Éloi Fournier **2020-04-16**

Projet final: *Making-of*

1. Liens utilisés

- http://m.ledevoir.com/article- (suivi de tous les numéros d'article parus en 2019)
- https://www.lemonde.fr/archives-du-monde/ (suivi des 365 dates de 2019)
- https://stackoverflow.com/questions/28143395/python-requests-get-invalidschema-error
- https://docs.python.org/3/library/urllib.parse.html
- http://jhroy.ca/uqam/edm5240/BeautifulSoup-DocAbregee.pdf
- https://bit.ly/2yW7cs1
- https://stackoverflow.com/questions/54634465/error-10060-using-python-requests-library
- https://www.filesmerge.com/merge-csv-files

2. Démarche

2.1 Le sujet

J'ai choisi ce sujet, tout d'abord, en raison d'une étude réalisée par Marie-Éva de Villers il y a de cela plusieurs années. Mme de Villers avait comparé les mots utilisés en 1997 dans les journaux *Le Monde* (France) et *Le Devoir* (Québec) afin d'observer les différences de vocabulaire entre les deux pays.

Son étude relevait les différents mots utilisés par ces journaux, mentionnant que les mots se recoupent à 77 % entre ces deux médias papier. Elle a d'ailleurs publié un ouvrage, *Le vif désir de durer*, à ce sujet quelques années plus tard.

C'est donc cette façon de faire qui a inspiré mon projet. Cependant, en raison de contraintes de temps et de technique, mon étude s'est limitée aux 100 mots les plus utilisés par *Le Monde* et *Le Devoir*.

2.2 Les technologies

Pour arriver à réaliser le projet, j'ai utilisé la bibliothèque BeautifulSoup pour extraire tous les articles publiés sur les sites web du *Monde* et du *Devoir* en 2019. Celle-ci permet, en séparant les différentes classes de texte et en spécifiant celles qui doivent être trouvées par BeautifulSoup, de récupérer le nécessaire pour avoir le corpus de tous les articles. La fonction *parser* de BeautifulSoup était essentielle à la réussite du projet.

Pour la deuxième étape du travail, j'ai utilisé la bibliothèque spaCy. Cette dernière réalise le traitement de la langue, et était très utile pour en arriver à une fine analyse des mots les plus souvent utilisés par *Le Devoir* et par *Le Monde*.

2.3 Les contraintes

Plusieurs difficultés techniques ont rendu le travail de récupération des données très ardu. Tout d'abord, des coupures de connexion wi-fi m'ont forcé à recommencer à zéro la cueillette d'articles à plusieurs reprises. Par la suite, après discussion avec le professeur Jean-Hugues Roy, il a déterminé qu'il serait préférable d'obtenir plusieurs fichiers csv. C'est donc ainsi que j'ai procédé. À la fin de la période de moissonnage, j'ai trouvé un site web permettant d'assembler mes fichiers csv pour en créer un seul par journal, ce que j'ai fait avec joie.

De plus, le processus pour trouver les mots les plus utilisés a été très difficile pour mon ordinateur. Cela a pris de nombreuses heures à Python et m'a découragé de pousser le travail plus loin, car je n'aurais pas été en mesure de le remettre à temps. C'est pourquoi mes résultats ne comptent que les cent mots les plus fréquents pour *Le Devoir* et *Le Monde*.

3. Les résultats

Malheureusement, après de très longues heures d'attente durant le processus de Python, les mots ne sont pas sortis. C'est possiblement à cause d'une coquille dans le codage. Cependant, pour pouvoir être en mesure de remettre le travail à temps, je ne peux malheureusement pas présenter les résultats de mon outil. Il s'agit peut-être d'une seule petite chose à changer… Mais les longues heures requises pour arriver au résultat me contraignent à remettre ce travail de manière légèrement incomplète, bien que l'outil ait été totalement conçu et qu'il soit prêt à être utilisé.