EDM 4466

Journalisme de données II

Présenté à Jean-Hugues Roy

MAKING-OF

Par

Simon Duclos

UQÀM

Le 16 avril 2020

TABLE DES MATIÈRES

1. INTRODUCTION.....3

2. MA DÉMARCHE.....4

2.1 POURQUOI CE SUJET?.....5

2.1.2 LES OUTILS ET TECHNOLOGIES UTILISÉES.....6

3. LES PROBLÈMES ÉRPOUVÉS.....7

4. CONCLUSION.....7

5. BIBLIOGRAPHIE.....8

1. Introduction

Au sein de ce document *Making-of*, j'expliquerai en détail le parcours que j'ai effectué pour ce travail de script final, dans le cadre du cours *Journalisme de données II (EDM 4466)*. Le but de cet exercice était de trouver un sujet donné et d'en extraire une analyse quelconque en utilisant divers processus de script, notamment avec l'utilisation du logiciel Python.

2. Ma démarche

2.1 Pourquoi ce sujet?

Tout d'abord, bien que ce sujet m'ait été proposé par toi au départ, c'est un sujet qui malgré le fait qu'il ne soit pas venu à mon esprit m'a semblé très intéressant pour plusieurs raisons.

Le principal intéressé du projet, Monsieur Horacio Arruda, médecin spécialiste en santé communautaire, est devenu en l'espace de quelques semaines, une véritable star de la sphère médiatique. Depuis les conférences de presse quotidiennes en lien avec la COVID-19, il a su, de par sa façon de s'exprimer, le ton avec lequel il communique et les termes utilisés à captiver les gens, plus particulièrement les milléniaux, qui ont fait de lui plusieurs « *mêmes* ».

Outre un intérêt envers le sujet de l'étude, tel que mentionné plus haut, mon but était de faire une étude du langage de M. Arruda et ainsi d'être en mesure de comprendre comment il fait pour capturer l'imaginaire des Québécois.

2.1.2 Les outils et technologies utilisées

Pour essayer de réaliser mon objectif, j'ai eu recours à divers outils et technologies. Tout d'abord, il y a bien sûr, le logiciel Python qui est la base de ce projet d'étude et qui permet de réaliser des scripts, sur lequel je me suis principalement concentré. Au sein de ce logiciel, il y a cependant, plusieurs outils et méthodes d'analyse qui peuvent être faites et que j'ai tenté d'utiliser pour permettre d'analyser le langage de Monsieur Arruda.

La Production de fichier(s) .CSV/.txt

Avant de débuter mon processus d'analyse, j'ai créé deux fichiers CSV/.txt (*Horacio.txt et Arruda.txt*), afin de pouvoir regrouper les différents intervenants au sein des conférences de presse de François Legault et ainsi pouvoir comparer le langage de chacun (le temps de parole, les mots les plus utilisés, etc.)

La Tokenization

Au sein de mon script, j'ai tenté d'utiliser un processus d'analyse du langage qu'on appelle «La *Tokenization*». Ce processus permet simplement de séparer un texte complet en mots simples que l'on appelle des «*tokens*».

Le Stemming

Au sein de mon script, j'ai tenté d'utiliser la notion de *Stemming*. Ce processus permet de ramener un mot à sa plus simple expression ou à sa racine. Par exemple, si au sein d'un texte on retrouve le terme « pêcheur », cette notion déterminera que le terme « pêcheur» et tous les autres synonymes, sont issus du mot «pêche», qui est le termine simplifié ou d'origine.

SpeechInMinutes

Concrètement parlant, SpeechInMinutes est un site web permettant à son utilisateur d'entrer un nombre de mot(s) spécifique(s) et de savoir combien de temps un interlocuteur a parlé. On peut choisir entre trois niveaux d'intensité selon la vitesse à laquelle la personne donnée parlait (lent, normal ou rapide).

En utilisant ce site web par exemple, j'ai découvert que M. Arruda aurait parlé, en moyenne, pendant 21 minutes (assument qu'il utilisait un débit moyen) lors de la conférence de presse du 12 mars dernier. Évidemment, j'aurais souhaité regrouper toutes les conférences de M. Arruda et en faire une moyenne me permettant de savoir le nombre de minutes parlées durant l'ensemble des conférences, en opposition avec le temps de parole des deux autres intervenants aux conférences de presse, soit M. François Legault et la ministre de la Santé et des Services sociaux, M. McCann.*

*Tu peux te référer au document *Arruda.txt* contenant la conférence de presse en question avec les 2784 mots prononcés par M. Arruda.

3. Les problèmes éprouvés

Au cours de ce projet j'ai eu évidemment, bien des embûches, puisque je ne suis pas arrivé à extraire les informations que je voulais en dégager. La plupart du temps, cesdites embûches prenaient souvent la forme de code d'erreurs dans mon terminal Python. Voici donc une liste des nombreuses erreurs auxquelles j'ai fait face et qui m'ont empêché de mener à terme ce projet.

Codes d'erreurs

- «Object not subscribable» : https://stackoverflow.com/questions/216972/what-does-it-mean-if-a-python-object-is-subscriptable-or-not
- «D is not defined» : https://stackoverflow.co«m/questions/2612948/error-in-python-d-not-defined
- «List index is out of range» : https://stackoverflow.com/questions/16005707/index-error-list-index-out-of-range-python et https://www.stechies.com/fix-list-index-out-range-python/
- «Readline has no attribute redisplay» : https://github.com/pyreadline/pyreadline/issues/49

4. Conclusion

Pour conclure, je n'ai pas nécessairement été en mesure de mener ce projet à terme, par contre, la procédure du projet m'a permis d'éclaircir plusieurs questions que j'avais auparavant quant aux différents processus que j'ai tenté d'utiliser dans mon script et sur le fonctionnement général de Python, à travers les nombreuses vidéos tutoriels que j'ai regardés.

5. Bibliographie/Liens consultés

Voici une liste des différents liens consultés au cours de ma recherche (excluant les liens consultés afin de résoudre les codes d'erreurs montrés ci-haut).

Assemblée nationale du Québec*

- Les conférences de presse quotidienne de François Legault : http://www.assnat.qc.ca/fr/actualites-salle-presse/index.html

*Notez que je n'ai pas inclus tous les liens de toutes les conférences de presse de François Legault au sein de ce document, mais que j'ai effectué **une recherche avancée**, me permettant de trouver toutes les conférences de presse de ce dernier via le lien ci-haut et qu'elles sont rassemblées dans le document *Horacio.csv* (du 12 mars 2020 au 8 avril 2020 inclusivement).

SpeechInMinutes

-SpeechInMinutes.com : http://speechinminutes.com/

Tutoriels – EDM4466

- Capsule 1 et Capsule 2 sur le moissonnage de données (VOIR) : https://journalisme-uqam.gitbook.io/edm4466-h2020/travaux/tutoriels

-Comment lire le contenu d'un fichier CSV : https://github.com/Journalisme-UQAM/edm4466-h2020/blob/master/lectures/lirecsv2.py

-Les principales commandes SQL : https://gist.github.com/jhroy/21acbdf067adc6721b20fbb8aabe020a#file-mysql-requetes-sql

-Notes sur le cours 1 pour Python : https://github.com/Journalisme-UQAM/edm4466-h2020/blob/master/lectures/tal-nltk.py

-Notes sur le cours 2 pour Python : https://github.com/Journalisme-UQAM/edm4466-h2020/blob/master/lectures/Cours2-H2020.pdf

-Tutoriel de traitement automatique du langage avec NLTK : https://github.com/Journalisme-UQAM/edm4466-h2020/blob/master/lectures/tal-nltk.py

-Tutoriel sur l'utilisation de BeautifulSoup : http://jhroy.ca/uqam/edm5240/BeautifulSoup-DocAbregee.pdf

YouTube

-24 Python NLTK Tokenization : https://www.youtube.com/watch?v=Y39v2W6N6Lw

-Creating Word Count Frequency in Python (Tokenization) : https://www.youtube.com/watch?v=UgXe0dUGUIs

-CSV Files in Python || Python Tutorial || Learn Python Programming : https://www.youtube.com/watch?v=Xi52tx6phRU

-Machine Learning - Text Classification with Python, nltk, Scikit & Pandas : https://www.youtube.com/watch?v=5xDE06RRMFk

-Natural Language Processing in Python: Part 5 -- Stemming and Lemmatization : https://www.youtube.com/watch?v=P2PMgnQSHYQ

-Natural Language Processing (NLP) Tutorial with Python & NLTK : https://www.youtube.com/watch?v=X2vAabgKiuM

-Natural Language Processing With Python and NLTK p.1 Tokenizing words and Sentences : https://www.youtube.com/watch?v=FLZvOKSCkxY

-Stemming And Lemmatization Tutorial | Natural Language Processing (NLP) With Python | Edureka : https://www.youtube.com/watch?v=p1ccbR2P_xA

-Stemming | Natural Language Processing with Python and NLTK : https://www.youtube.com/watch?v=54k2JV3HlFc

-Tokenizing Words Sentences with Python NLTK : https://www.youtube.com/watch?v=A5n7tsZctwM