Université du Québec à Montréal

PROJET FINAL

Travail présenté
à Jean-Hugues Roy
Dans le cadre du cours EDM5240
le 17 avril 2019

Par Hannah Raphaëlle Petitclerc-Vilandré
PETR18618703

Making-of

Le sujet ou « le comment du pourquoi »

L'expression "Fakenews" est devenue populaire pendant la campagne présidentielle américaine de 2016, et depuis, on a l'impression que la crise de confiance envers les médias augmente de jour en jour. On a voulu creuser et vérifier cette tendance à partir de l'API de Twitter. Puisqu'il est possible de « moissonner » des milliers de *tweets*, et de *retweets* sur cette plateforme nous pouvons ainsi analyser les *tweets* adressés aux médias, aux journalistes, ainsi que les *hashtags*, les concernant, les plus utilisés.

- Nous voulions des données en français et pour ce faire nous avons décidé de nous concentrer sur les médias et les journalistes français et québécois.
- Nous voulions répondre aux questions suivantes : quel est le sentiment global du public vis-à-vis des médias -plutôt négatif, ou positif- ? Est-ce pire ou mieux au Québec ou en France ? Dans le cas échéant -quels médias accusent une crise du public plus importante – quel type de média la presse écrite ? La radio ? La télévision- ?
- Enfin, nous avons décidé qu'il serait judicieux de faire une analyse de texte pour mieux comprendre ce que l'on reproche à ceux-ci. Quels mots sont les plus utilisés quand on s'adresse à un média et à un journaliste (positif ou négatif) ? Quels sont les reproches formulés à leur égard ? Quels Hashtags leur sont associés?

→ Les données récupérées (moissonnées sur une courte période de temps et sur une seule plateforme) ne permettent qu'une analyse partielle du phénomène. Néanmoins, cela permet de dégager des tendances et faire le portrait des utilisateurs de Twitter sur une période donnée.

Les outils utilisés

Nous avons utilisé **l'API** de Twitter, car c'est assez facile d'accès et d'utilisation. Elle permet de faire du moissonnage assez efficacement des tendances (#) des journalistes et des médias(@).

Nous avons utilisé **Vadersentiment** pour analyser *les sentiments* de chacun des tweets récoltés. Cela permet de calculer la valeur entre -1 (sentiment très négatif) et 1 (sentiment très positif) de chaque mot, et même émoticône. Le logiciel fait ensuite une moyenne et donc attribue aux *tweets* une valeur, qui elle se situe également entre – 1 et 1. Si le *tweet* est neutre, il a une valeur égale à zéro.

Nous avons utilisé la boite à outils **NLTK (Natural Language Toolkit**) qui permet la création de programmes pour l'analyse de texte. Que nous avons associé à l'opération de « **tokenization »** qui permet dans ce cas-ci le découpage des *tweets* en mots pour mieux en dégager les tendances.

Finalement, nous avons utilisé **Sublime Text** pour écrire nos scripts dans **Python** et enfin **Terminal**. Parce que nous étions, de toute évidence, plus à l'aise et expérimentés avec ce langage-là.

Le travail d'équipe et la répartition des tâches

Notre équipe était composée de 7 personnes : trois personnes de Polytechnique Montréal et quatre personnes de l'UQAM. En raison du nombre élevé, la répartition et les rencontres étaient plutôt difficiles à organiser (vous pouvez voir le plan dans les fichiers du bilan de notre première rencontre). Nous avons donc créé on groupe Facebook pour pouvoir échanger le maximum d'informations entre nos rencontres physiques.--> pas la meilleure façon pour expliquer une erreur dans un script ou l'améliorer ☺

La répartition des tâches s'est faite naturellement avec les forces et les horaires de chacun. L'écriture du premier script a été un travail collectif -tout le monde à essayer de l'écrire : bien que Jean-Batiste a réussi à trouver la formule gagnante ! Je tiens à souligner son acharnement et son travail ici parce qu'il le mérite largement !

- La répartition s'est faite comme ceci : Jean-Batiste à fait rouler le script pour les médias et les journalistes et moi pour les Hashtags.

- Nous avons donc fait une demande de clefs fournies par Twitter (pour accéder à son API) et écrit individuellement un script personnalisé correspondant à nos recherches respectives. (une ligne du script diffère)

- Pour ce qui est des #Hashtags, après moult recherches, j'ai sélectionné ceux qui étaient les plus utilisés, tout en essayant d'équilibrer au maximum entre ceux positifs et ceux négatifs. (Tâche difficile puisqu'il n'existe pas beaucoup de Hashtags félicitant le travail des journalistes et des médias !)

Pour ce qui est du deuxième script (*NLTK* avec la fonction *word_tokanize*) moi et Jean Batiste travaillons toujours dessus, mais sans grand succès pour le moment. Il y aura deux personnes qui analyseront les médias et les journalistes français, deux pour ceux québécois.

Pour ce qui est des entrevues (nous en avons quatre). Mélissa en a fait trois et Théo une. Je me suis occupée de deux verbatim (ils sont dans la liste des fichiers).

- Seul bémol : Les entrevues ont été faites sans montrer la visualisation (pas concluante) donc nous n'avons pas de commentaires à ce sujet de la part des experts…

Problèmes rencontrés

Un problème mineur d'entente avec les élèves de polytechnique. Nos niveaux de programmation étant très différents, nous avons ressenti une frustration de leur part. Pas très ouverts à nous aider à faire fonctionner un script qu'ils avaient écrit, mais trop complexe pour nous. Nous avons finalement réussi à l'adapter à notre script initial. (Leur contribution nous a permis de récupérer les scripts au-delà des 500 derniers *tweets*, c'est à dire sept jours en arrière. De plus, nous avons pu rajouter le « lecteur de sentiment » à notre script -avec votre aide!)

Autre problème : nous n'avions pas les mêmes échéanciers et donc ils étaient très stressés à l'idée de ne pas recevoir les données assez vite pour créer la visualisation. Somme toute, nous avons réussi de notre côté et donc cela a apaisé les tensions.

Nous avons également rencontré un problème avec le logiciel utilisé: *Vadersentiment.* Comme expliqué ci-dessus, il devait calculer « le sentiment positif ou négatif » de chaque *tweet* et lui donner une valeur entre -1 et 1. Lorsque nous avons vu la visualisation, nous avons constaté que la marge d'erreur était beaucoup trop grande et que cela ne permettait pas de dégager des tendances.

- La première raison qui explique cet échec : nos données en Français ont dû être traduites en anglais pour être analysées par le logiciel - et donc beaucoup de problème de mauvaises traductions.-
- La deuxième raison : nous avons remarqué que plusieurs *tweets* parlaient négativement ou positivement du contenu de l'information et non du média (ou du journaliste) qui l'avait transmise ce qui fausse largement les résultats.
- Nous avons donc été contraints de recommencer à zéro et écrire un nouveau script permettant d'analyser de manière plus qualitative nos données (script 2).

Liste des fichiers dans mon « repo »:

- Trois fichiers de type .csv qui correspondent à toutes nos données récoltées.
- Deux *verbatim* d'entrevues
- Trois scripts Python (écrit dans *Sublime Text*) .py
- Un script PANDA écrit par les élèves de polytechnique dans le carnet Jupyter .Ipynb
- Plan de rencontre des étudiants de polytechnique (Design.pdf)
- Lien vers la visualisation : https://gadiben.github.io/DatavizAlter/index.html?fbclid=IwAR3cYxt6L02GbLSmI_U5mdfHj4mwNLH_fH34p6Hh_xkgHLf0wpt7vwLo0G0

Les liens vers tutoriels et ressources en ligne :

Premier script :

- http://socialmedia-class.org/twittertutorial.html

Deuxième script :

- https://www.youtube.com/watch?v=Xi52tx6phRU
- https://www.nltk.org/
- https://marcobonzanini.com/2015/03/09/mining-twitter-data-with-python-part-2/