Etude des corrélations dans les élections sur Twitter

Objectifs

Nous souhaitons analyser les Tweets à propos de Joe Biden et Donald Trump pour déterminer le sentiment au niveau des États pour chaque candidat, puis comparer cela aux résultats de vote afin de déterminer s'il existe une relation entre le sentiment Twitter et les tendances de vote. Ensuite, nous voulons afficher ces données sur des cartes.

Nous avons pour objectif d'explorer si les données de Twitter pourraient être utilisées pour prédire le résultat des élections et les limitations d'utilisation de ce jeu de données. En visualisant les tendances de vote et les opinions des utilisateurs de Twitter, nous espérons obtenir un aperçu des opinions du public américain lors de l'élection présidentielle de 2020 et savoir si ces opinions se traduisent dans les résultats réels des votes. Nous utiliserons l'analyse des sentiments et l'analyse de corrélation pour nous aider à atteindre ces conclusions.

Objectifs Spécifiques :
- Objectif 1 : Analyse des sentiments sur les Tweets de l'élection présidentielle de 2020
- Objectif 2 : Analyse des relations entre les résultats d'analyse des sentiments et les tendances de vote
- Objectif 3 : Visualisation spatiale des sentiments sur Twitter et des tendances de vote

Trois sections Objectif 1, Objectif 2 et Objectif 3 seront dédiées à la préparation de données, de fonctions et de graphes. La section Analyse sera dédiée à l'analyse des différentes graphes et statistiques crées dans les sections précédentes.

Les Données

Les données de Tweets (tweets.zip) se composent de deux fichiers .csv, l'un des Tweets avec #DonaldTrump et l'autre des Tweets contenant #JoeBiden, nommés 'hashtag_donaldtrump.csv' et 'hashtag_joebiden.csv' respectivement. Les Tweets ont été collectés en utilisant l'API Twitter open source via `statuses_lookup` et `snsscrape` pour rechercher des mots-clés. pour défendre vos travaux

Les données contiennent un total de 21 colonnes, comprenant des informations sur le Tweet lui-même, le compte de l'utilisateur, les données géographiques extraites de l'emplacement de l'utilisateur et la date et l'heure auxquelles les données du Tweet ont été extraites via l'API. Le fichier 'hashtag_donaldtrump.csv' contient 958 580 Tweets uniques, et le fichier

'hashtag_joebiden.csv' contient 768 423 Tweets uniques. Par conséquent, plus de 1,5 million de Tweets ont été extraits au total en utilisant l'API Twitter.

Les colonnes sont les suivantes :
- `created_at` : Date et heure de création du Tweet
- `tweet_id` : ID unique du Tweet
- `tweet` : Texte complet du Tweet
- `likes` : Nombre de likes
- `retweet_count` : Nombre de Retweets
- `source` : Outil utilisé pour publier le Tweet
- `user_id` : ID de l'utilisateur créateur du Tweet
- `user_name` : Nom d'utilisateur du créateur du Tweet
- `user_screen_name` : Nom d'écran du créateur du Tweet
- `user_description` : Description de soi par le créateur du Tweet
- `user_join_date` : Date d'inscription du créateur du Tweet
- `user_followers_count` : Nombre de followers de l'utilisateur
- `user_location` : Lieu donné sur le profil du créateur du Tweet
- `lat` : Latitude extraite de user_location
- `long` : Longitude extraite de user_location
- `city` : Ville extraite de user_location
- `country` : Pays extrait de user_location
- `state` : État extrait de user_location
- `state_code` : Code de l'État extrait de user_location
- `collected_at` : Date et heure auxquelles les données du Tweet ont été extraites de Twitter

Toutes ces données ne seront pas utilisées dans l'analyse requise pour atteindre nos objectifs, et toute donnée qui sera nettoyée ou supprimée sera clairement indiquée.

Nous utiliserons également les résultats finaux des votes de l'élection présidentielle de 2020 aux États-Unis (ap_votes.csv). Ces données proviennent de l'Associated Press - l'AP suit les comptes des votes aux élections américaines depuis 1848 et ses données sont largement considérées comme précises. Les variables de ce jeu de données sont :

- state : État auquel le compte des votes correspond
- state_abr : Abréviation à deux lettres du nom de l'État
- trump_pct : Pourcentage des votes remportés par Donald Trump
- biden_pct: Pourcentage des votes remportés par Joe Biden
- trump_win : Variable binaire indiquant si Donald Trump a remporté le vote dans un État
- biden_win : Variable binaire indiquant si Joe Biden a remporté le vote dans un État

Le jeu de données final que nous utiliserons, exclusivement pour le troisième objectif, est un fichier de formes (shapefile) des frontières des États américains (y compris des territoires non constitués comme Guam et les Samoa américaines), téléchargé à partir du Bureau du Recensement des États-Unis (URL : https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html ou https://data.texas.gov/dataset/US-States-Cartographic-Boundary-Map/r5bg-j3b6/data).

Les colonnes incluent :
- `STATEFP`, `STATENS`, `AFFGEOID`, `GEOID` : Divers numéros d'identification géographique
- `STUSPS` : Abréviation à deux lettres du nom de l'État
- `NAME` : État auquel la frontière géographique correspond
- `LSAD` : Code de description de l'aire légale/statistique (tous désignés 00 au niveau de l'État)
- `ALAND` : Aire, en mètres carrés, des portions terrestres de l'État
- `AWATER` : Aire, en mètres carrés, des portions aquatiques de l'État
- `geometry` : Polygone (ou multipolygone) représentant la frontière de l'État

Une limitation réside dans le fait que `user_location` est autodéclaré et peut donc ne pas être précis au moment du Tweet. Puisque Twitter ne met pas automatiquement à jour les lieux des utilisateurs, les Tweets provenant d'utilisateurs ayant déménagé entre les États sans mettre à jour leur emplacement Twitter ou ayant mal rapporté leur lieu de résidence (ce que les utilisateurs peuvent choisir de faire pour diverses raisons) pourraient être résumés sous l'État incorrect lorsque l'emplacement a été extrait. Cela ne devrait probablement pas avoir un impact majeur sur les résultats de notre analyse.

Objectif 1 – Analyse des sentiments

Dans cette section, il est demandé d'importer des fichiers CSV (tweets.zip) contenant des données sur les Tweets concernant l'élection présidentielle américaine. En particulier ceux avec "#donaldtrump" et "#joebiden" dans le Tweet. Les données seront filtrées pour ne retenir que les Tweets en provenance des États-Unis, rédigés en anglais, et avant la fermeture des bureaux de vote. Les données seront ensuite nettoyées et les mots vides (stop words) seront supprimés de tous les Tweets afin de mettre davantage l'accent sur les mots qui définissent le sens du Tweet lors de l'analyse de sentiment. La

subjectivité et la polarité de chaque Tweet seront spécifiées, et des nuages de mots seront tracés pour comparer et identifier les similarités entre les deux ensembles de données.

Préparation des données

1- En utilisant la bibliothèque Pandas, chargez les deux fichiers csv contenant les tweets sur Biden et Trump. Chaque candidat est associé à un dataframe (structure de données de Pandas)
2- Les colonnes 'tweet_id', 'collected_at', 'user_description' et 'collected_at' ne sont pas pertinents à notre étude. Supprimez-les.
3- Filtrer les Tweets pour ne garder que ceux publiés aux États-Unis les Tweets publiés avant la fermeture des derniers bureaux de vote (3 novembre 2020 à 20h).
4- Supprimez tous les Tweets non rédigés en anglais. Utilisez la fonction detect de la bibliothèque langdetect pour déterminer la langue du Tweet.
5- Nettoyer les tweets en éliminant les mentions (@), les hashtags (#), les liens (hyperliens) et les retweets (RT), et en convertissant l'ensemble du texte en minuscules. Pour ce faire, utiliser des expressions régulières qui doivent être appliquées au dataframe (structure de données de Pandas) contenant les tweets.

Un mot vide en anglais (*stop word*) est un mot très courant qui n'apporte pas de sens significatif au texte, comme "the," "is," "at," ou "and." Ces mots sont souvent supprimés dans le traitement automatique du langage pour se concentrer sur les mots porteurs de sens.

6- Utilisez la bibliothèque nltk pour supprimer les mots vides de la langue anglais sans omettre d'inclure les mots suivants "donaldtrump", "trump", "donald", "biden", "joe", "joebiden", "amp", "president", "vote", "voting", "election". Ces mots n'ajoutent pas de pertinence dans notre analyse de sentiments.

Analyse des sentiments

L'analyse des sentiments repose sur l'examen de deux critères principaux : la subjectivité et la polarité. La subjectivité mesure dans quelle mesure un texte reflète des opinions, des émotions ou des jugements personnels, par opposition à des faits objectifs.

- **Texte subjectif** : exprime des sentiments ou des points de vue personnels (ex. : "Je trouve ce film incroyable.").

- **Texte objectif** : repose sur des faits vérifiables, sans opinion (ex. : "Le film dure deux heures.").

En traitement automatique du langage, la subjectivité permet de distinguer les informations factuelles des contenus émotionnels ou opinionnels. La polarité évalue l'orientation émotionnelle d'un texte, en déterminant s'il est positif, négatif ou neutre. Elle est généralement mesurée sur une échelle de -1 à 1 :

- -1 : Texte très négatif (ex. : "Je déteste ce produit.").
- 0 : Texte neutre, sans opinion marquée (ex. : "Le produit est rouge.").
- 1 : Texte très positif (ex. : "Ce produit est excellent.").

Ainsi, la polarité permet de quantifier l'attitude émotionnelle d'un texte.

7- Utilisez la bibliothèque TextBlob pour déterminer la subjectivité et la polarité de chaque tweet. Créez trois nouvelles colonnes : subjectivity, polarity et pol_state. Un tweet a une valeur “negative” si la polarité est strictement inférieure à 0, “positive” si la polarité est strictement supérieure à 0, sinon “Neutral”.

Un **nuage de mots** (*word cloud*) est une représentation visuelle des mots les plus fréquemment utilisés dans un texte ou un ensemble de textes :

- **Taille des mots** : Plus un mot est fréquent, plus sa taille dans le nuage est grande.
- **Disposition** : Les mots sont disposés de manière artistique, souvent dans une forme ou un contour spécifique.
- **Couleurs** : Différentes couleurs peuvent être utilisées pour distinguer les mots ou rendre le nuage plus attrayant.

8- Créez une fonction, appelée getWordCloud, qui prend en paramètre un dataframe et génère le nuage de mots des tweets associés. Cela nous permettra d’analyser les mots les plus souvent tweetés associés à chaque hashtag (#donaldtrump ou #joebiden).

9- Créez une fonction, getInfoPolarity, qui :
   a. Calcule le pourcentage de Tweets **positifs**, **négatifs** et **neutres** à partir d'un ensemble de données contenant une colonne indiquant l'analyse des Tweets (avec des valeurs comme "Positive," "Negative," et "Neutral").
   b. Affiche ces pourcentages sous forme de texte, précisant les valeurs pour chaque type de polarité.
   c. Génère un **diagramme circulaire (pie chart)** , via matplotlib ou seaborn, pour visualiser ces pourcentages, avec :
      i. Des tranches colorées pour chaque catégorie.
      ii. Les pourcentages affichés sur les tranches.
   d. Un titre personnalisé basé sur le nom du candidat

Objectif 2 – Analyse des relations entre les sentiments des tweets et les tendances de vote

A partir des données issues de l'objectif 1, une analyse de corrélation sera effectuée pour examiner la relation entre les sentiments des Tweets et les résultats de vote dans chaque État.

10- Calculer la polarité moyenne de tous les Tweets dans chaque État pour chaque candidat.
11- Crée une fonction qui fusionne les données de vote (ap_votes.csv) avec les données des Tweets, de la consigne précédente, en utilisant les codes des États (colonne state_abr pour les données de vote et state_code pour les Tweets). Appliquez cette fonction sur chaque candidat pour aboutir à deux dataframes trump_merged et biden_merged.
12- Créez une fonction
    a. qui prend en paramètre deux variables (colonnes) où l'une est indépendante (x) et l'autre dépendante (y) ;
    b. qui retourne une variable dépendante estimée ;
    c. La fonction d'estimation basée les moindres carrées doit être développée.
13- Appliquer cette fonction entre le pourcentage de votes (trump_pct, biden_pct) et la polarité pour chaque candidat.
14- Créez une fonction, appelée scatterplot, qui projette les données en nuage de points (scatterplot) entre le pourcentage des votes et la polarité des tweets et trace la droite estimée avec la régression linéaire.
15- Créez trois fonctions:
    a. **WinnerPolarity** : Cette fonction prend en entrée un DataFrame et le nom d'un candidat. Elle renvoie un tableau des polarités des Tweets dans les États où le candidat a gagné le vote.
    b. **LoserPolarity** : Cette fonction prend également un DataFrame et le nom d'un candidat, mais elle renvoie un tableau des polarités des Tweets dans les États où le candidat a perdu le vote.
    c. **GetPolarities** : Cette fonction calcule la moyenne ou l'écart type des polarités des Tweets pour un candidat, en fonction de l'option choisie.
16- Créez une fonction, showBarChart, qui trace un graphique en barres des catégories en fonction de la polarité des Tweets. Cette fonction prend en entrée les catégories que nous souhaitons tracer (dans notre cas, "Trump Win" et "Biden Win") par rapport à la polarité moyenne des Tweets dans ces états, avec des barres d'erreur définies par l'écart type.

Objectif 3 – Création de cartes choroplèthes des données Twitter et de vote

Nous allons élaborer une fonction pour créer des cartes de nos données. Ce seront des cartes choroplèthes des états américains (à l'exception de l'Alaska et d'Hawaï), mettant en évidence les différences dans les sentiments et les tendances de vote à travers les États-Unis. La visualisation spatiale de nos données permettra d'interpréter les motifs spatiaux entre les deux ensembles de données et constitue une méthode efficace pour communiquer les données au grand public. Lors de l'élection présidentielle de 2020, les cartes ont été largement utilisées par les médias pour informer le public sur l'état des affaires, des prédictions et des sondages de

sortie jusqu'aux résultats finaux. Elles sont également utilisées par des stations de diffusion comme la BBC pour les élections au Royaume-Uni.

17- Chargez la carte des Etats-Unis munie de ses frontières inter-états (fichier shapefile de l'archive cb_2018_us_state_500k.zip) en utilisant Geopandas. Supprimez les états d'Alaska et Hawai.
18- Calculez la subjectivité moyenne de tous les Tweets dans chaque état.
19- Fusionnez les moyennes de polarité et de subjectivité par état, ainsi que la part des voix du candidat avec le fichier de formes (shapefile) pour associer les valeurs de ces champs aux limites géographiques de l'état. Cela crée l'ensemble de données final à utiliser dans cet objectif.
20- Créez la fonction qui génère une carte choroplèthe pour les données relatives aux états des États-Unis.
21- Créez la fonction qui permet de tracer deux cartes sur le même axe, avec une barre de couleur partagée. Cela permet de comparer la même variable entre Donald Trump et Joe Biden.

Les cartes produites par les fonctions de cette section seront tracées dans la section Analyse, avec des descriptions des paramètres utilisés et l'interprétation des résultats.

Analyse

Dans cette section, il est demandé de rédiger un rapport reportant les résultats obtenus dans chaque objectif (graphes, statistiques et modèles de régression) ainsi que l'interprétation associée. Dans ce rapport, on doit aussi inclure une discussion qui confronte les différents résultats pour aboutir à une analyse globale sur la relation entre le sentiment sur Twitter et le vote réel.

Une conclusion est demandée qui inclut

- les opérations qui ont été effectuées
- les limites rencontrées
- les perspectives potentielles à cette étude

Rendu et soutenance

Délivrables attendus pour le 6/01/2026 :

1. Un rapport détaillé couvrant la section *Analyse* ;

2. Le code Python exécutable (fichiers Python ou notebooks).

Modalités de la soutenance :

- Une présentation orale d'une durée de 10 à 15 minutes pour défendre vos travaux, suivie de 10 minutes de questions.

- Le plan suggéré pour la présentation est le suivant :
    - **Contexte et problématiques**
    - **Technologies et outils utilisés**
    - **Concepts et outils mathématiques appliqués**
    - **Analyse de la relation entre le sentiment sur Twitter et le vote réel**, illustrée par les graphiques du projet
    - **Limites et difficultés rencontrées**
    - **Conclusion et perspectives**