UNIVERSITÉ PARIS DESCARTES | USPC Université Sorbonne Paris Cité

Sujet de PPD 2016-2017

MIAGE M2 FA

Visualisation des résultats d'un co-clustering

Contexte du sujet

Le *co-clustering* est une technique qui permet de regrouper simultanément des ensembles de mots et de documents. Elle est très utilisée en *Text Mining*. Si des algorithmes performants sont déjà disponibles, l'interprétation des résultats pose encore problème et doit ête supportée par des outils.

Travail à effectuer

L'objectif du projet est de proposer des solutions de visualisation facilitant l'interprétation des résultats d'un *co-clustering*.

La visualisation à réaliser se présente sous forme d'unee matrice réorganisée par blocs, interactive telle que :

http://bokeh.pydata.org/en/latest/docs/gallery/les_mis.html

Le cahier des charges est décrit ci-dessous.

Chaque ligne représente un cluster de documents et chaque colonne représente un cluster de mots.

Au survol d'un bloc, les statistiques de ce bloc sont affichées :

— nombre de mots

— nombre de documents

— critère interne (ex : pkl)

— éventuellement un titre qui représente le bloc (si donné en entrée)

On peut sélectionner des blocs dans cette matrices. C'est de cette sélection

que dépend une autre visualisation détaillée de l'ensemble des blocs sélectionnés.

La visualisation détaillée de blocs peut être, par exemple, un diagramme en bâtons des termes les plus importants.

D’autres idées sont les bienvenues, un ensemble d’exemples est disponible à

l’adresse suivante : https://github.com/d3/d3/wiki/Gallery.

La bibliothèque Bokeh (http://bokeh.pydata.org) permet d’écrire du code Python et de produire une visualisation interractive. Bokeh contient également une partie Javascript qu’il faudra peut-être exploiter si l’API Python est trop limitée mais l’essentiel du travail se fera en Python.

Outils et techniques à mettre en œuvre

- Bibliothèque Python de co-clustering (CoClust).
- Bibliothèque de visualisation Bokeh Python + D3

Encadrant(e)

François Role