# *MAKING-OF* DU REPORTAGE

Par Thomas Dufour (DUFT30099602)

- **Pourquoi ce sujet?**

Au début de la session d'hiver, Camille et moi sommes tombés sur un article d'un média américain nommé *The Pudding* refaisant l'histoire du 20e siècle à travers la couverture médiatique du New York Times. L'idée était de voir quel pays était le plus cité par le célèbre média américain au cours des cent dernières années. De la Première guerre mondiale à la seconde, de la crise des missiles cubains à l'attentat du 11 septembre 2001, l'article donne une bonne idée de l'histoire récente des États-Unis. Dans le cadre de ce cours sur le journalisme de données, nous avons voulu réaliser le même exercice avec un média québécois (Le Devoir) afin de voir si les pays au centre des préoccupations journalistiques étaient différents au Québec que chez nos voisins du sud. Voilà l'idée au coeur de notre démarche.

- **Quel.le.s outils ou technologies avez-vous utilisées?**

Nous avons obtenu les données nécessaires à ce reportage de deux manières. Premièrement, nous avons utilisé un code permettant d'accéder à l'API du Devoir de 1910 à 2011 et de comptabiliser les pays cités, puis nous avons rédigé un autre code permettant d'aller chercher tous les articles du journal entre 2011 et nos jours.

Pour écrire la première partie du code, nous nous sommes inspirés d'un script de Jean-Hugues Roy, professeur dans le cadre de ce cours [1]. Le script de Jean-Hugues permet de se connecter à chaque édition du Devoir entre 1910 et 2011, d'extraire le texte et de *tokeniser* tous les mots. Ce dernier processus consiste en un code qui *parse* à travers un texte et qui isole les *tokens* (mots repérés par le script).

Nous avons modifié le code de Jean-Hugues puisque notre objectif était différent. Une fois les documents *tokenizés,* nous avons écrit un code permettant de compter le nombre d'occurrences de chaque pays dans les listes de mots. En utilisant une boucle *for* dans un *csv* comprenant tous les noms de pays du monde (en ajoutant l'URSS, la Tchécoslovaquie et la Yougoslavie), nous avons pu créer un autre csv comprenant le nombre d'occurrences des pays dans chaque édition du Devoir. Le script a roulé sur deux ordinateurs pendant au moins deux jours.

Une fois les résultats sommaires obtenus, il nous a fallu les comptabiliser afin qu'ils puissent être utilisés par les étudiants de Polytechnique. En utilisant la fonction *Groupby* dans *Pandas,* nous avons regroupé les données par mois. Nous avons ensuite divisé les occurrences par le nombre d'éditions mensuelles afin de pondérer nos données. Nous

[1] https://github.com/jhroy/CdJ_LeDevoir

avons dû modifier l'ordre des mois et des années afin de nous conformer au format attendu par les étudiants de Polytechnique.

Dans un deuxième temps, nous avons écrit un script permettant d'aller chercher tous les articles du Devoir entre 2011 et 2019. Le script utilisait UrlLib2 pour se connecter aux liens du Devoir et BeautifulSoup afin d'extraire le texte du code HTML des pages. La fin de script comptabilisait le nombre de pays avec la même méthode que celle citée plus haut. Faute de temps pour la compilation, les résultats de ce script n'ont pas été compilés dans l'article final.

- **Pourquoi les avoir choisi.e.s?**

Énumérons les outils afin d'expliquer leur choix dans le cadre de notre travail. La librairie *Pandas* est sans aucun doute l'outil qui a été le plus utile dans ce travail. Comme nous avions des fichiers de centaines de milliers de lignes, *Pandas* a permis d'être plus efficace dans la modification, le traitement et le calcul des données. Il nous évitait aussi d'avoir recours à une sémantique plus complexe afin d'arriver au même résultat.

Pour ce qui est de UrlLib2 et de BeautifulSoup, ces deux outils ont été choisis pour leur simplicité et leur efficacité. BeautifulSoup a été particulièrement utile pour retrouver dans les articles les extraits de texte désirés.

- **Quels problèmes avez-vous éprouvés?**

Le parcours menant au reportage n'a pas été sans embûches. D'abord, nous avons éprouvé de la difficulté à extraire le texte de fichiers PDF. Il arrivait souvent que le code « *tokenize* » mal les mots et que plusieurs occurrences ne soient pas comptabilisées. Nous avons tenté de contourner ce problème en enlevant le caractère « - » du texte. Bien que cela ait aidé pour la précision du code, ce dernier n'était pas parfait.

La collaboration avec l'équipe de Polytechnique a été assez difficile. Les deux étudiants n'ont pas communiqué avec nous avant d'élaborer leur visualisation. Nous n'avons donc pas pu leur donner des conseils qui auraient permis de rendre la visualisation plus digeste. J'ai eu l'impression que le travail n'était pas collaboratif, mais plutôt compartimenté.

Somme toute, j'ai beaucoup appris au cours de ce processus. J'ai développé mes capacités à jouer avec une base de données et interagir avec un ficher csv. Bien que j'avais déjà quelques notions en informatique, j'ai tout de même eu l'impression que l'élaboration de ce reportage représentait un défi.

# Bibliographie:


https://pandas.pydata.org/

https://pypi.org/project/pycountry/

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.iloc.html

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.divide.html

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

https://docs.python.org/3/library/index.html

https://github.com/jhroy/CdJ_LeDevoir

http://www.atlas-monde.net/tous-les-pays/

https://stackoverflow.com/

https://www.nltk.org/

http://numerique.banq.qc.ca/patrimoine/

https://www.ledevoir.com/