Projet de *computer science for big data*

INTRODUCTION

Présentation des données

Les données exploitées sont en partie disponibles sur le site Kaggle. Ce jeu de données agrège les tweets relatifs au réchauffement climatique collectés entre le 27 avril 2015 et le 21 février 2018. Au total, 43 943 tweets ont été annotés. Chaque tweet est étiqueté en toute indépendance par trois réviseurs. Ce jeu de données contient uniquement les tweets sur lesquels les trois réviseurs se sont mis d'accord (les autres ont été écartés). Les tweets sont classifié selon l'un des quatre groupes suivant[1] :

- Ne croit pas au réchauffement climatique `{sentiment: -1}` ;
- Sans opinion `{sentiment: 0}` ;
- Croit au réchauffement climatique `{sentiment: 1}` ;
- Relaie des faits d'actualité à propos du réchauffement climatique `{sentiment: 2}`.

Cependant, une manipulation spécifique est nécessaire afin de récolter plus d'information sur ces tweets. Ce second jeu de données est constitué de données structurées au format JSONLines avec de nombreuses informations (heure de création, profil d'utilisateur, hashtags...). Puisque le changement climatique est un sujet lié aux événements en temps réel et à l'influence de l'environnement social, ce second jeu de données offre plus de détails sur le contexte. Le logiciel Hydrator permet de récupérer la totalité de ces informations. En effet, ce logiciel utilise l'API de Twitter permettant ainsi de télécharger autant de tweets (et tout leur contenu) que l'on souhaite, palliant ainsi le problème de la limitation de 20 tweets téléchargeable « à la main » par jour depuis Twitter directement. Une clé de l'API de n'importe quel compte Twitter est nécessaire, ce qui l'autorise à télécharger les données des tweets en récupérant un grand nombre d'identifiants.

Problématique

L'objectif de cette étude est, dans un premier temps, d'utiliser MongoDB pour trouver certaines « tendances » sociales qui pourraient affecter la prise de conscience des gens sur le réchauffement climatique ; et dans un deuxième temps, d'utiliser Spark afin de transformer le contenu des tweets et de préparer une analyse plus avancée.

[1] Le code couleur utilisé fait référence à la légende des FIGURES 3 et 4.

Outils utilisés

En plus la jointure de deux sources de données, abordée dans la partie Présentation des données, d'autres outils sont utilisés pour le traitement des données.

Pour la construction du jeu de données, deux sources de données ont été regroupées pour obtenir le plus d'information possible sur ces tweets classifiés selon les quatre groupes possibles. Dans un premier temps, il faut installer MongoDB sur Windows et créer une base de données vierge par l'intermédiaire de l'invite de commandes Windows (voir Annexe). Ensuite, les données sont importées dans cette nouvelle base sous forme de collection à partir d'un fichier au format JSON. Cette base et collection crées, il faut ensuite s'y connecter depuis un script Python à l'aide des fonctions du module `pymongo`.

La première partie de l'étude est une analyse exploratoire du contenu des tweets afin d'y desceller des caractéristiques pouvant être déterminantes dans la classification. Pour cette partie, différents modules sont utilisés pour valoriser certains traitements de données et les résultats présentés dans la partie Analyse exploratoire avec MongoDB.

La seconde partie est consacrée à l'utilisation de Spark à travers le module `pyspark`, afin de construire le pipeline évolutif de nettoyage et de transformation de texte pour d'autres apprentissages ultérieurs. Les résultats sont présentés dans la partie exploration et nettoyage de texte avec Apache Spark.

ANALYSE EXPLORATOIRE AVEC MONGODB

Ordre de grandeur des tweets

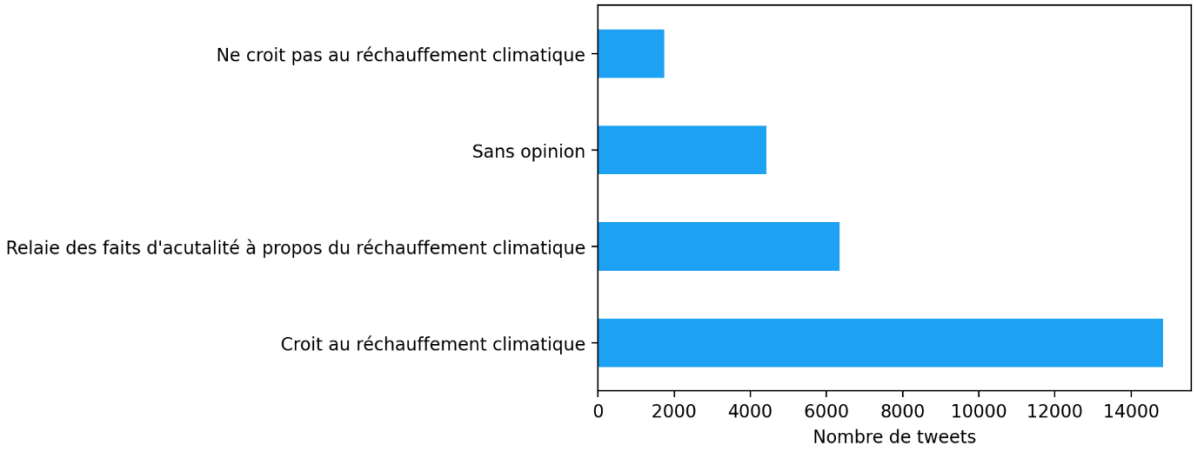


FIGURE 1 : Nombre de tweets par opinion sur le réchauffement climatique

Il y a un déséquilibre flagrant entre les différents groupes, avec une surreprésentation des tweets illustrant la croyance de l'internaute au réchauffement climatique.

Classement des *hashtags* les plus utilisés

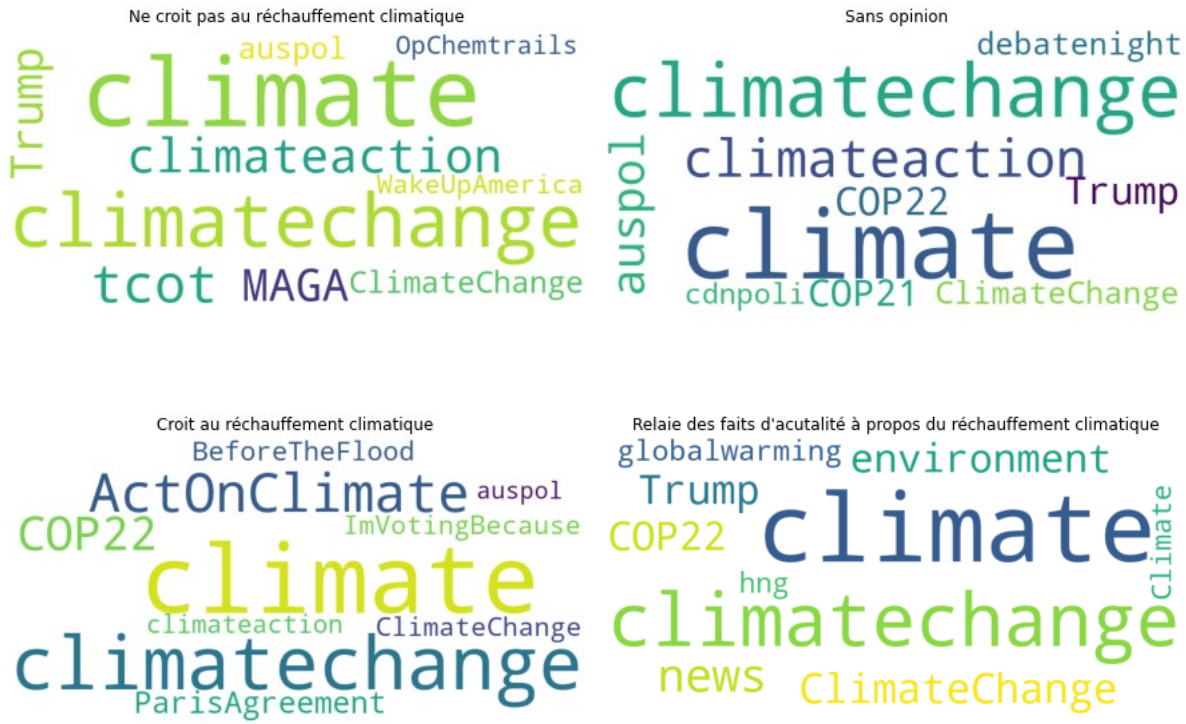


FIGURE 2 : Top 10 des # les plus utilisés par opinion sur le réchauffement climatique

Dans une première lecture globale des quatre nuages de mots, on repère tout d'abord des similitudes dans les *hashtags* les plus utilisés (ceux qui paraissent les plus gros). On y retrouve : *#climate, #climatechange, #climateaction*. Cependant, le nuage qui illustre l'opinion perçue des tweets comme « Ne croit pas au réchauffement climatique » se distingue des trois autres au niveau de certains *hashtags*. En effet, on retrouve dans les trois autres nuages, des *hashtags* de rassemblement autour de la question du réchauffement climatique avec *#COP22*.

Dans une seconde lecture plus spécifique à chaque nuage, on voit que la distinction de l'opinion « Ne croit pas au réchauffement climatique » reste forte avec des *hashtags* relevant du complotisme politique ou climatosceptique. *#OpChemtrails* fait référence à un blog de théories complotistes sur le climat, quand *#MAGA* (*make America great* again) et *#WakeUpAmercia* symbolisent le militantisme pour Donald Trump. À l'inverse, les trois autres opinions font plutôt référence à des conférences ou mouvements politiques en faveur d'une prise de conscience et d'actions sur le réchauffement climatique avec #COP21, #ParisAgreement ou #ImVotingBecause.

Cette visualisation des *hashtags* par nuages de mots est donc assez juste dans la représentation des opinions.

Séries temporelles sur le nombre de tweets

Évolution mensuelle des tweets (vision globale)

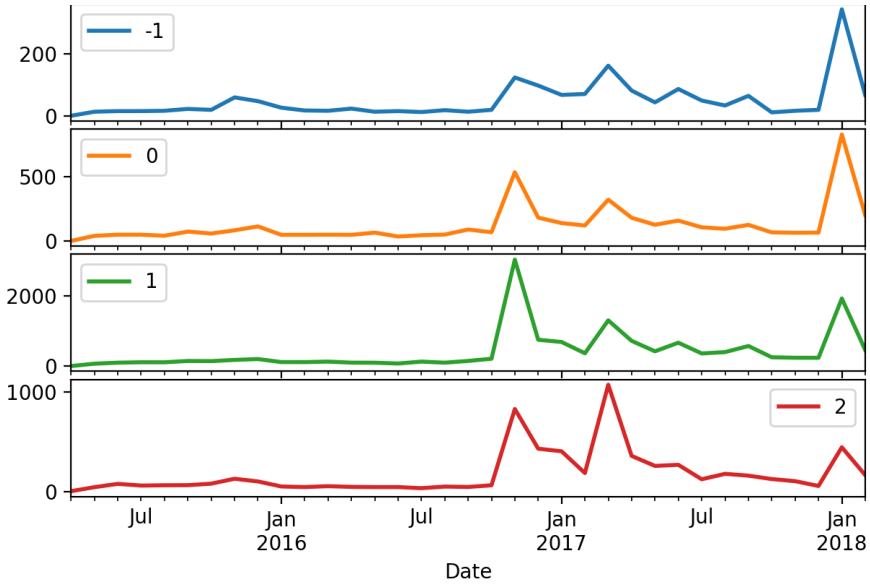


FIGURE 3 : Nombre de tweets par "sentiment" sur le réchauffement climatique

D'après la FIGURE 3, le nombre de tweets par mois a traversé des période d'engouement sur la question du réchauffement climatique en fin 2016 et début 2017, ainsi qu'en hiver 2018. La hausse du nombre de tweets est proportionnelle selon l'opinion bien qu'elle soit légèrement moins marquée pour le « sentiment » -1, illustrant la non-croyance au réchauffement climatique.

Évolution journalière des tweets (vision détaillée)

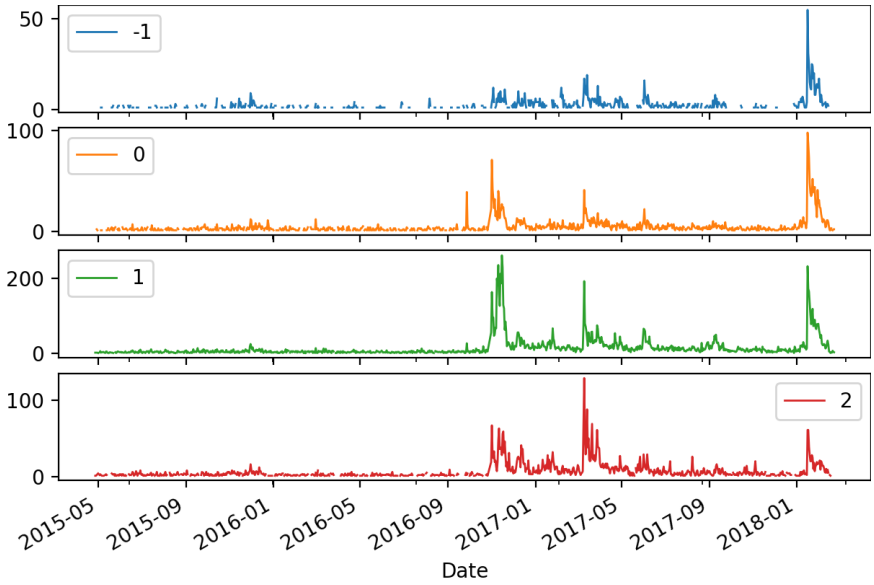


FIGURE 4 : Nombre de tweets par "sentiment" sur le réchauffement climatique

La FIGURE 4, permet de voir en détail l'évolution de ces hausses du nombre de tweets, interprétés comme un engouement des utilisateurs de Twitter autour du réchauffement climatique et des enjeux environnementaux.

Ces hausses apparaissent à des périodes clés. En effet, l'accord de Paris sur le climat, signé en avril 2016, entre en vigueur le 4 novembre 2016[2]. De plus, la COP22 a lieu en novembre 2016[3]. Ces évènements peut être une explication de cette première hausse.

ANALYSE AVANCEE AVEC SPARK

Traitement des données avec **pymongo**

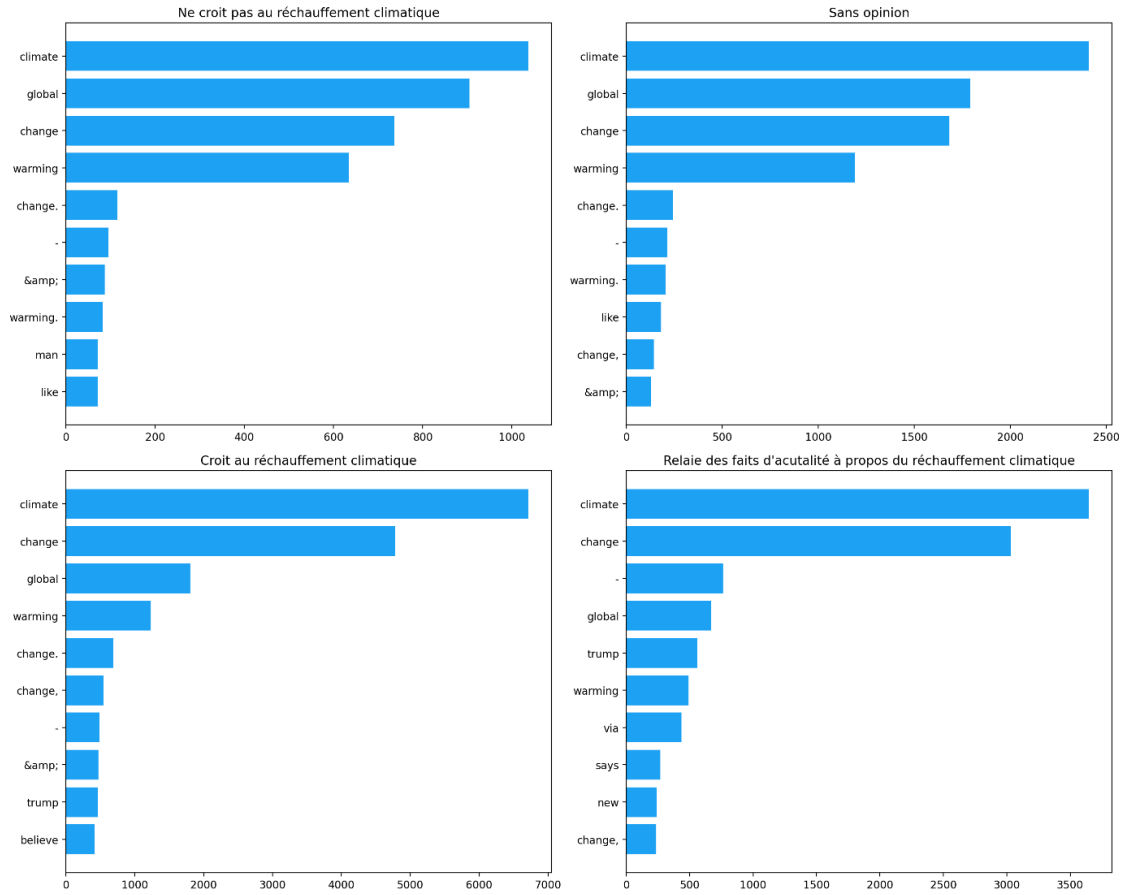
Tout d'abord, l'environnement Spark est hébergé localement. Une session Spark est créée dans le but de charger notre ensemble de données comme un *dataframe* Spark. Cet objet peut être considéré comme une table distribuée sur un cluster et possède une fonctionnalité similaire aux *dataframes* de R et Pandas. En préparation des données textuelles pour *Natural Language Processing* (NLP), *"tweet message*" est alors :

- *tokenisé* (décomposer le texte en mots). Il donne une structure à un texte précédemment non structuré.
- Les mots courants qui apparaîtront probablement dans tout texte (stop words) sont supprimés car ils ne nous apprennent pas grand-chose sur nos données.
- Les *hashtags*, liens web, *tags* dans les *tokens* sont également supprimés car ils n'ont pas beaucoup de sens.
- Toute ligne vide est supprimée.

Classement des mots clés les plus employés

Enfin, la méthode map/reduce est utilisée pour calculer la fréquence des mots pour chaque sentiment. Les 10 mots les plus utilisés sont illustrés dans la partie Classement des mots clés les plus employés ci-après.

[2] https://fr.wikipedia.org/wiki/Accord_de_Paris_sur_le_climat
[3] https://fr.wikipedia.org/wiki/Conf%C3%A9rence_de_Marrakech_de_2016_sur_les_changements_climatiques



Les mots les plus utilisés ne font pas beaucoup de différence entre chaque classe de sentiments (i.e. chaque opinion). Ce sont les mots les plus révélateurs du sujet du réchauffement climatique comme *climate*, *change*, *global*, *warming*, etc… À l'exception du tweet d'actualité, des mots plus typiques apparaissent comme *trump*, *says*, *news*, etc…

Cependant, le nettoyage du texte n'est pas parfait car il y a encore des caractères inutiles "-", et des *tokens* différents pour des mots ayant la même définition. Il est possible d'aller plus loin en utilisant la méthode de "*Stemming*" et "*Lemmatization*" ans le NLP.

ANNEXE

Annexe 1

[guide de présentation du repo Git]