Modèles graphiques causaux interactifs pour les données textuelles

Amine Ferdjaoui[*,**], Séverine Affeldt[*], Mohamed Nadif[*]

[*] Centre Borelli UMR 9010, Université Paris Cité, 75006 Paris, France.
[**] SogetiLabs, 147 Quai du Président Roosevelt, 92130, Issy-les-Moulineaux.
<prénom.nom>@u-paris.fr

Résumé. Nous proposons de reconstruire des modèles graphiques causaux à partir de données textuelles via un nouveau package Python appelé `WordGraph`. Ce package facilite l'exploration de grands corpus de documents par des visualisations interactives sous forme de modèles graphiques de mots. Le pipeline WordGraph exploite à la fois les *widgets jupyter* et le *notebook jupyter* pour aider les utilisateurs sans expérience en Python à prendre rapidement en main le pipeline `WordGraph`, qui est entièrement personnalisable. `WordGraph` est disponible via un dépôt GitHub, qui fournit également une courte vidéo présentant l'utilisation de notre système.

1 Introduction

1.1 Co-clustering et données textuelles

Soit un ensemble d'objets décrits par un ensemble de caractéristiques organisés sous forme d'une matrice de données. La tâche consistant à partitionner simultanément les deux ensembles est communément appelée *co-clustering* ou *biclustering*. Elle vise à révéler des blocs/co-clusters d'intérêt dans une matrice de données et aboutit généralement à des clusters de lignes et de colonnes plus pertinents et interprétables qu'avec le clustering qui ne s'appuie que sur l'une des deux dimensions. Depuis les travaux de Hartigan (Hartigan, 1972), le co-clustering a trouvé des applications dans de nombreux domaines tels que l'analyse textuelle et la bio-informatique (Affeldt et al., 2020, 2021). Ainsi, il est particulièrement bien adapté aux matrices de données documents×termes, qui sont par essence de grande dimension, *sparses* et de nature directionnelle. Les algorithmes pour de telles matrices de co-occurrences peuvent être dérivés de différentes approches. Les méthodes de type spectral, qui traitent la matrice d'entrée comme un graphe bipartite entre les documents et les mots, approximent la coupe normalisée de ce graphe à l'aide d'une relaxation réelle (Dhillon, 2001). Les méthodes basées sur un modèle dérivé des modèles de blocs latents (LBM) appropriés (Govaert et Nadif, 2013, 2018), reposent sur des algorithmes d'*expectation-maximization* (Salah et Nadif, 2019). Le co-clustering peut également utiliser des méthodes basées sur la factorisation matricielle telles que la factorisation matricielle non négative (NMF) (Febrissy et al., 2022) ou la tri-factorisation (NMTF) (Salah et al., 2018). Les méthodes basées sur la théorie de l'information,