CORPEX : Analyse exploratoire d'un corpus biomédical à l'aide de la classification croisée

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

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

Résumé. Nous proposons une interface d'aide à l'analyse de corpus via la visualisation interactive de *coclusters* afin d'accompagner l'exploration des thématiques pour un ensemble de textes. Les saisies de l'utilisateur permettent la création ou le chargement d'un corpus de documents, son nettoyage et l'étude interactive et simultanée des termes et des documents. Cet article détaille les fonctionnalités en lien avec la génération dynamique de corpus, notamment dans un cadre biomédical, et également le chargement de matrices documents-termes pour des corpus déjà pré-traités. L'analyse du corpus par la classification croisée (*co-clustering*) et la visualisation conjointe des termes et des documents, suivant le co-partitionnement obtenu sur les deux ensembles, sont des outils efficaces pour une compréhension rapide des sujets abordés dans un corpus. La sauvegarde automatique des résultats permet de relancer facilement différentes analyses par un *co-clustering* approprié et d'obtenir des vues croisées des thématiques à différents niveaux de granularité.

1 Introduction

L'information est aujourd'hui disponible en abondance sous forme textuelle, dans de nombreux domaines. Le Traitement Automatique des Langues (TAL) permet l'automatisation à grande échelle de tâches telles que l'annotation ou la classification. Les méthodes existantes, aisément accessibles via de nombreuses librairies de programmation en `R` ou `Python`, permettent d'analyser et de valoriser de larges corpus comportant par exemple des articles de presse, des compte-rendus d'entretiens ou des commentaires de consommateurs.

Dans le domaine biomédical, de très nombreux articles sont aujourd'hui disponibles en ligne et leur exploitation peut permettre d'identifier des relations d'intérêt pour une éventuelle meilleure prise en charge des patients. Toutefois, on produit de nos jours bien plus d'articles biomédicaux qu'on ne peut en lire. A titre d'exemple, la plateforme PubMed comprend plus de 34 millions de citations pour la littérature biomédicale provenant de MEDLINE, de revues de sciences de la vie et de livres en ligne. Recouper l'ensemble des documents mis à disposition nécessite l'emploi d'approches de TAL avancées. Avec `CORPEX` (*CORPus EXploration*), nous mettons à la disposition de la communauté des chercheurs, mais également des praticiens, une interface ergonomique et légère pour l'exploration de corpus, notamment dans un contexte biomédical.