INF8111 – Fouille de données
Automne 2019 – Groupe 01

Travail pratique 2

Extraction et analyse d'une base de données de tweets : Méthodes de clustering appliquées à des tweets sur l'actualité

Présenté à
Théo Moins

Par

Boisvenue, Étienne	1798942	etienne.boisvenue@polymtl.ca
Le Page, Pierre-Étienne	1689986	pierre-etienne.le-page@polytml.ca
Wong, Jade	1738089	jade.wong@polymtl.ca

Polytechnique Montréal
11 novembre 2019

INTRODUCTION SOMMAIRE

Lancée en 2006, la plateforme Twitter compte aujourd'hui plus de 330 millions de membres s'exprimant notamment sur des sujets d'actualité [1]. L'objectif de ce travail était de regrouper les tweets d'un jeu de données selon le sujet d'actualité qu'ils couvrent. La méthodologie exploratoire employée consiste à générer des résultats simples, puis à faire varier les hyperparamètres et à évaluer la qualité des amas d'une sélection d'expériences. Enfin, la distribution temporelle d'amas obtenue concorde avec des événements d'actualités des dernières années.

1. PRÉSENTATION DU JEU DE DONNÉES

Le jeu de données utilisé est celui de *Littman et al* intitulé « News Outlet Tweet Ids » [2]. Le plus grand nombre de tweets hydratables en 3 heures l'ont été (900 000 tweets, dont 803 951 étaient en anglais). Un échantillon aléatoire de 100 000 tweets qui présentait les mêmes distributions d'attributs que la population a été utilisé pour réduire le temps de calcul. Le **Tableau 1** présente les attributs retenus et les justifications correspondantes.

Tableau 1 Attributs retenus et justifications

Attribut	Justification
created_at	Permet l'analyse de l'évolution des amas dans le temps
id	Identifie de manière unique le tweet traité.
text	Fournit le contenu textuel du tweet qui sera traité.
user['name']	Permet de visualiser les contributeurs principaux d'un amas.
user['location']	Permet d'identifier les sujets d'actualité propre à une région géographique.
followers_count	Permet d'identifier les personnes plus influentes et/ou prolifiques.
in_reply_to_status_id	Agit comme métrique d'évaluation de la qualité des amas en étant une forme d'agglomération organique.
lang	Permet de limiter l'analyse aux tweets rédigés en anglais

2. PREPROCESSING

Premièrement, pour éviter un biais, les enregistrements ont été mélangés aléatoirement et seuls les tweets en anglais ont été retenus pour ne traiter qu'une seule langue. Deuxièmement, les attributs ont été transformés vers des formats adéquats (e.g. *created_at* en *numpy.datetime64*). Troisièmement, les URLs, *hashtags*, nombres et ponctuations ont été retirés, car ils ne contribuent pas à l'analyse de la langue anglaise. Le **Tableau 2** présente les autres étapes et les justifications correspondantes.

Tableau 2 Étapes 4 à 7 du preprocessing

Description	Justification
Tokenization	Séparation du texte en séquence de jetons pour isoler chaque mot du texte.
Troncature (*stemming*)	Réduction le vocabulaire en combinant les mots d'une même famille sous un seul jeton.
Retrait des *stop words*	Contribution négligeable à l'analyse et réduction de la taille du dictionnaire.
Bag-of-words avec la pondération TF-IDF avec et sans bigram	Identification des jetons plus rares et susceptibles d'être évocateurs à l'analyse du thème des amas.

3. MÉTHODOLOGIE, RÉSULTATS ET ANALYSES

Avant tout, les graines (*seeds*) de chacune des librairies utilisant des méthodes stochastiques (i.e. *random, numpy, sklearn*) ont été fixées pour assurer la rigueur et la reproductibilité des

résultats. Une méthodologie *data-driven* plutôt que *hypothesis-driven* a été employée. Afin d'illustrer le cheminement des explorations, cette section identifie consécutivement la méthode employée, les résultats obtenus et leur analyse. Tous les résultats présentés dans cette section ont été obtenue avec l'échantillon décrit à la section 1.

3.1 Résultats préliminaires

Tout d'abord, les 10 000 premières observations de l'échantillon, réduites à 100 dimensions par la décomposition en valeurs singulières (SVD), ont été visualisées à l'aide de t-SNE (**Figure 1** ci-contre), selon la distance euclidienne. Il aurait été intéressant de réduire à un plus grand nombre de composantes pour expliquer plus de variance, mais avec le temps réservé pour cette étude et ses objectifs, une réduction à 100 dimensions avec une variance expliquée de 9,89 % permettant d'exécuter la SVD en 12,2 s a été privilégiée (détails en annexe). Ces paramètres ont tout de même permis une visualisation révélatrice des données. Il est possible d'observer que les données forment naturellement des amas, justifiant la pertinence d'agglomérer ces tweets. Leur forme ronde indique que K-moyennes est un algorithme approprié pour l'agglomération. Malgré tout, les méthodes d'agglomération hiérarchique de types *single-linkage* et *complete-linkage* ont été explorées. La première s'avère peu efficace pour la mesure de similarité entre deux tweets, sauf pour la détection de quelques tweets produits par des machines (*bot*) (voir l'annexe pour plus de détails). La seconde n'assure pas la convergence de la méthode du coude, au contraire de K-moyennes [3][4]. Ainsi, pour le reste de l'analyse, la méthode de K-moyennes a été privilégiée.

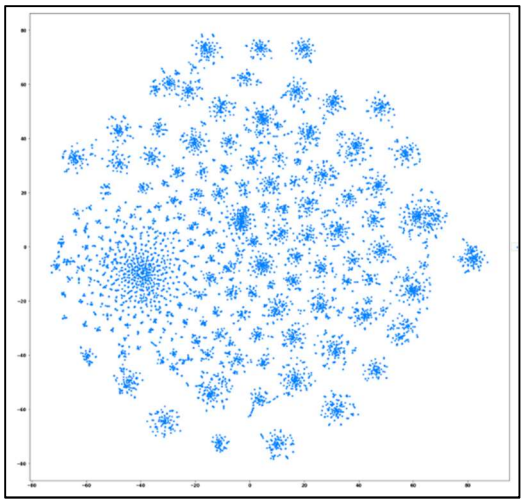

Figure 1 t-SNE résultant d'un échantillon aléatoire de 10 000 éléments de la matrice TF-IDF

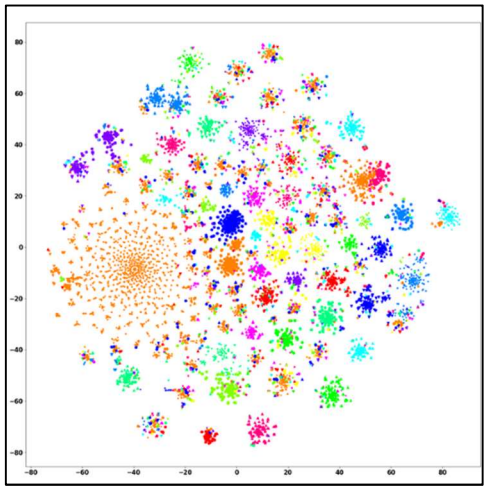

Figure 3 Nombre de tweets par auteur dans l'amas *trump* et la proportion de leurs tweets qui mentionnent *trump*.

Figure 2 Visualisation des 50 amas de K-moyennes selon t-SNE avec les hyperparamètres *perplexity*=34 et *early_exageration*=3.

L'algorithme de K-moyennes a tout d'abord été appliqué sur la matrice de TF-IDF pour agréger les tweets en 2 amas distincts (K = 2) avec un maximum de 100 itérations et une seule configuration d'initialisation aléatoire. Les sujets *Donald Trump*, *45e président des États-Unis* et *Autres tweet* ont été identifiés. 99,8% des tweets de l'amas Donald Trump contenaient le jeton *trump*, comparé à 0,1% pour le second amas. Les cinq tweets de l'amas ne contenant pas le jeton *trump* étaient tout de même connexes à la présidence américaine (voir l'annexe). Ensuite, les tweets de l'amas *trump* ont été agrégés par auteur, tel qu'illustré par la **Figure 3**. Cette figure présente le nombre de tweets mentionnant *trump* par auteur ainsi que la proportion de leurs tweets respectifs qui font partie de l'amas *trump*. Cette liste

Auteur	Fréquence	Proportion (%)
The Hill	220	45.27
POLITICO	109	32.06
The Independent	99	9.93
Washington Examiner	88	33.08
CNN Politics	81	40.50

d'auteur pourrait servir de recommandation à un utilisateur voulant rester informé au sujet du président.

3.2 Résultats approfondis

La section précédente présente les résultats obtenus en appliquant naïvement l'algorithme K-moyennes pour K = 2, sans évaluer la robustesse et la qualité des amas générés. Cette section présente les améliorations apportées à la méthodologie et les *insights* résultants.
En observant la distribution dans le temps de la base de données (l'échantillon de 100 000 tweets), il est flagrant qu'elle suit une loi exponentielle, ce qui concorde avec l'augmentation des dernières années du nombre d'utilisateurs actifs sur Twitter [1].

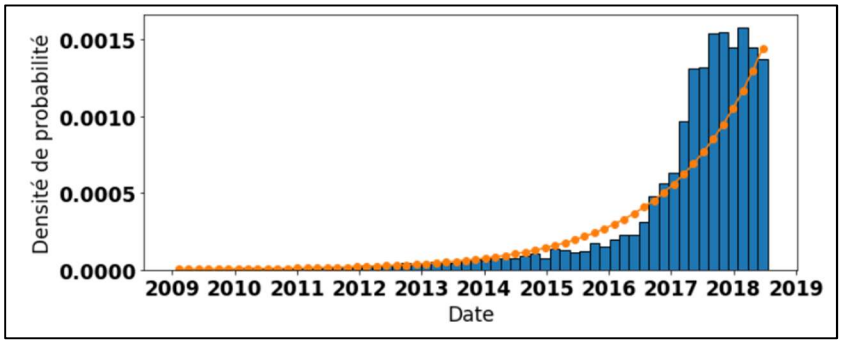

Figure 4 Distribution de la base de données dans le temps ainsi qu'une loi exponentielle (en orange) modélisant la tendance selon un coefficient de détermination de 85%.

Cette tendance a été retirée de tous les résultats subséquents. Cette transformation met en valeur les déviations, spécifiques au sujet d'un amas, de la tendance générale. La **Figure 4** présente la distribution temporelle des tweets dans l'amas *trump*. Les dates de son élection et de son entrée en fonction (assermentation) sont identifiées en rouge et en vert, respectivement. Le graphique du bas présente les résultats normalisés selon la tendance exponentielle générale. Enfin, il est possible de remarquer un maximum local le jour des élections et une grande réactivité face aux tweets controversés de Trump durant sa présidence.

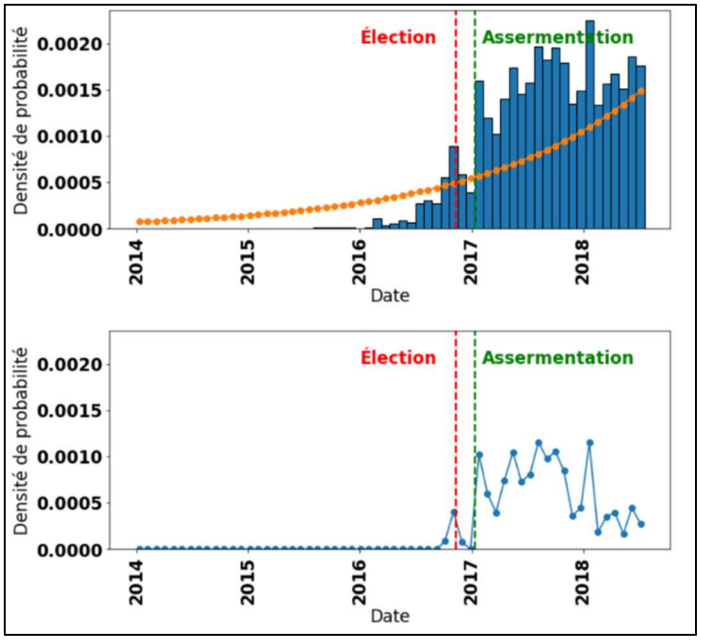

Figure 5 Distribution temporelle des tweets dont le thème principal est Donald Trump et identification d'événements politiques corrélés.

Le **Tableau 3** présente les mots de vocabulaires les plus communs de la base de données. Ils sont tous reliés au champ lexical de l'actualité et des États-Unis. Ils aident très peu à la détermination du thème spécifique d'un amas.
Afin de déterminer le nombre optimal d'amas, une courbe d'inertie en fonction de K a été générée, telle que présentée à la **Figure 6**. Toutes les valeurs pour K = 2 à 70 ont été testée. Puis, pour observer le comportement de l'inertie lorsque K est très grand, quelques valeurs supplémentaires ont été testées. Les données ont été réduites à 100 composantes, selon une SVD avant de performer l'agglomération.

Le point d'inflexion de la courbe (le coude) se situe autour de K = 80. Les analyses subséquentes se concentrent sur les expériences K = 20, K = 50 (voir **Figure 2**) et K = 350. Cette sélection permet d'étudier le contenu d'amas issus des différents régimes de la courbe d'inertie. Pour chaque amas de chaque expérience, la sous-matrice terme-document correspondante a été extraite. Parmi les jetons présents dans au moins 10% des tweets de l'amas, seuls ceux dont l'IDF est supérieur à 5 ont été conservés. Ce filtrage permet de faire ressortir les jetons relativement rares qui sont présents dans plusieurs tweets de l'amas. Le seuil IDF ≥ 5 a été établi en observant la distribution de l'IDF de tous les termes du vocabulaire. Les jetons très fréquents, donc peu spécifiques à l'amas, possèdent un IDF d'environ 4, alors que les jetons présents dans uniquement un tweet possèdent un IDF de 11,5, soit ln(100 000/1). Malgré le fait que le jeton *trump* possède un IDF inférieur au seuil fixé, il est quand même possible de retrouver les amas le concernant puisque le champ lexical entourant ses fonctions possède des IDF plus élevés (i.e. *presid*=5,12 et *donald*=6,04). Le **Tableau 4** présente un exemple de l'information extraite avec la méthodologie pour quelques amas. Les thèmes des amas ont été confirmés en lisant manuellement des tweets de l'amas. Les thèmes sont récurrents à travers les expériences. Plus K augmente, plus les amas sont spécifiques. L'expérience K = 20 possède un amas centré sur *La politique américaine* contenant 1531 tweets. Ces tweets sont distribués à travers plusieurs thèmes plus spécifiques dans l'expérience K = 350, tels le *Plan proposé par les républicains pour réformer le système de santé* et les *Nouvelles lois adoptées par le Sénat*, contenant respectivement 32 et 221 tweets. Avec DBSCAN, ces 32 tweets auraient probablement été considérés comme du bruit, alors qu'ils couvrent un même thème spécifique et pertinent. Certains amas sont formés autour d'un seul jeton commun, par exemple l'amas 105 de l'expérience K = 350 regroupant les tweets contenant le jeton *fire*.

Tableau 3 Unigrams et bigrams du vocabulaire dont l'IDF est le plus élevé.

Ngram	(IDF)	
new	3,98	
trump	4,05	
say	4,22	
us	4,46	
year	4,54	
high school	5,92	
donald trump	6,12	
presid trump	6,33	

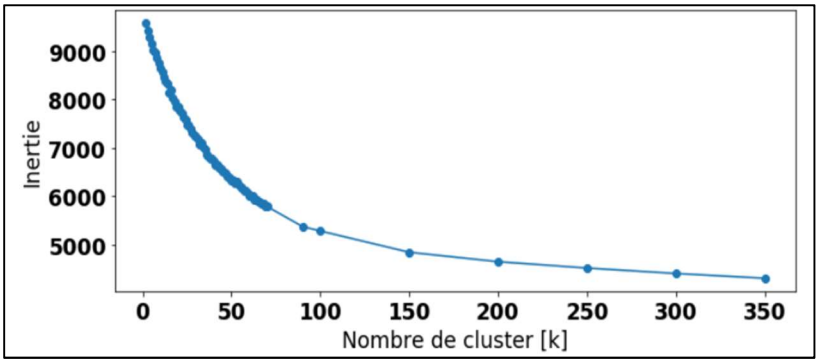

Figure 6 Inertie (somme intra-amas des distances au carré) en fonction du nombre d'amas. Données générées avec K-moyennes, selon 20 initialisations différentes et 100 itérations maximum par initialisation.

Tableau 4 Identification des thèmes de certains amas et illustration de la méthodologie employée.

K	Numéro d'amas	Nombre de tweets	Thème	IDF	Jeton	% de tweets dans l'amas contenant le jeton
50	48	571	Corée du Nord	5,63	north	0,88
				6,30	korea	0,71
				5,85	south	0,17
350	22	32	Plan de soins de santé proposé par les républicains	5,20	plan	1,00
				5,88	health	0,88
				6,30	care	0,38
				6,27	gop	0,22

Effectivement, cet amas regroupe des tweets à propos des feux de forêt en Californie de 2017 et des tweets mentionnant le renvoi d'employés (*to fire someone*). Similairement, le jeton *bill* a regroupé des tweets à propos de nouvelles lois (*legislative bills*) et le fondateur de *Microsoft*, *Bill Gates*. La **Figure 7** suivante illustre la distribution dans le temps de l'amas 105 de l'expérience K = 350. La date étiquetée comme étant *Feux forêt Californie* est le 8 juillet 2017, soit la date d'intensité maximale des feux [5].
L'écart le plus flagrant de la tendance exponentielle se trouve à la date d'intensité maximale des feux de forêt en Californie. Le thème de l'amas a donc été correctement identifié. D'autres distributions temporelles sont disponibles en annexe.

CONCLUSION

Pour conclure, les tweets ont été regroupés par thème d'actualité en utilisant K-moyennes. Pour valider les thèmes des amas, l'évolution de la densité de probabilité a été comparée avec les dates d'événements connexes au sujet. La tendance exponentielle dans le temps de la population a été retirée des densités de probabilités. La faiblesse principale de la méthodologie est la perte de l'aspect syntaxique des jetons. Pour la surmonter, une étape de *Word embedding* aurait pu être réalisée. Ainsi, les jetons auraient été représentés par un vecteur traduisant leur sémantique pour les contextualiser et ainsi distinguer l'utilisation du jeton *bill* dans les contextes de *Bill Gates* et de *legislative bill*. Pour estimer le point d'inflexion, il aurait été plus rigoureux de l'obtenir avec la librairie *scikit-yb* qui possède une implémentation de la méthode du coude, que visuellement [6]. Comme avancements possibles, il serait intéressant d'appliquer des méthodes d'apprentissage semi-supervisé, soit imposer le regroupement des tweets selon l'attribut *in_reply_to_status_id*.

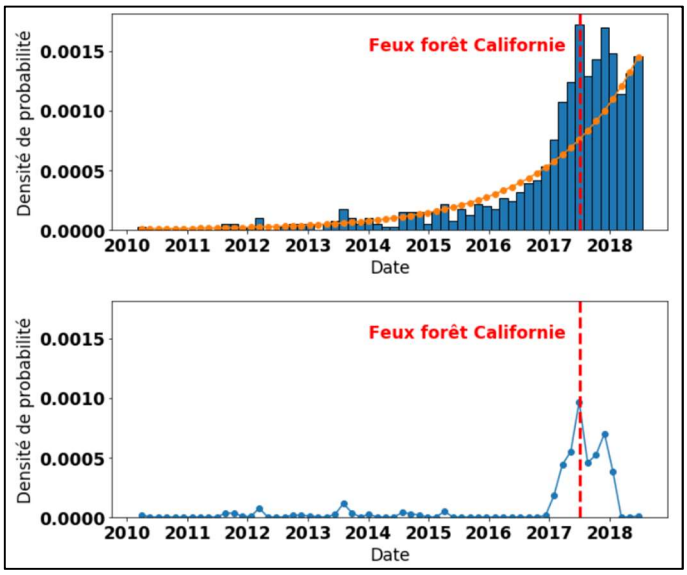

Figure 7 Distribution des tweets de l'amas dont le thème principal est *fire*.

RÉFÉRENCES

[1] Clement, J. (14 août 2019). Twitter: number of monthly active users 2010-2019. Tiré de https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[2] Littman, Justin; Wrubel, Laura; Kerchner, Daniel; Bromberg Gaber, Yonah, 2017, "News Outlet Tweet Ids", https://doi.org/10.7910/DVN/2FIFLH, Harvard Dataverse, V3, UNF:6:I38WJ5vqwDky1fkEOeexvQ== [fileUNF]

[3] Bottou, L. & Bengio, Y. (2007). Convergence Properties of the K-moyennes Algorithms. Tiré de http://www.iro.umontreal.ca/~lisa/pointeurs/kmeans-nips7.pdf

[4] SciPy Hierarchical Clustering and Dendrogram Tutorial (10 novembre 2019) | Jörn's Blog. (2019). Tiré de https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/?fbclid=IwAR1iX2sYKvTTOYJH-XY6Rr-BU4s-UX6fF6sYwfLps-Zm4-qNNtaZnvYqWDQ

[5] Wikipedia. (10 novembre 2019). 2017 California wildfires. Tiré de https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[6] Yellowbrick. (2019). Elbow Method, revision dd795b49. Tiré de https://www.scikit-yb.org/en/latest/api/cluster/elbow.html