherchons à les caractériser en terme de vocabulaire et de sujets spécifiques.

2 Collecte et représentation des textes

Les rapports 8-K qui constituent notre corpus¹ ont été téléchargés avec l'outil EDGAR de la SEC². 248 rapports ont ainsi été obtenus pour l'entreprise *Wells Fargo* (WFC), celle-ci ayant le plus publiée durant la période de l'indice S&P 500³, pour les années 2015 et 2016. Puis, suivant un pipeline classique, les textes ont été pré-traités pour obtenir un dictionnaire de 3778 racines distinctes.

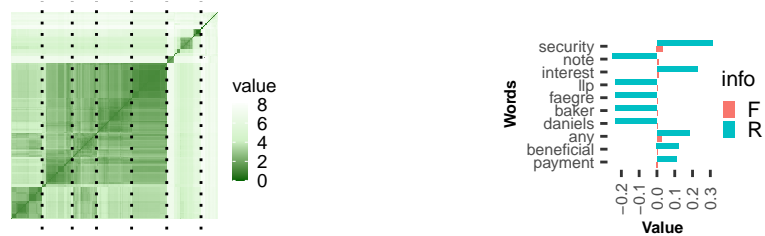
Ensuite, les textes sont représentés de deux manières différentes. La première, qui s'inspire du seul article (Lee et al., 2014) qui à notre connaissance s'intéresse aux rapports 8-K, est basée sur une représentation par sac de mots. La deuxième, repose sur le transport optimal, champ de recherche très actif, et notamment la distance Sinkhorn. Pour obtenir une représentation vectorielle de cette distance, un plongement euclidien en appliquant une *multidimensional scaling* métrique est réalisé.

1. Disponible à cette adresse <https://github.com/FloFloB/article-EGC-2021>.

2. Securities and Exchange Commission, organisme fédéral américain de réglementation et de contrôle des marchés financiers.

3. S&P 500 ou Standard & Poor's 500 est un indice boursier américain basé sur les capitalisations boursières des 500 plus grandes entreprises ayant des actions cotées au NYSE (New York Stock Exchange) ou au NASDAQ (National Association of Securities Dealers Automated Quotations).

Comparaison de représentations de textes en vue d'une analyse exploratoire



(a) Sinkhorn selon classes HDDC.

(b) Grandes rotations d'une classe Uni.

FIG. 1: Exemples de visualisations.

3 Classification et résultats

Dans l'objectif d'étudier l'intérêt de différentes représentations et pour rendre généralisable notre méthodologie, nous avons retenu l'algorithme *High-dimensional data clustering (HDDC)* (Bouveyron et al., 2007). En effet, les représentations vectorielles de texte ayant généralement une grande dimension dégradant les performances des algorithmes de classification, celui-ci a été spécifiquement conçu pour lutter contre le fléau de la dimension. HDDC est un modèle de mélange gaussien avec, comme spécificité principale, la sélection automatique, à l'aide du critère BIC, d'une projection spécifique en basse dimension pour chaque classe.

L'utilisation d'HDDC conduit à la sélection d'une solution à 3 classes pour les unigrammes et 7 classes pour le transport optimal. De manière empirique, nous avons noté qu'une course à la réduction de dimension amenait à l'élargissement du nombre de classes et à une difficulté accrue pour les analyser. Pour comprendre les projections ainsi obtenues, nous adaptions notre analyse selon les représentations. Pour les unigrammes, les matrices de rotation (cf exemple figure 1b, F pour fréquence et R pour rotation) permettent d'avoir une vision claire des mots générant les plus fortes rotations et donc de l'importance qu'ils ont dans la projection de chaque classe. Pour le transport optimal, l'étude se porte sur la matrice des distances Sinkhorn, figure 1a, et notamment des plans de transport afférents. Une analyse de l'entropie de ces derniers, permet de mettre en lumière les similitudes mais aussi les différences lexicales entre deux textes. Une entropie forte signifiant qu'un mot n'est que peu présent dans le texte sur lequel il est projeté et pour lequel il n'y a pas de mots similaires et le contraire si l'entropie est faible.

Notre méthodologie permet donc de mettre en lumière l'intérêt de différentes représentations sur l'analyse exploratoire de texte et d'offrir une alternative intéressante aux traditionnels *topic models*.

Références

- Bouveyron, C., S. Girard, et C. Schmid (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis* 52(1), 502–519.
- Lee, H., M. Surdeanu, B. Maccartney, et D. Jurafsky (2014). On the importance of text analysis for stock price prediction. In *Proc. of the Ninth Int. Conf. on Lang. Res. & Eval. (LREC'14)*.