Université du Québec à Montréal
PROJET FINAL

Travail présenté à
Jean-Hugues Roy
Dans le cadre du cours
-
— EDM5240 —
-
Technologies de l'information
appliquées au journalisme

Par
DEMJ23108808 — Jean-Baptiste Demouy
En équipe avec ,
Hannah R. Vilandré
Mélissa Aubert
Théo Sardaigne
Programme
Baccalauréat en Communication (journalisme)

SURTITRE : Une analyse de sentiment des journalistes sur Twitter

TITRE : Médias et médias sociaux

1. LE SUJET

Le sujet choisi se veut une analyse des sentiments des internautes utilisateurs de Twitter des journalistes.

L'idée est partie de l'API Twitter proposée dans le cadre du cours et avec laquelle j'ai commencé à m'amuser après que tu nous aies montré la technique. Vu que nous avons choisi d'être en équipe avec Polytechnique, nous avons commencé à imaginer comment rassembler des milliers et des milliers de tweets et essayer d'en extraire ce que les gens y disent des journalistes.

On voit la situation des journalistes se dégrader de par le monde avec la montée du populisme. En France. J'étais même tombé sur un article qui parlaient de jeunes qui arrêtaient le journalisme à cause de la précarisation et de la violence. Je n'ai plus le lien, mais juste en tapant « précarisation journalisme » dans *Google*, on trouve des articles comme, « Comment le journalisme s'est détérioré » dans *Challenges*, « L'indépendance du journalisme, toujours plus précaire » de l'*AJIQ* ou encore « Le journalisme en crise » au *Devoir*.

Et avec la montée du populisme, d'un président comme Macron qui dit en pleine conférence de presse autour du scandale Benalla, que « les journalistes ne recherchent plus la vérité », on voit qu'il y a un vrai problème démocratique. La liste est encore longue d'exemples.

2. Technologie utilisée et problèmes rencontrés (tous les .py et .ipynb seront joints en fin de pdf)

Nous sommes partis de l'API Twitter avec la technique qui nous avait été donnée en cours. J'en ai profité pour utiliser cette API durant le moissonnage de mi-session. C'est à cette occasion que je me suis rendu compte que tout fonctionnait bien à part les retweets. Ceux-ci ne sortaient qu'avec un extrait. J'ai donc cherché pendant des heures jusqu'à trouver la solution.

Aussi, avec le script original, nous étions limités aux 500 derniers tweets. Il a fallu trouver le moyen d'aller plus loin pour récupérer plus de données et sur plusieurs jours. L'équipe de polytechnique

a amélioré mon script originel en trouvant la solution pour aller jusqu'à 7 jours en arrière. En ce faisant, ils ont grandement complexifié le script pour notamment analyser les sentiments des tweets, vaderSentiment. Pour la visualisation de données, il fallait, au-delà de la collecte, un moyen de donner un sens à ces tweets.

Mais nous nous sommes vite retrouvés bloqués devant la complexité de leur script. Ce qui a créé quelques frustrations entre les deux équipes.

Pour remédier à cela, j'ai réussi à intégrer le code pour récupérer 7 jours de données. Et avec de l'aide, je dois l'avouer, j'ai réussi à intégrer vaderSentiment à mon python.

Donc maintenant, le moissonnage pouvait commencer. J'ai donc créé une boucle pour deux séries d'utilisateurs twitter, les médias et les journalistes. J'ai pu récupérer environ 40 000 tweets par médias dans deux csv différents. 80 000 tweets en tout.

Il a juste fallu faire un peu de nettoyage pour se débarrasser des doublons, environ 4000 par .csv, récupérés en utilisant Google Docs.

Ce qui a donné le résultat suivant.

Mais le résultat a été plutôt décevant même si le design est très intéressant et clair.

Le conseil a donc été d'analyser nos données avec nltk avec la fonction tokenize pour analyser nos fichiers csv. Ce à quoi nous nous attelons depuis quelques jours.

3. Entente avec l'équipe de polytechnique

L'entente avec l'équipe de polytechnique a été assez particulière. Tout d'abord, la différence de niveau technique dans l'élaboration de scripts. Étant parti de notre script initial, ils l'ont complexifié d'une telle façon que nous nous sommes retrouvés devant un script que nous ne parvenions pas à faire marcher et devant une équipe incapable de nous guider à travers celui-ci.

Cela a créé quelques tensions et quelques échanges un peu plus acides. Mais le résultat développé par l'équipe est tout de même intéressant même si le résultat des données que nous cherchions avec la fonction *vaderSentiment* fut décevante.

Autre déception est que nous avons runné deux scripts différents, celui des médias et un peu plus tard celui des journalistes. L'équipe de poly ayant une deadline avant la nôtre, n'ont intégré dans

leur design que les médias. Ils nous avaient parlé d'intégrer les journalistes par la suite mais ils n'ont jamais ajouté ces données et après leur rendu, nous n'avons plus entendu parler d'eux.

4. Documentation

Lien github vers toute la documentation :
https://github.com/JBDemouy/JBDemouy_FinDeSession

Lien Dropbox vers les fichiers .py .ipynb et .csv : https://bit.ly/2IEORCV

Lien vers les deux google docs pour le tri des tweets :

- Tri pour les journalistes : https://bit.ly/2IKVSSL
- Tri pour les medias : https://tinyurl.com/y5xlhueq

Drive de notre collaboration avec Poly.

Sources de documentation :

- retweet entier
- vaderSentiment
- nltk corpora and corpus
- nltk corpus
- nltk tokens
- nltk_stop_words
-