Partie I:

Téléchargement et Pré-traitement des tweets

Ce partie est réalisée par la programme get_tweets en utilisant l'API Tweeter et le module GOT (get-old-tweets) pour télécharger les Tweets des comptes qui nous intéressent, décrits dans le fichers en format csv ou json. Les données que j'ai récupré sont des tweets de Donald Trump en 2020 et en 2016. Pour mieux analyser l'impact sur l'élection, j'ai aussi récupéré les données qui parle de Donald Trump après l'élection de 2020.

Partie II:

Analyse des Tweets et réalisation du graphe de statistique

Avant toute analyse il faut explorer les données, pour comprendre quel type d'analyse faire selon la question qu'on se pose. On utiliser des simples visualisations comme des histograms et tirer les premières simples conclusions et contrôler la cohérence avec le sujet

Cette partie est réalisé par la programme Pre_processing.py.

Avec les function décrit pour nettoyer les données et analyser les données en une dimention.

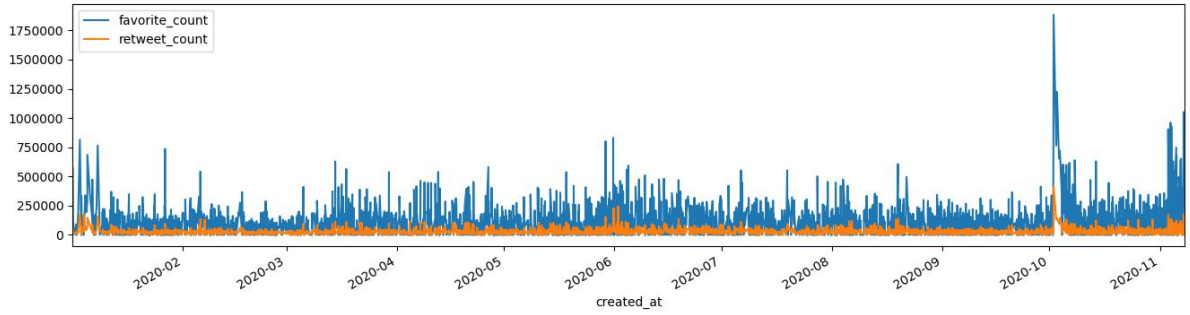
Les outils concerné:

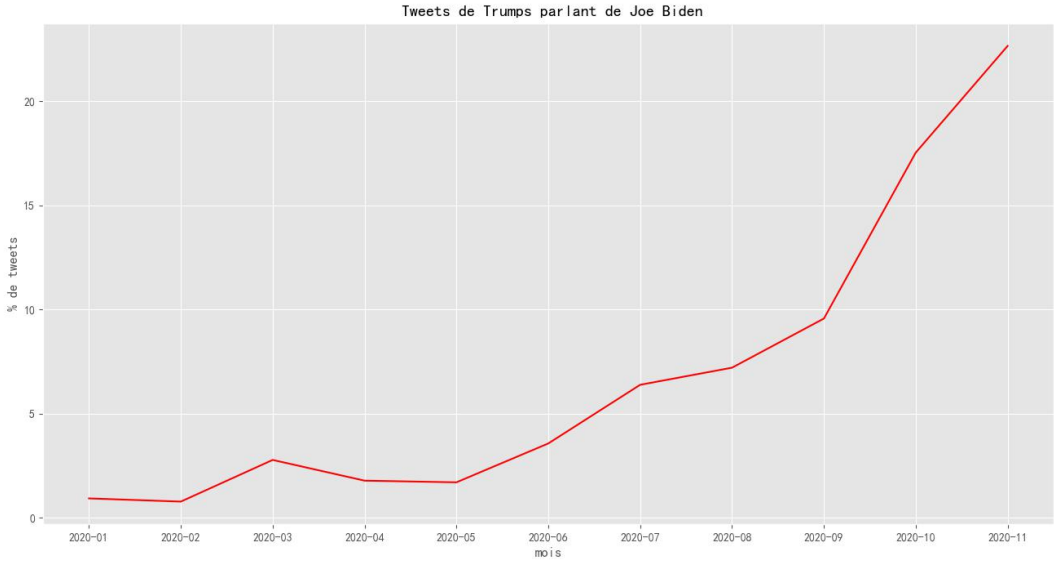
1. Les packages de python pour l'analyse stastique:

Numpy , pandas

2. Les packages de python pour la visualisation :
3. Matplotib,char_studio.

- Les résulats de prétraitement :





Ci-desus deux graphiques sont lié avec l'évoluation des tweets.
Le première graphique parle de nombre des favorite tweets et des retweets qui évolue au file de temps, on peut observer qu'il y a une pic en mois de l'election. Et le deuxième graphique montre le pourcentage de tweets qui parle de son rival Biden, il a beaucoup augmenté en moi de l'election. On peut voir que le pourcentage qui parle de Biden a augmenté de 5% en Janvier à environ 25% en Novembre. On peut voir que l'utilisation de réseau social a influcencé la participation politique pour les citoyens et les candidats profitent les reseaux sociaux comme Tweet pour influencer les citoyens et renforcer ses opinions politique.

Partie III:

Lemmatisation des Tweets et réalisation du graphe de sentiment

Après l'exploration de données, pour mieux connaître notre données, il faut approfondir l'analyse. Considérant les caractéristique des tweets :

- La longueur des tweets : la longueur maximale d'un tweet est de 140 caractères, c'est la spécifité de ce réseau social ainsi les

utilisateurs en tendance à exprimer leurs sentiments en utilisant des mots qui ont une valeur sémantique importante

contiennent des liens, ces derniers sont représentés comme une seule unité(token).

- Le hachtags : l'hashtag est beaucoup utilisé sur tweeter, c'est le symbole clé qui permet la diffusion de l'informations, suivie de quelques mots qui encapsule une information.
- L'arobase « @ » : ce symbole est disponible dans l'intégralité des tweets, il permet de déterminer le nom ou le pseudo de l'utilisateur. Donc on décide de extraire les hastages , l'arobase et les mots fréquents pour mieux connaître la tendance de tweets de Trumps.
- Http: Ce simbole est utilisé pour partager les liens qui permet de maximiser la propagation de certains information.

Cette partie est réalisée par ce fichier EBA-explorer.py qui contient la function de nettoyer des données, l'analyse de sentiments et le textes et la visualisation de données.

Les modules pour nettoyer les données:

TweetTokenizer, stopwords , String

Les packages pour analyser les sentiments et calculer les fréquence de mots:

FreqDist, TextBlob , TweetTokenizer

Les packages pour la visualisation :

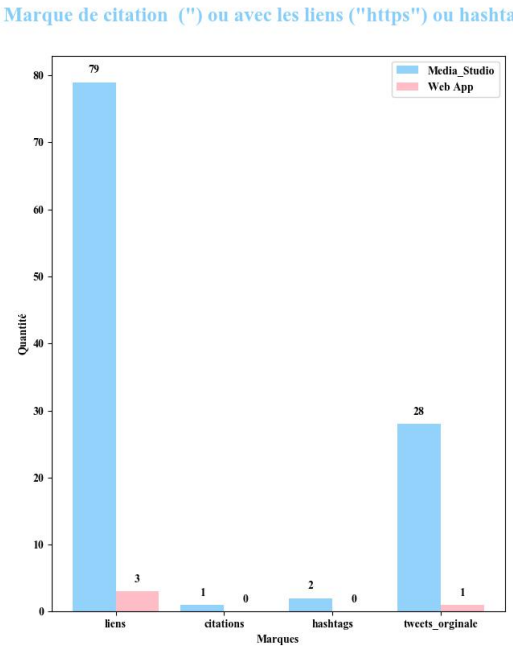
Matplotlib , wordcloud, seaborn

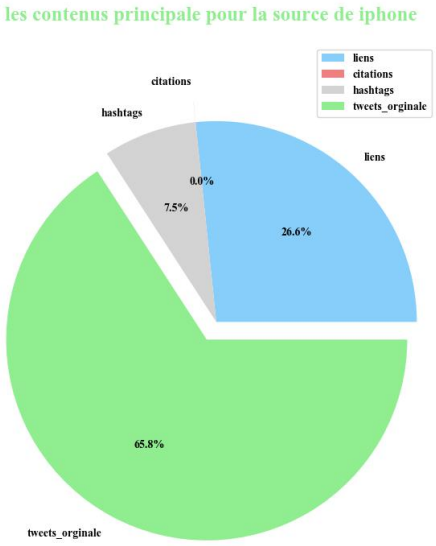
Le programme EBA-explorer.py récupère le fichier picke généré par get_tweeet.py, et filtre les mots en générant une stopList contenant,et enlever les ponctuation et les mots inutiles.

Il boucle ensuite sur tous les tweets pour calculer la fréquence de chacun de leurs mots.En calculant la polarity (<0, >0 =0) pour classifier les mot négative, positive ou neutre.

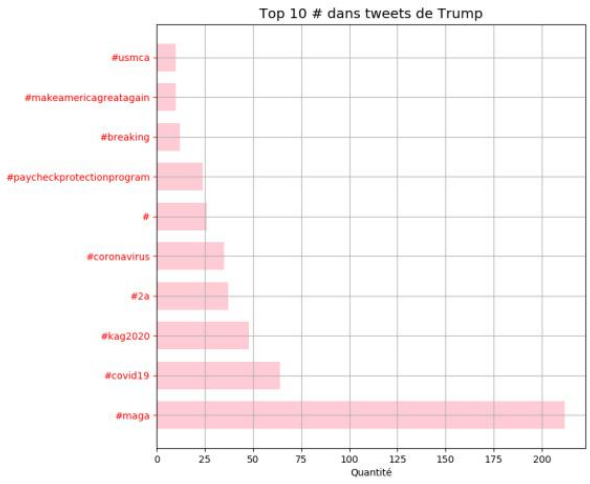
- Les résulats de l'analyse de text:

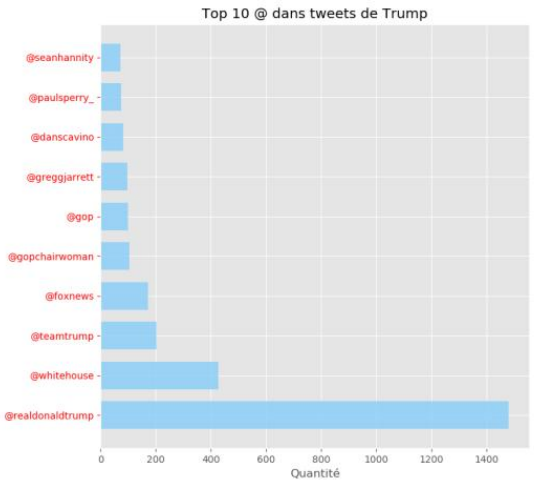
1. Voici c'est le graphique qui analyse le source et les type de contenus de ces tweets en identifiant les simboles comme '# ' et 'http' pour extraire les données. On peut remarquer que la plus part de tweets qui viennent de iphone, ce sont des tweets originals (seulment des textes, pas de hastag et liens), par contre pour le source Media studio, il présent les liens officiels (comme les paroles de présitent).





2. Ci-desous est la graphique qui calcule les nombre de top 10 hastags et top 10 arobase.

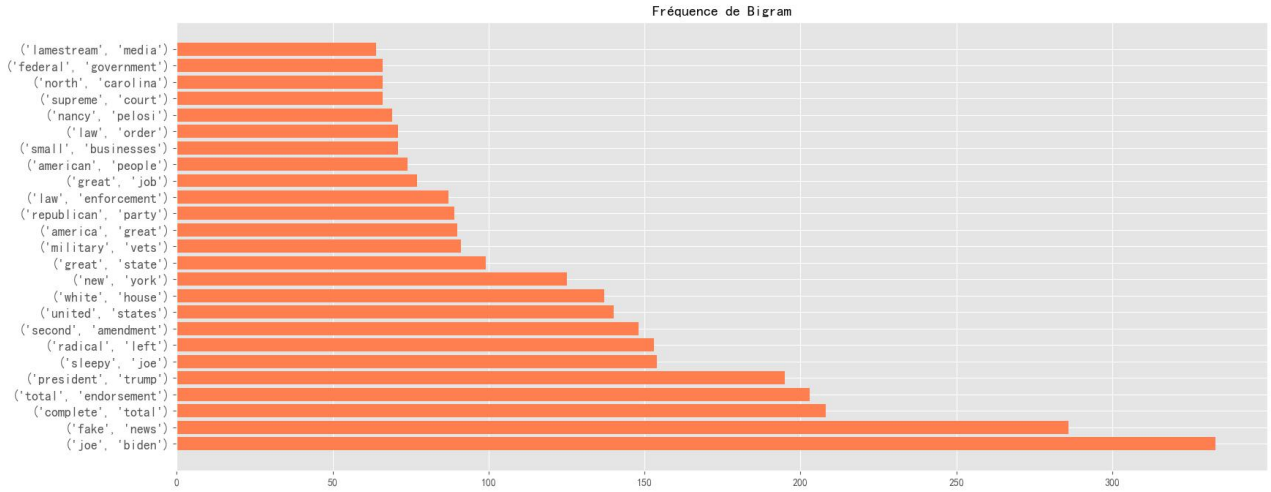




1. On peut voir dans les top 10 hastags et top 10 @, ce sont des tweets lié avec “maga” (make american great again) , covid, et lui-même, ça correspond aussi avec les foyers de conflits dans la société, et ça reflet aussi un partie de son

stratégie d'influencer les citoyens par réseaux sociaux.

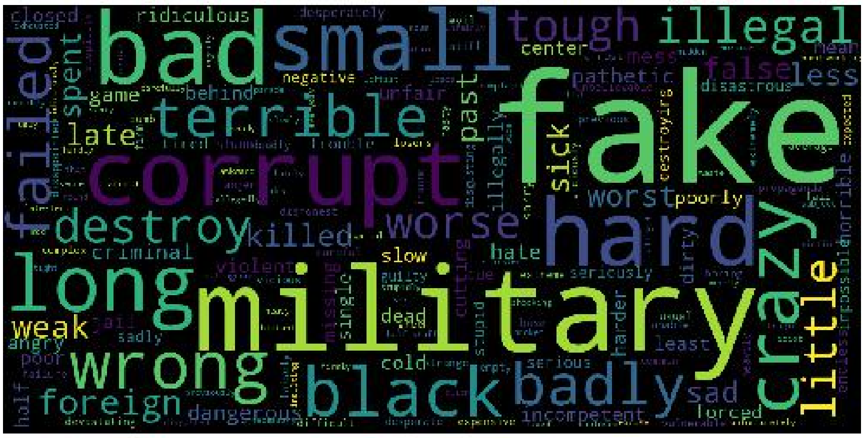
Pour avoir connaissance de la distribution des contenus de ces tweets, on a nettoyé des données pour avoir les texts brut et calculer les fréquence de mot Bigram(les mots qui apparaient ensemble). Le résulat correspond avec les analyses avant, Trump parle beaucoup de son rival et les confits dans la société. Ce qui aussi reflet le fait que comment les candidats manipulent et influence les opinions des citoyens.

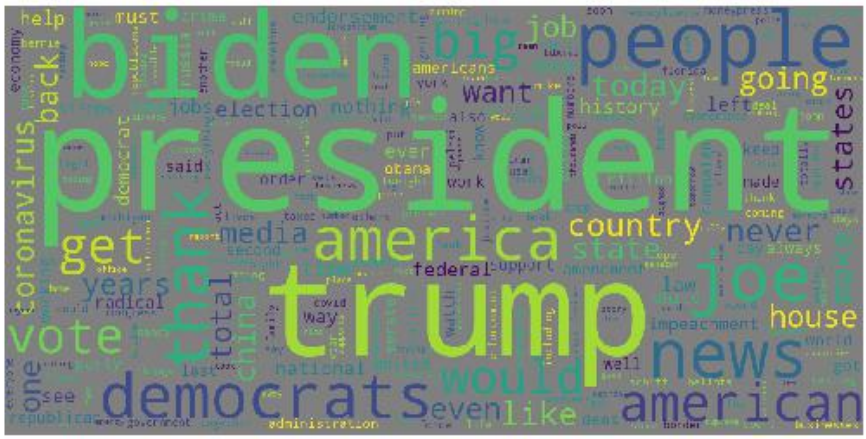


2. De plus, on a essayé d'intégrer des informations externes dans l'ensemble de formation, sous la forme des scores de polarité pour les tweets. On a classifié les mots fréquents de tweets de Trump sous trois labels : positive, négative,neutre

   Voici les résultat de mots classifiés:

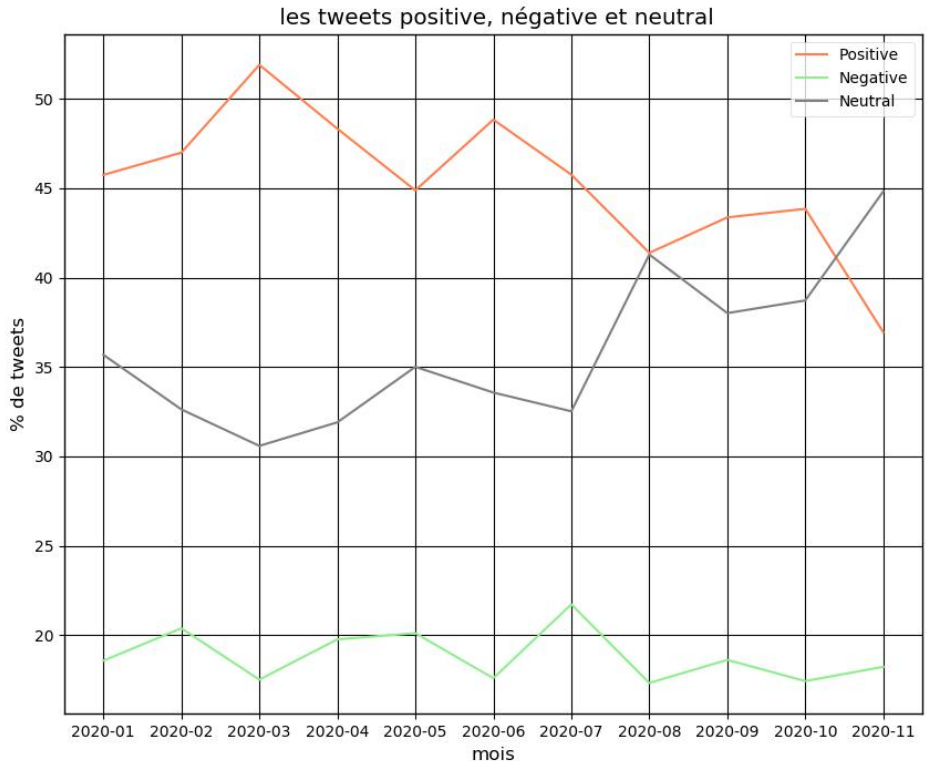

great
new
many
complete
good
strong
first
much
incredible
win
best
right
live
better
really
important
far
economic
love
proud
fast
general
real
true
high
early
tremendous
wants
happy
strongly
action
greatest
primary
ready
social
fantastic
wonderful
absolutely
quickly
beautiful
amazing
wins
clear
full
enjoy
thanks
legal
special
safe
highly
peaceful
completely
sure
major
brave
latest
easy
fair
top
smart
successful
winning
exactly
nice
safely
young
whole
wow
huge


fake
military
bad
small
corrupt
hard
terrible
long
wrong
black
crazy
destroy
illegal
tough
failed
worse
badly
little
foreign
killed
weak
spent
late
past
sick
worst
poorly
false
pathetic
less
disastrous
ridiculous
closed
negative
behind
unfair
game
criminal
violent
dirty
slow
dead
serious
sad
incompetent
forced
dangerous
angry
poor
cold
seriously
horrible
least
hate
center
mess
half
missing
stupid
harder
cutting
sadly
single



Le premier graphique, ce sont le mot nuage, qui décrit les mots postive les plus fréquents dans les tweets de Trump, on peut remarquer ce mot nuage correspond avec les analyse avant sur les contenu, comme le mot ‘great’ qui lié avec la hastag ‘maga’ , dans le deuxième graphique, le mot nuage des vocabulaires négatives, comme le mot fake, milltary, qui correspond avec les confits dans la société américanne. Le troisième graphique décrit le mot neutre, ce qui apparaient le plus souvent est lié avec les votes pour la président et son rival.

3. Pour bien montrer la tenance avec l’approchment de l’élection, on a fait un graphique sur l’évoluation de nombre des mots negative , positive, et neutre.



On peut voir que avec l'approchement de l'élection, le pourcentage de mots positive a décendue, par contre, le pourcentage de mots négatives a augmenté, ce qui reflet aussi la situation instable de la société américanne en mois de l'élection et la stratégie de Trump pour avoir plus de soutient en parlant les soucis de la société.

Partie IIII:

Construire la modèle LDA pour classifier les topic à partir de ces tweets.

Pour améliorer la qualité de l'analyse de textes de tweets, on a décidé utiliser le modèle de machine learning pour classifier les textes. La méthode que j'ai choisi pour classifier les texts en sujets différent est LDA (Latent Dirichlet Allocation). C'est une méthode non-supervisée générative qui se base sur les hypothèses suivantes :

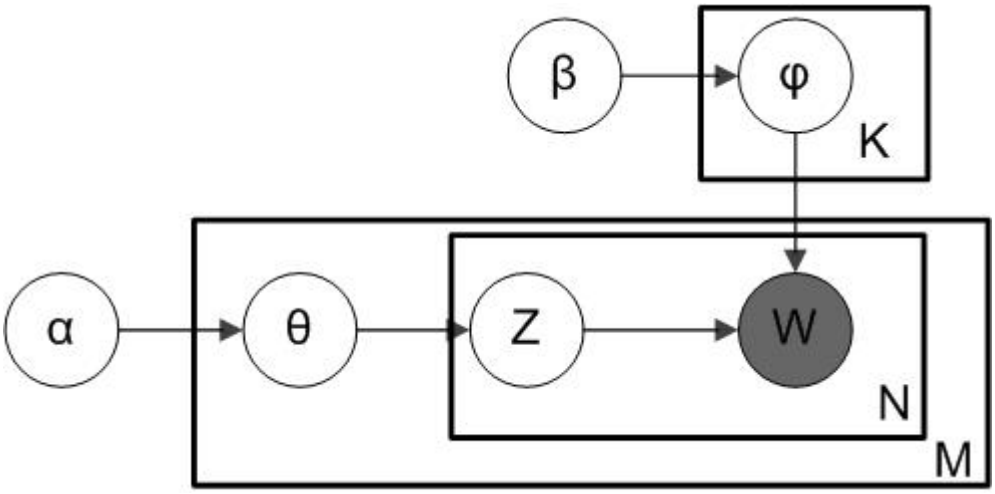
Chaque document du corpus est un ensemble de mots sans ordre (bag-of-words) ;

Chaque document m aborde un certain nombre de thèmes dans différentes proportions qui lui sont propres $p(\theta m)$;

Chaque mot possède une distribution associée à chaque thème $p(\varphi k)$. On peut ainsi représenter chaque thème par une probabilité sur chaque mot.

Puisque l'on a accès uniquement aux documents, on doit déterminer quels sont les thèmes, les distributions de chaque mot sur les thèmes, la fréquence d'apparition de chaque thème sur le corpus.

Une représentation formelle sous forme de modèle probabiliste graphique est la suivante :

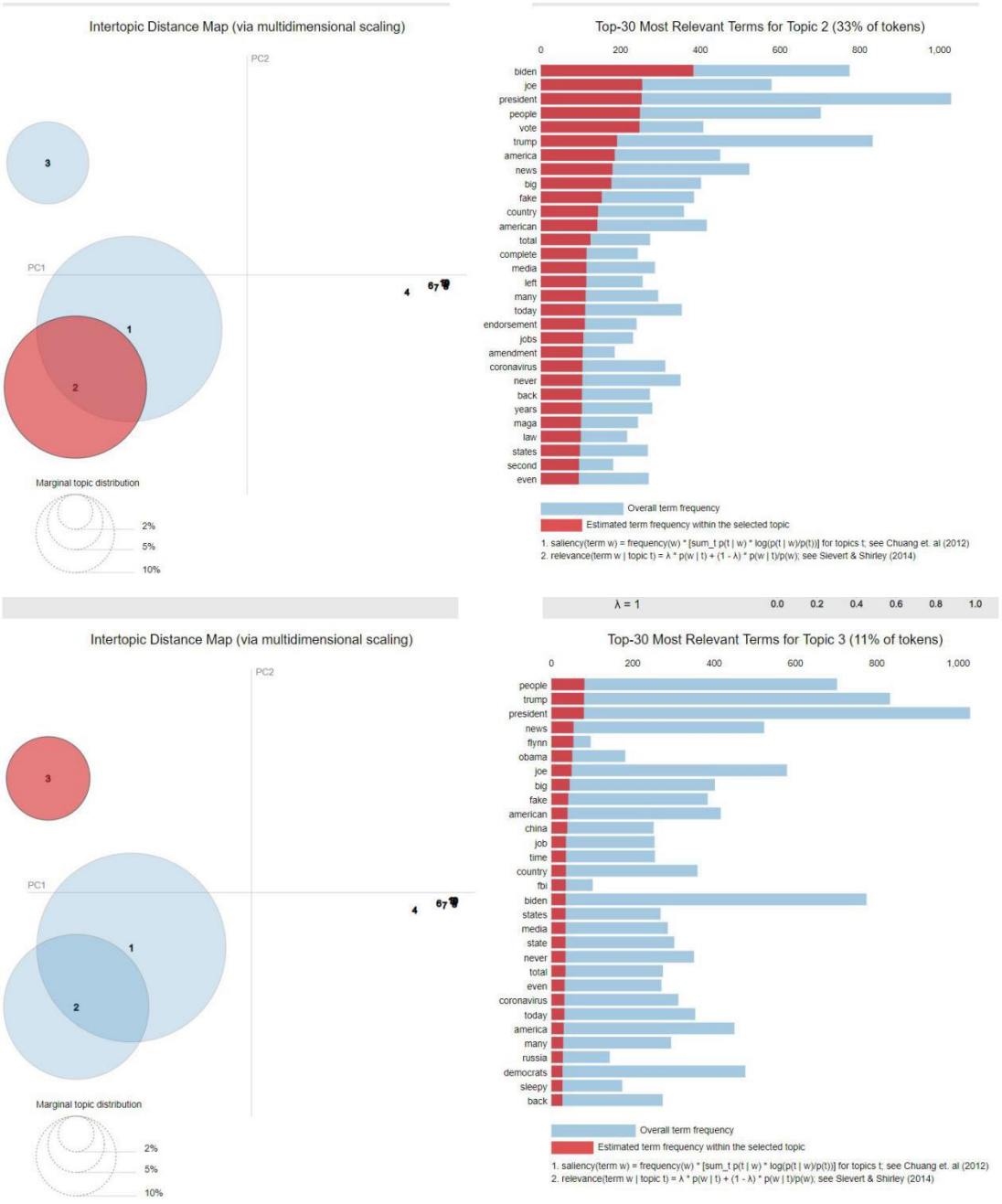


La fichier Topic-analyse.py contient la prétraitement des données et la modélisaton pour classifier les textes et aussi la l'évoluation des résulats. À la fin , en utilisant pyLIDiv pour

Les Lib concernés:

Gensim , pyLDAvis.gensim,nltk.

- Les résulats de la claasification de sujets en analysant les texts.



On peut voir qu'Il y a 3 circle qui représente 3 sujets principales dans tweets de Trump.A la droite, on peut voir les pourcentage des mots, qui consistent ces sujets. Le première le plus parlant, est l'élection, les mots lié avec ce sujuet c'est lui-même et son propositon(maga) , le deuxième sujet est lié avec son rival

Biden, par contre, le mot le plus parlant est fake news. La troisème sujet on peut remarquer c'est le covid-19, le mot dans cette sujet le plus parlant c'est la chine.