Catherine Savoie

**Projet final - EDM4466**

Mon projet final consistait à faire la comparaison entre les expressions les plus couramment utilisées dans le journal *Le Devoir*, publié ici au Québec, et le quotidien français *Le Monde* lors de la dernière année. L'objectif d'un tel travail est de mettre de l'avant les différences linguistiques au point de vue de l'écriture entre deux journaux équivalents, établis sur deux continents différents.

**Étape # 1 :**

D'abord, il faut passer du temps à naviguer à travers des archives en ligne du *Devoir* pour comprendre la structure de ce site web. Dans le fichier **« travailfinal.py »** de l'outil Python, j'ai écrit un script pour obtenir les liens de tous les articles publiés par le journal en 2018. Mon travail a été principalement de faire une boucle avec le fichier CSV qui contient tous les numéros d'identification des articles publiés au cours de l'année 2018. L'objectif était donc d'ajouter ces numéros à la fin de l'URL suivant : http://m.ledevoir.com/article-.

À la fin de ce script, le fichier « **ledevoir.csv** » a été créé pour enregistrer et regrouper les nouveaux liens URL dans un nouveau document CSV.

**Étape # 2 :**

Maintenant, le quotidien *Le Monde*.

Un nouveau script, a été mis en place, du nom de « **travailfinal2.py** ». La particularité avec les archives du journal *Le Monde* est qu'ils sont regroupés pour chaque jours d'une année, donc leurs archives ressemblent à un grand calendrier annuel. Il ne suffit que d'appuyer sur le lien d'une journée pour être mené vers l'ensemble des liens URL de chaque article publié durant cette même journée. Une formule bien différente du journal *Le Devoir.*

Une boucle a été créée pour extraire l'URL de chaque article de 2018 à l'aide de *Beautiful Soup* et de la fonction *string.* Pour se faire, il a fallu trouver où se trouvait l'hyperlien de chaque article dans structure du document HTML du journal. Par la suite, il faut de nouveau procéder à l'exportation des données dans le fichier CSV « **lemonde.csv** » à l'aide d'une formule similaire à celle du journal *Le Devoir.*

**Étape # 3**

Après la création des deux scripts suivis des deux fichiers CSV qui n'ont jamais réussi à enregistrer dans mon ordinateur, j'ai tout de même fait l'essai de réaliser deux nouveaux scripts Python pour « traiter » les données récoltées plus haut. Le premier script, qui porte le nom de « **traitementdevoir.py** », a été créé pour l'analyse des données du *Devoir,* alors que le second se nomme « **traitementlemonde.py** » lié évidemement au journal *Le Monde*.

En fin de parcours, ma tentative de traitement des données s'est annoncée infructueuse étant donné l'impossibilité de créer un lien avec chacun des fichiers csv. J'ai tout de même tenter de réaliser deux scripts qui auraient pu traiter les données d'un document CSV fonctionnel. Évidemment, la difficulté de réaliser un script de type NLTK était énorme en raison de l'impossibilité d'obtenir le résultat des deux boucles des scripts de début de parcours **« travailfinal.py »** et « **travailfinal2.py** ».

Sources:

Toutes les notes de cours disponibles sur le site web *Journalisme de données II*
codesDevoir2018.csv
Archives du Devoir : https://www.ledevoir.com/recherche?expression=&rechercher=
Archives du Monde: https://www.lemonde.fr/archives-du-monde/
Loops : https://www.w3schools.com/python/python_for_loops.asp
Boucle : https://www.youtube.com/watch?v=BrknhzrHm8w
Boucle : https://www.youtube.com/watch?v=excGUISppC4
NLTK: https://www.youtube.com/watch?v=gRk53jBPYvE