STICB545 – Traitement automatique de corpus

Devoir 2 – Extraction d'information

Nassim Derras – M LING

Matricule : 000334827

Ce devoir est sauvé dans le dossier *devoir 2*

1.

Fait. Je n'ai téléchargé que les ***100*** premiers fichiers pdfs (dû à un manque d'espace mémoire). Voir dossier../data/pdf

2.

Fait. Voir dossier ../data/txt

3.

Dossier de donnés téléchargés de l'UV: ./bulletins/

Année choisie : **1855**

4.

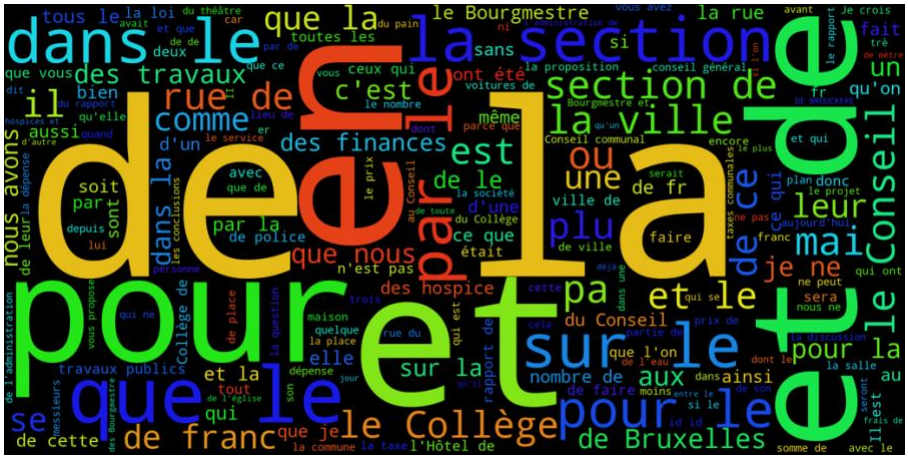
Ici, j'ai d'abord agrégé les fichiers 13 fichiers de 1855 en un seul (via le script *aggregator_bY.py*). Ensuite, *s1_keywords_1.py* lit le fichier *1855.txt* afin d'en tirer les **keywords.**

5.

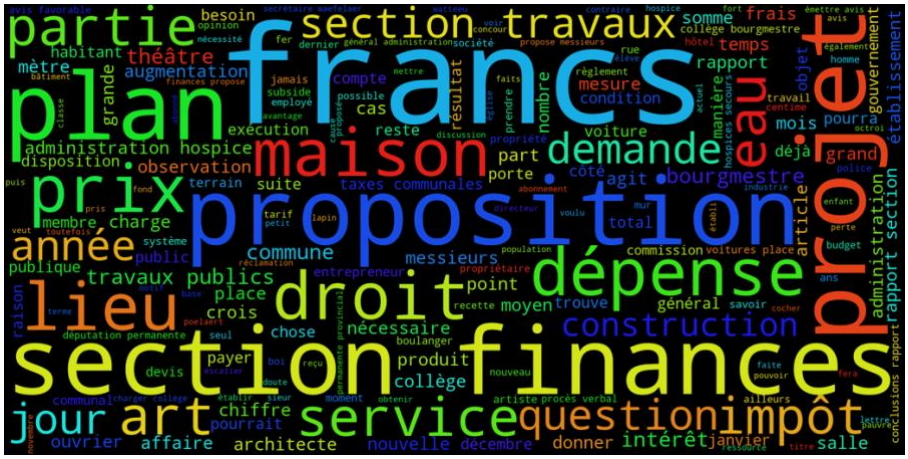
On va ici ajouter le script *filtering.py* afin de garder que les mots « intéressants ». Comme commentaire, j'ai décidé de laisser ces scripts de façon séparée et d'introduire manuellement l'année à chaque fois (bien évidemment, cela peut être amélioré). On lance donc le ficher *filtering.py* pour l'année 1855 afin de générer un fichier *1855_keywords.txt* que l'on sauve dans le dossier *tmp*.

Les stopswords ont été enrichi grâce à la table fournie par countwordsfree.com au format *json*. En plus de ces stopwords « externes », j'ai ajouté une liste de mots afin qu'il n'apparaissent pas dans le *wordcloud.*

Wordcloud sans le filtering



Wordcloud avec le filtering.



6.

La fonction *s3_ner.py* a été utilisé afin de déterminer les dix personnes, organisations et lieux les plus récurrents en 1855. Inutile d'afficher les résultats mais il est à noter que le script fourni de meilleurs résultats en utilisant le fichier *1855.txt* que le fichier *1855_keywords.txt*. (peut-être dû au fait que les majuscules sont retirées dans *1855_keywords.txt*. On peut penser aussi que les noms « belges » avec un petit « de » seront « mal nommés ».)

7.

Dix phrases ont été sélectionné dans le fichier *1855.txt*. le script *s4_sentiment.py* analyse les taux de positivité et de subjectivité de chacune de ces phrases.

8.

…

9.

Fait.