Introduction :

Pour extraire des informations de tweets (entre janvier et mars 2021) liés au vaccin COVID où les opinions sont très peu structurées, hétérogènes et sont soit positives ou négatives, ou neutres et identifier les facteurs déterminants du changement de sentiments Pour explorer les conversations et les « sujets » abstraits qui se produisent dans les tweets collectés à l'aide de la modélisation des sujets et de l'analyse de texte soutenue par des événements révolutionnaires dans la chronologie.

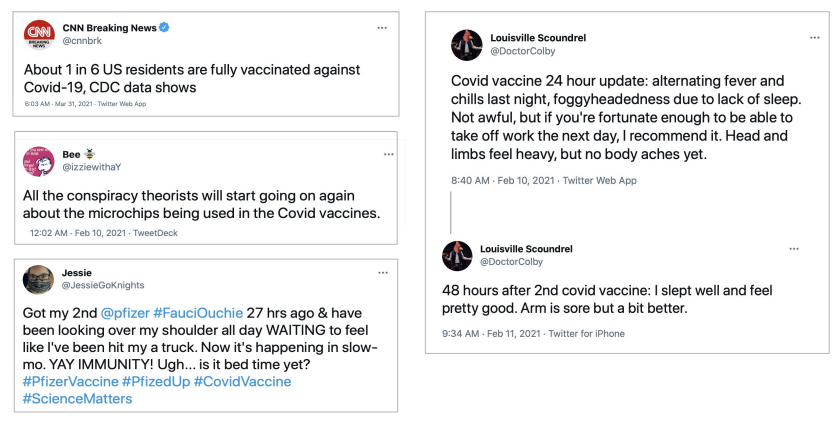
La vaccination est une pierre angulaire de la prévention des maladies infectieuses transmissibles; Cependant, les vaccins ont traditionnellement rencontré la peur et l'hésitation du public, et les vaccins COVID-19 ne font pas exception. Il a été démontré que l'utilisation des médias sociaux joue un rôle dans la faible acceptation des vaccins.

Objectif :

Le but de cette étude est d'identifier les sujets et les sentiments dans la discussion publique sur le vaccin COVID-19 sur les réseaux sociaux et de discerner les changements saillants dans les sujets et les sentiments au fil du temps afin de mieux comprendre les perceptions, les préoccupations et les émotions du public qui peuvent influencer l'atteinte des objectifs d'immunité collective.

Nous utilisons pyspark , qui est l'API Python pour Spark. Ici, nous utilisons Spark Structured Streaming, qui est un moteur de traitement de flux construit sur le moteur Spark SQL et c'est pourquoi nous importons le module pyspark.sql. Nous importons ses classes ; SparkSession pour créer une session de flux, une fonction et des types pour rendre disponible une liste de fonctions intégrées et de types de données. Nous utilisons également textblob pour la classification du texte des tweets.

COVID Vaccine Conversations on Twitter :



1. Extaction des données :

Les tweets ont été téléchargés à partir d'un ensemble de données de discussion à grande échelle sur Twitter COVID-19 du 11 mars 2020, le jour où l'Organisation mondiale de la santé a déclaré COVID-19 pandémie, jusqu'au 31 janvier 2021. Nous avons utilisé python pour nettoyer les tweets et conserver tweets contenant les mots-clés vaccination, vaccinations, vaccine, vaccines, immunisation, vacciner et vacciné. L'ensemble de données final inclus dans l'analyse consistait en 1 499 421 tweets uniques provenant de 583 499 utilisateurs différents. Nous avons utilisé R pour effectuer une allocation Dirichlet latente pour la modélisation de sujets ainsi que l'analyse des sentiments et des émotions à l'aide du lexique des émotions du Conseil national de recherches du Canada.

snscrape est une bibliothèque qui permet à quiconque de récupérer des tweets sans avoir besoin de clés API personnelles. Il peut renvoyer des milliers de tweets en quelques secondes et dispose de puissants outils de recherche qui permettent des recherches hautement personnalisables.

1. Data Cleaning

CRÉATION DES MÉTADONNÉES ET CLASSIFICATION

a. Removed punctuation using remove\_punct function with library re

b. Removed URLs and emojis in Tokenization using library re

c. Removed stopwords using nltk

d. Lemmatization of Tweets using nltk.WordNetLemmatizer()

Pour analyser le sentiment d’un tweet, nous pouvons alors extraire ce qui nous semble être important comme par exemple :

Le nombre de lien “http(s)” ou “www.”.

Le nombre d’adresse mail. (le nombre d’adresse mail cité dans le tweet pourrait avoir une importance concernant le sérieux du tweet.)

Le nombre d’hashtag.

Le nombre de citation d’utilisateur.

Le nombre de lettre en majuscule

Le nombre de chaine lettre en majuscule.

Le nombre de point d’exclamation.

Le nombre de chaîne de point d’exclamation.

Le nombre de point d’interrogation.

Le nombre de chaîne de point d’interrogation.

Le nombre de point ect (…).

Removed punctuation, stopwords,  lemmatization,

Les variables explicatives étant désormais prêtes à être utilisé, nous n’avons plus qu’à faire tourner notre modèle de régression logistique afin de classifier les sentiments positifs ou négatifs des tweets.

1. Analyse des sentiments :

Nous utilisons également textblob pour la classification du texte des tweets.

TextBlob est un module NLP sur Python utilisé pour l’analyse de sentiment. La fonction de TextBlob qui nous intéresse permet pour un texte donné de déterminer le ton du texte et le sentiment de la personne qui l’a écrit.

Pour chaque tweet nous aurons une métrique qui donne la polarité de ce tweet.

1. Régression logistique

Désormais, nous allons commencer notre travail de classification de sentiments des tweets à l’aide d’une régression logistique.

Pour ce faire, l’idée est comme d’habitude de diviser notre jeu de données en un échantillon d’entraînement (80%) dans lequel nous allons apprendre les paramètres du modèle et un échantillon test (20%) dans lequel nous allons les tester.

Néanmoins, les variables explicatives étant des données textuelles (les tweets), nous allons au préalable créer un nouveau variable explicative numérique en lien avec les tweets (sentiment) afin de pouvoir ensuite prédire les labels.

[Analyze The Sentiment of Tweets From Twitter Data and Tweepy in Python | Earth Data Science - Earth Lab](https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/analyze-tweet-sentiment-in-python/)

[Introduction au NLP avec Python pour l'analyse de sentiments (larevueia.fr)](https://larevueia.fr/nlp-avec-python-analyse-de-sentiments-sur-twitter/)

[NLP Twitter - Analyse de Sentiment - DataScientest](https://datascientest.com/nlp-twitter-analyse-de-sentiment)

[Text Data Cleaning - tweets analysis | Kaggle](https://www.kaggle.com/ragnisah/text-data-cleaning-tweets-analysis)

[Final Project Report.pdf](file:///C:\Users\HPR\Downloads\Final%20Project%20Report.pdf)

[Sentiment Analysis using TextBlob · Twitter Sentiment Analysis Visualization Tutorial (gitbooks.io)](https://wafawaheedas.gitbooks.io/twitter-sentiment-analysis-visualization-tutorial/content/sentiment-analysis-using-textblob.html)

Lizd9o textblob sentiment

[Sentiment Analysis of tweets, Wordclouds, Textblob | Kaggle](https://www.kaggle.com/shailaja4247/sentiment-analysis-of-tweets-wordclouds-textblob)

[Sentiment-Analysis-using-Pyspark-on-Multi-Social-Media-Data/pyspark\_logistic\_reg\_twitter.ipynb at master · chaithanya21/Sentiment-Analysis-using-Pyspark-on-Multi-Social-Media-Data (github.com)](https://github.com/chaithanya21/Sentiment-Analysis-using-Pyspark-on-Multi-Social-Media-Data/blob/master/pyspark_logistic_reg_twitter.ipynb)

[pyspark-twitter-sentimental-analysis/Sentimental\_Analysis.ipynb at master · apurva-modi/pyspark-twitter-sentimental-analysis (github.com)](https://github.com/apurva-modi/pyspark-twitter-sentimental-analysis/blob/master/Sentimental_Analysis.ipynb)

[Twitter Sentiment PySpark | Kaggle](https://www.kaggle.com/fatihgulsen/twitter-sentiment-pyspark)