Charles Mathieu
Journalisme de données II
16 mars 2020

***Making of* de mon travail final**

Pour mon travail final, j'ai décidé d'écrire un reportage court portant sur l'analyse des communiqués de presse publiés par l'Autorité des marchés financiers (AMF). J'ai choisi ce sujet parce qu'au *Journal de l'assurance*, on fait fréquemment référence à cet organisme de régulation.

En fait, je trouvais que le Journal avait besoin d'un article un peu plus *relax.* Dans le sens où le but de cette recherche n'était pas de trouver un problème dans la rédaction de leurs communiqués. Je souhaitais faire une analyse de données qui pouvait simplement permettre de savoir le nombre de fois que l'industrie de l'assurance est mentionné par l'AMF et de quoi il est question lorsque c'est le cas. Le tout devrait intéresser nos lecteurs.

Voici l'URL menant à la section « Actualités » de l'Autorité : https://lautorite.qc.ca/grand-public/salle-de-presse/actualites/. À cet endroit, il est possible de filtrer les communiqués pour aller chercher seulement les textes ayant le sujet « assurance ». C'est sur cette page que j'ai pu regrouper les différentes informations concernant le nombre total de communiqués publiés depuis 2004 et le nombre de fois que les sujets sont mentionnés dans les nouvelles.

Les informations importantes ont toutes été regroupées dans un fichier Google Sheets que voici : https://docs.google.com/spreadsheets/d/1jYgtSJsGMg4eZHrs63ce4qqO1VRcx3imwyliaWz5Uc8/edit?usp=sharing.

Premier script : déroulement et difficultés

Pour réaliser mon travail, j'ai fait deux scripts différents. L'un d'entre eux avait pour but, à l'aide de BeautifulSoup, de générer un fichier CSV contenant tous les communiqués et les mises en garde diffusés par l'Autorité. J'ai fait une première requête à cette URL :

« https://lautorite.qc.ca/grand-public/salle-de-presse/actualites/?tx_solr[filter][0]=category%3A50&tx_solr[page]= ».

J'ai par la suite généré une liste de 42 pages sur lesquelles on retrouve au total 419 liens allant vers des communiqués de l'AMF. Sur ces pages, j'ai pu extraire la date, le titre, et l'URL de chaque article pour les placer dans ma liste. Il ne me restait que les sujets du texte et le texte en tant que tel. Je suis allé chercher le tout dans les nouvelles publiées en faisant une deuxième requête à chacune d'entre elles.

Le texte a été difficile à extraire. J'ai dû utiliser une autre méthode pour générer mon texte et mes données, en utilisant « html5lib » à la place du *parser* que nous avions utilisé lors de nos cours. Certaines parties du texte, qui contenaient une image menant à un lien, faisait stopper le *parser* à chaque fois, et prenaient ainsi seulement une fraction du texte.

Voici le lien m'ayant permis de régler le problème : https://www.crummy.com/software/BeautifulSoup/bs4/doc/#parser-installation.

Deuxième script et déroulement

Le deuxième script en était un de traitement automatique du langage en utilisant SpaCy. J'avais envie de voir quels mots étaient utilisés le plus souvent, alors j'ai utilisé cette bibliothèque pour être en mesure d'analyser les textes des communiqués. J'ai donc importé le fichier CSV que j'ai généré précédemment pour analyser seulement la quatrième colonne, qui contenait le texte.

Pour savoir quels produits étaient mentionnés par l'AMF, j'ai décidé de générer une liste de *bigrams* et d'aller chercher uniquement les paires de mots qui contenaient « surance » (pour « assurance » ou « insurance »). Ainsi, j'ai pu voir que « assurance dommage », est mentionné plus de fois que « assurance personne », ce qui peut être intéressant pour nos lecteurs. Finalement, pour compter la fréquence à laquelle on retrouvait ces *bigrams*, j'ai utilisé l'outil nommé Counter.

Le seul problème que j'ai eu avec ce script est celui des nombres et des espaces qui n'étaient pas inclus dans les *stop-words*. J'ai donc utilisé « token.like_num == False and

token.is_space == False » afin de les enlever, chose que nous n'avions pas vue en classe. Voici le lien m'ayant permis de régler mon problème : https://spacy.io/api/token.

Conclusion

En conclusion, je suis content du travail réalisé dans les circonstances actuelles. Évidemment, j'aurais bien aimé créer mon outil avec le Registraire des entreprises du Québec, mais c'est un projet que je garde pour après mon baccalauréat.

Néanmoins, la création de ce script m'a permis de mettre en pratique une grande partie de la matière apprise en classe. Je pourrais également me servir de mon code pour éventuellement me créer un outil m'avertissant aussitôt qu'un communiqué de l'AMF est publié. À suivre…