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 corpus [1] ont été téléchargés avec l'outil EDGAR de la SEC [2]. 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 500 [3], 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 un *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).