WEBMINING

Arthur Katossky
2018/2019
3e Année Filière Marketing

Sujet : En utilisant l'API de l'AFP, proposer un étiquetage des articles récents concernant la thématique "environnement" en sujets[1].

Projet à rendre pour le 1er mars 2019 (avant midi) sous la forme d'un document R Markdown.

1. Collectez, nettoyez et structurez les articles sur la thématique environnement via l'API de l'AFP.

   Vous pouvez notamment utiliser les bibliothèques `tidyverse` *et* `tidytext`.
2. Décrivez brièvement trois méthodes classiques d'étiquetage non-supervisé. Expliquez-en les limites pour la tâche considérée.
3. Appliquez une de ces techniques aux articles de l'AFP avec l'approche *bag-of-words* et interprétez les sujets obtenus.

   Vous utiliserez au moins deux façons de construire la matrice terme-document (ex: présence / absence, compte, TF-IDF...). Il faudra peut-être ruser (adapter le problème, ou adapter les données) pour que certaines méthodes classiques puissent s'appliquer.
4. Décrivez brièvement trois méthodes d'étiquetage non-supervisé adaptées au données textuelles, dont une au moins qui permette de dépasser l'approche *bag-of-words*.
5. Appliquez une méthode *bag-of-words* et une méthode alternative aux articles de l'AFP. Interprétez les sujets obtenus.
6. Dans quelle mesure est-il possible de juger l'efficacité de vos différentes approches? Quelle approche vous semble la plus convaincante?

   Vous pourrez baser votre discussion sur la section 4.9 du livre Web Data Mining *de Bing Liu (2011).*

Ressources :

- https://thinkr.fr/text-mining-et-topic-modeling-avec-r
- https://www.tidytextmining.com
- https://rmarkdown.rstudio.com/index.html

1 Les sujets (an anglais *topics*) sont non exclusifs : un article peut appartenir simultanément à plusieurs sujets. J'aurais tendance à appeler « catégories » un étiquetage exclusif. Il s'agit d'un problème *non supervisé*, puisqu'il n'existe pas de « vrai » étiquetage.