Projet TER

Interface web : Analyse avancée des Tweets

Aborder le problème l'apprentissage automatique de thématiques afin de répondre de manière efficace aux différents challenges posés par ce type de données textuelles (tweets) aux méthodes basées sur des modèles probabilistes ou sur la factorisation matricielle.

Rose-Milca CENAT
Université Paris Cité _ Master I AMSD

15/05/2023

TER : Interface web : Analyse avancée des Tweets

L'objectif de ce projet est d'aborder l'apprentissage automatique de thématiques (topics) à partir des tweets afin de répondre de manière efficace aux différents challenges posés par ce type de données textuelles aux méthodes basées sur des modèles probabilistes ou sur la factorisation matricielle.

La réalisation de ce projet comprend quatre étapes principales :

1. Constitution d'un benchmark de données Tweets à partir de Twitter

2. Création d'une Interface web avec Python Dash, Streamlit, ou R shiny pour l'analyse des Tweets

3. Implémentation des méthodes de nettoyage de tweets et intégration à l'interface web

4. Utilisation des approches de "Topics modeling" (LDA, NMF) et intégration à l'interface web

❖ Problématique :

Opinions des gens sur la réforme des retraites 2023 en France.

1. Constitution d'un benchmark de données Tweets à partir de Twitter

La collection des données (tweets) a été faite par l'API de twitter en utilisant la librairie tweepy.
Pour pouvoir y connecter un API key et un API Key Secret est dispensable. Les paramètres de recherches sont les suivantes :

search_words = "#retraites2023 OR #reformedesretraites OR #retraitesfrance exclude:retweets"
lang = "fr"
date_since = "2023-04-01"
date_until = "2023-05-05"
location = "France"
tweet_count = 2000

De ce fait, 2 000 tweets concernant la réforme de retraites postés en France et en français durant la période du 01 avril au 05 mai 2023 ont été collectés. A noter que les retweets ont été exclus, car plus tard, en effectuant les traitements, il ne restait que 79 tweets non retweetés.
Les données collectées ont été stockées dans un fichier csv, quatre informations des tweets ont été stockées : l'id, la date de création, le texte (tweet), l'utilisateur.

2. Création d'une Interface web avec Streamlit pour l'analyse des Tweets

La création de l'interface web est avec Streamlit. C'est une fenêtre avec 4 menu principal: Hello, Tweets Analysis, Tweets Cleaning, LDA NMF.

- **Hello**

Ce menu n'affiche que l'objectif et les réalisations du projet.

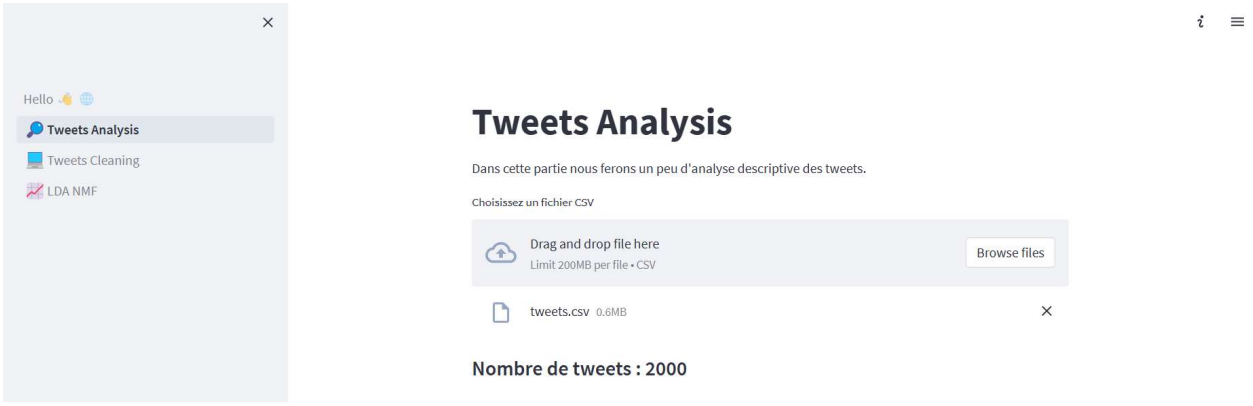
- **Tweets Analysis**

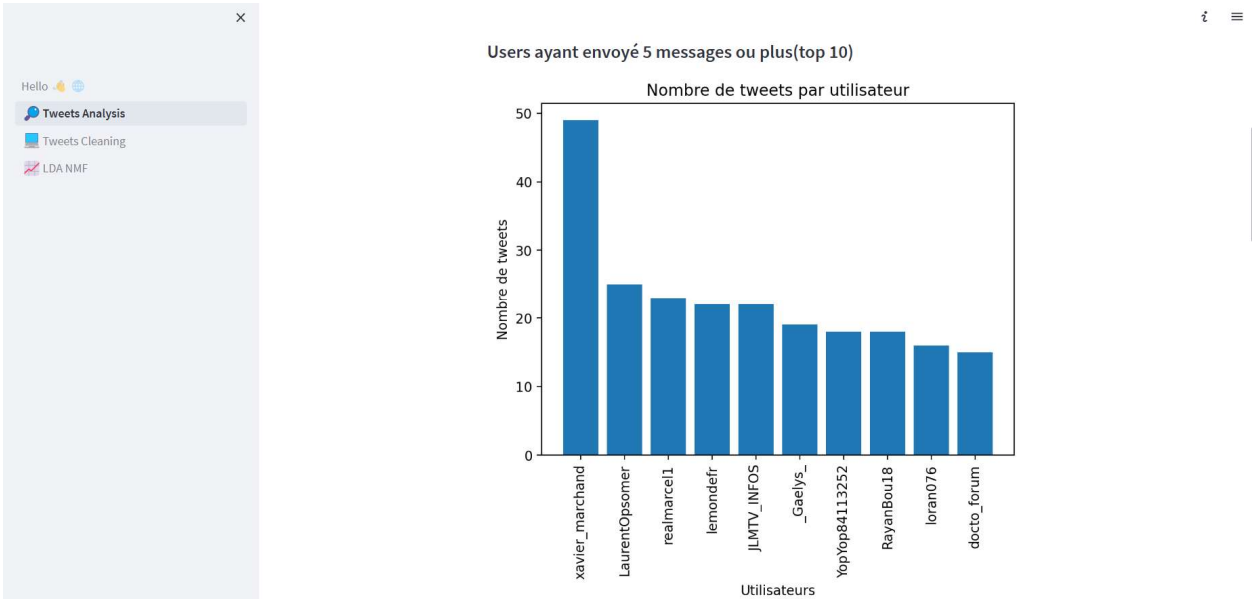
Ce menu est pour l'analyse descriptive des tweets.

L'utilisateur a la possibilité d'upload le fichier des tweets à analyser. Cette étape est obligatoire pour la suite. Après que le fichier soit upload, l'application affichera le nombre de tweets contenant dans le fichier, le nombre de message par utilisateur(les 10 premiers utilisateurs), un diagramme à barres des 10 premiers utilisateurs ayant posté 5 tweets ou plus, les hashtags (10) les plus fréquents, un diagramme à barres des mentions (5) les plus fréquentes, l'affichage ainsi qu'un diagramme des 10 mots les plus fréquents dans les tweets, sans nettoyage, cependant les caractères spéciaux ont été enlevés.

- **Avant que le fichier soit upload**

- **Après que le fichier soit upload**



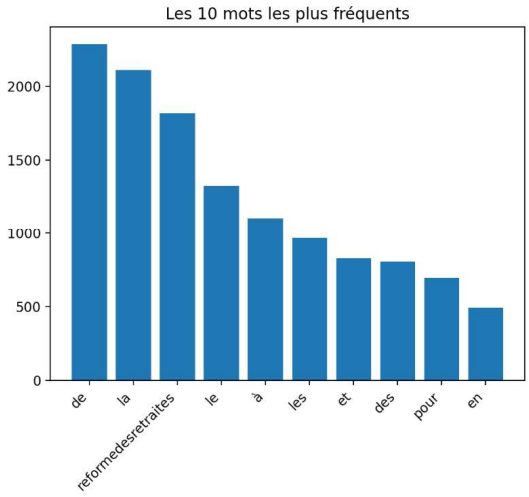


Hello 👋 🌐

Tweets Analysis

Tweets Cleaning

LDA NMF

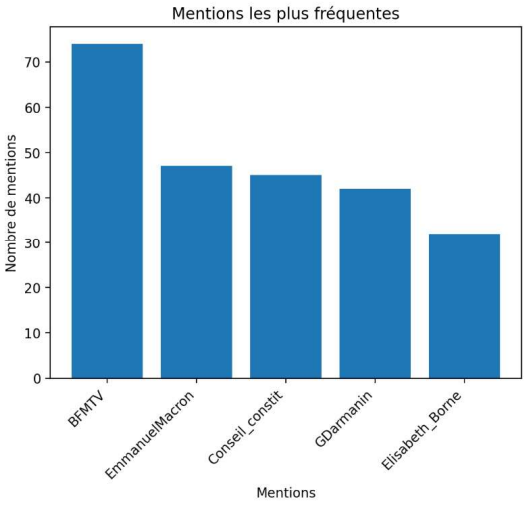


Hello 👋 🌐

Tweets Analysis

Tweets Cleaning

LDA NMF



Hello 👋 🌐

Tweets Analysis

Tweets Cleaning

LDA NMF

Hashtags plus fréquents (top 10)

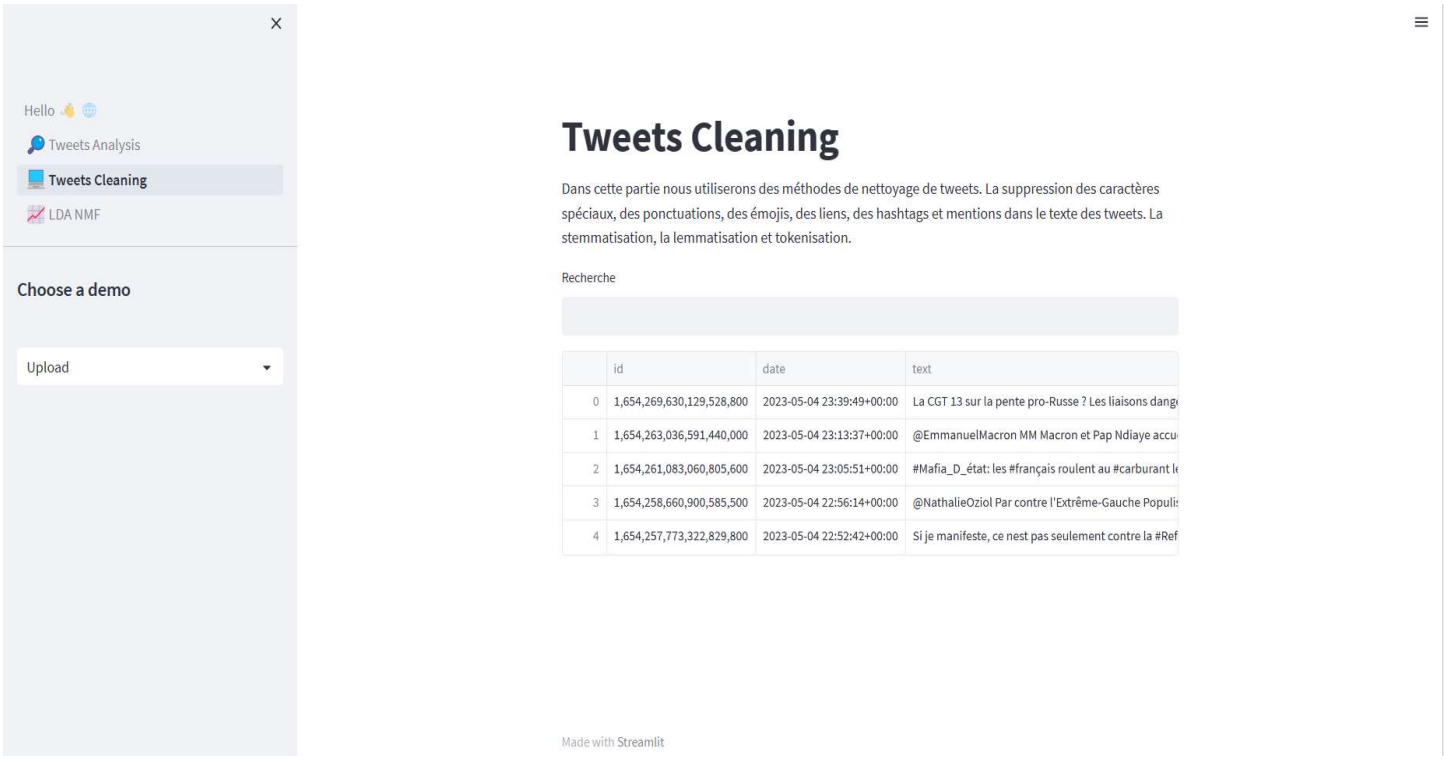
hashtags	count
#ReformeDesRetraites	1,701
#Macron	295
#RIP	142
#RéformeDesRetraites	130
#ViolencesPolicieres	109
#1erMai	98
#reformedesretraites	92
#casserolades	87
#manifestation	83
#France	79

3. Implémentation des méthodes de nettoyage de tweets et intégration à l'interface web

Dans cette partie, on utilise des méthodes de nettoyage des tweets. Ce menu comporte trois sous-menus : Upload, Preprocessing et Word Processor.

- **Upload**

Ce menu affiche dans un tableau l'id, la date, le texte, l'username, les hashtags et les mentions de chaque tweet. L'utilisateur peut faire des recherches sur le texte des tweets.



- **Preprocessing**

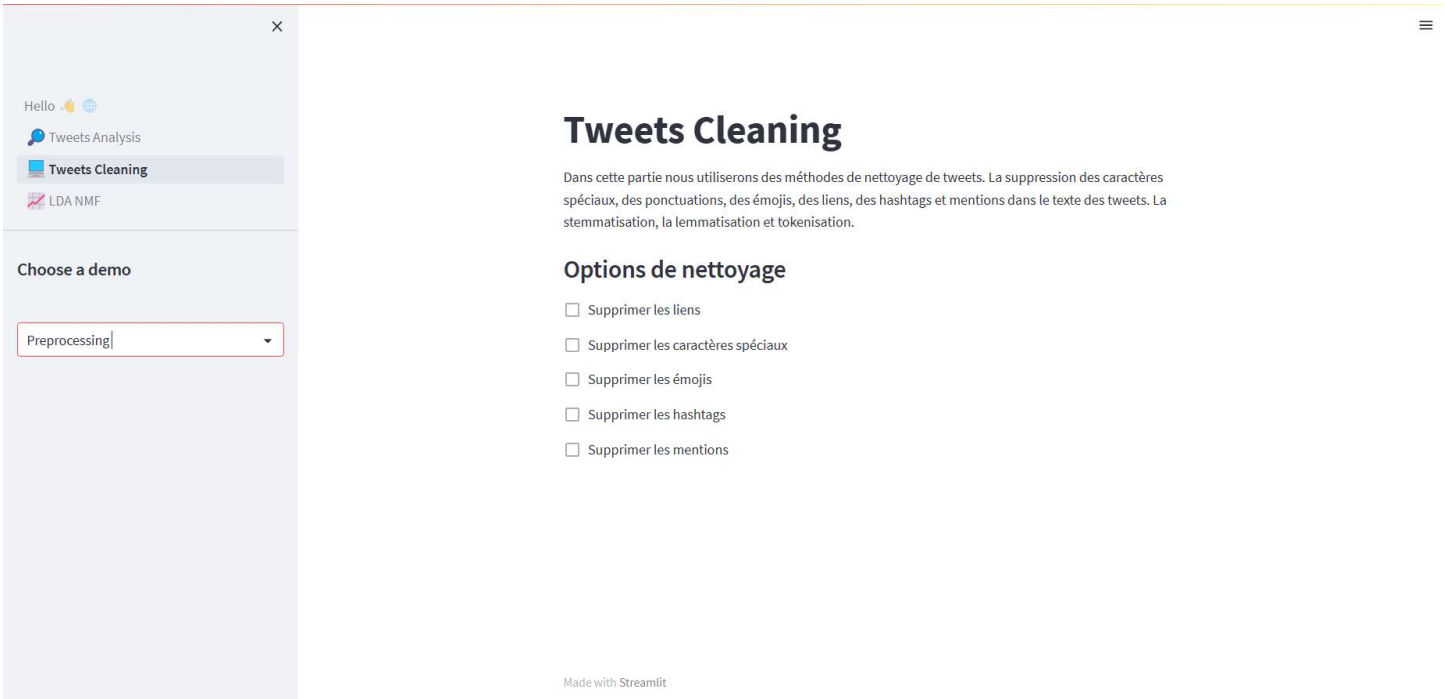
Dans ce sous-menu, les options de nettoyage sont sur la forme de case à cocher. Quand l'utilisateur coche au moins une case, il voit apparaitre un bouton lui permettant de télécharger un fichier contenant les mêmes données qu'il avait avant et trois nouvelles colonnes: hashtags, mentions et le texte traité en fonction de l'option qui avait été cochée. Il verra aussi un tableau affichant le texte avant et après le traitement de nettoyage appliqué.

Les options de nettoyage sont les suivantes :

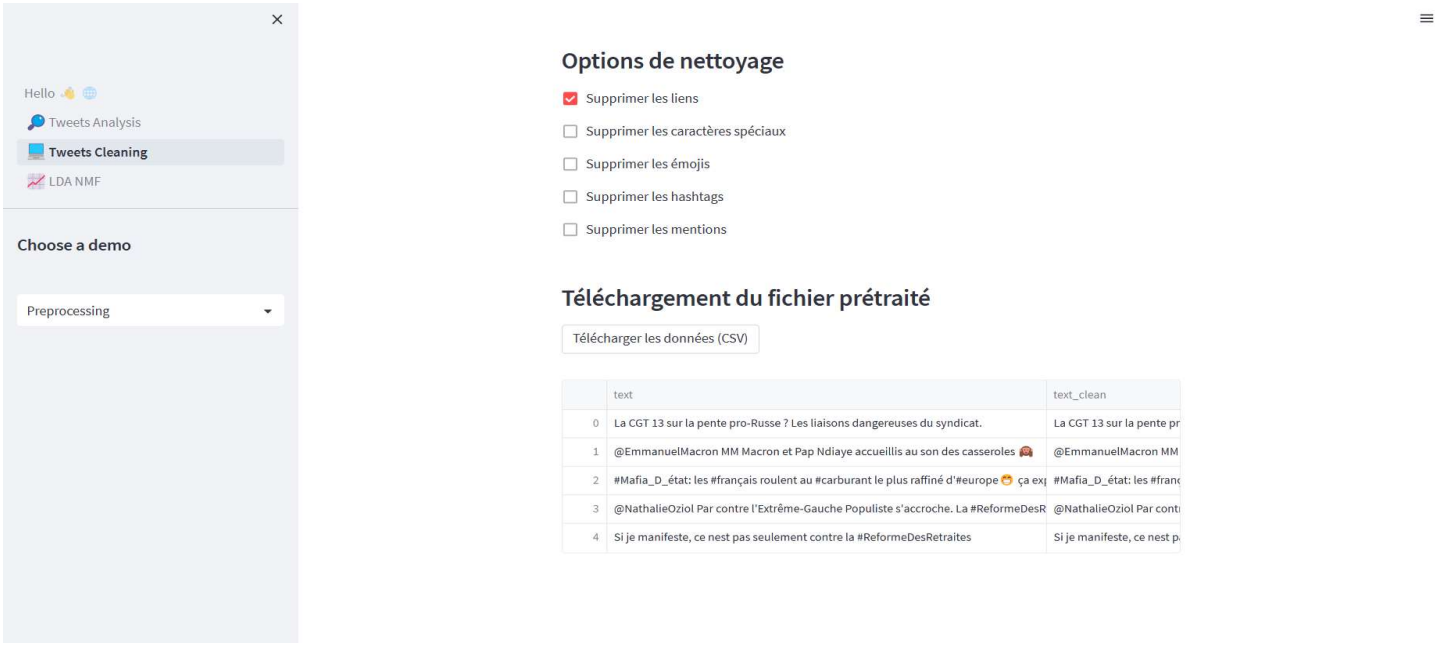
- ☒ Supprimer les liens
- ☒ Supprimer les caractères spéciaux
- ☒ Supprimer les des ponctuations
- ☒ Supprimer les émojis
- ☒ Supprimer les hashtags
- ☒ Supprimer les mentions

NB : Il y a les options Supprimer les caractères spéciaux et Supprimer les hashtags afin de laisser toute la liberté à l'utilisateur. S'il enlève que les caractères spéciaux, seul le dièse devant les hashtags sera enlevé, le texte restera. Mais s'il enlève les hashtags tout le texte sera effacé, ce qui aura un effet bien sur les analyses futures.

- **Avant que l'utilisateur coche une case**

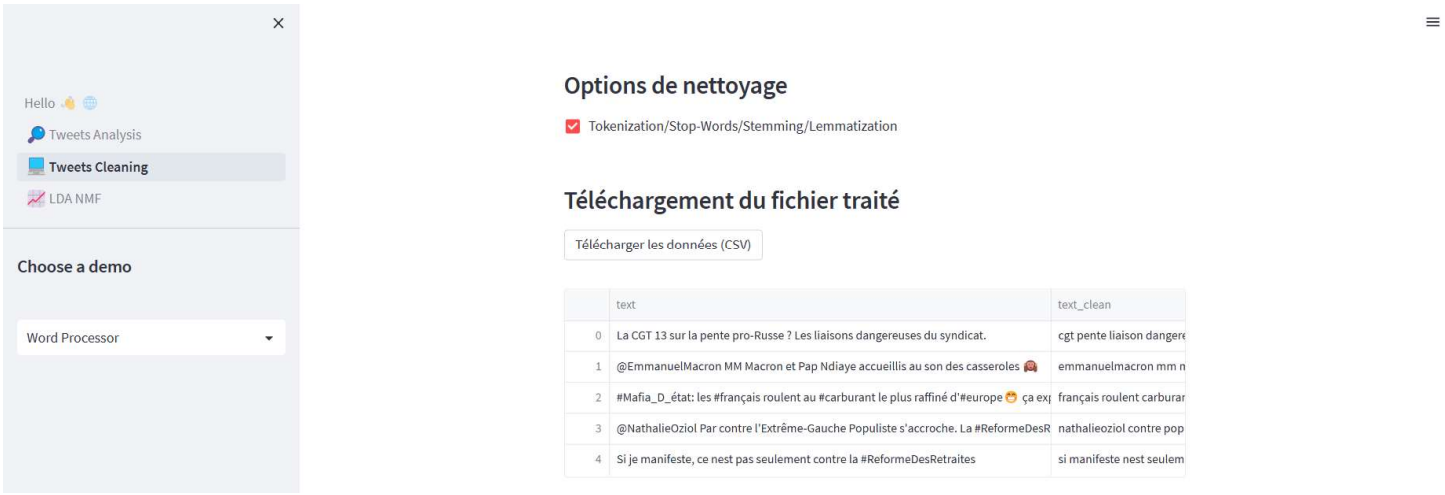


- **Après que l'utilisateur ait coché une case**



- **Word Processor**

Cette partie est dédiée à la Tokenization, la suppression des Stop-Words, au Stemming et la Lemmatization. On affiche un Wordcloud des textes (tweets) avant ces étapes. Ensuite, il y a une case à cocher : Tokenization/Stop-Words/Stemming/Lemmatization. Si l'utilisateur la coche, il verra un bouton lui permettant de télécharger le fichier nettoyé, un tableau affichant le texte et le texte nettoyé et pour finir un nouveau Wordcloud affichant les mots des tweets après les traitements.



4. Utilisation des approches de “Topics modeling” (LDA, NMF) et intégration à l’interface web

Ce menu a deux sous-menus: LDA et NMF.

- **LDA**

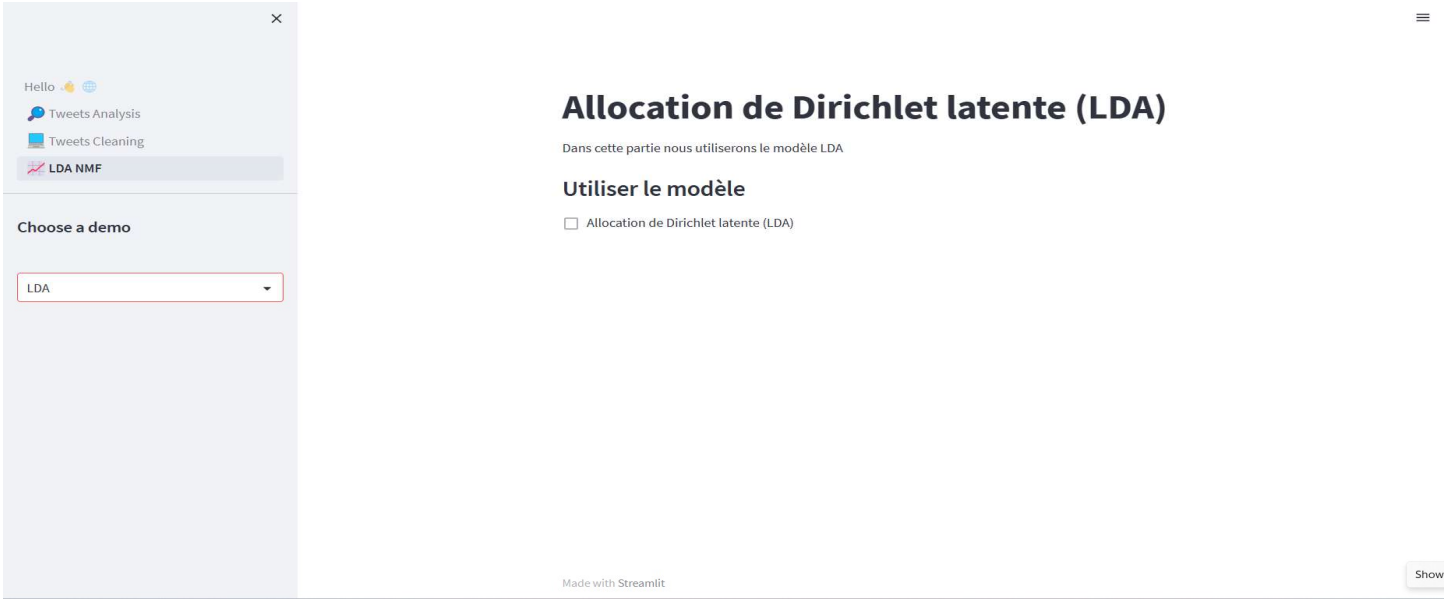
Dans ce sous menu, il ya une case à cocher "Allocation de Dirichlet latente (LDA)", si l’utilisateur la coche, cela déclenchera l’entrainement du modèle LDA. On utilise random.seed() afin que le modèle soit initialisé de la même manière à chaque exécution, ce qui garantit que les résultats seront cohérents et reproductibles.

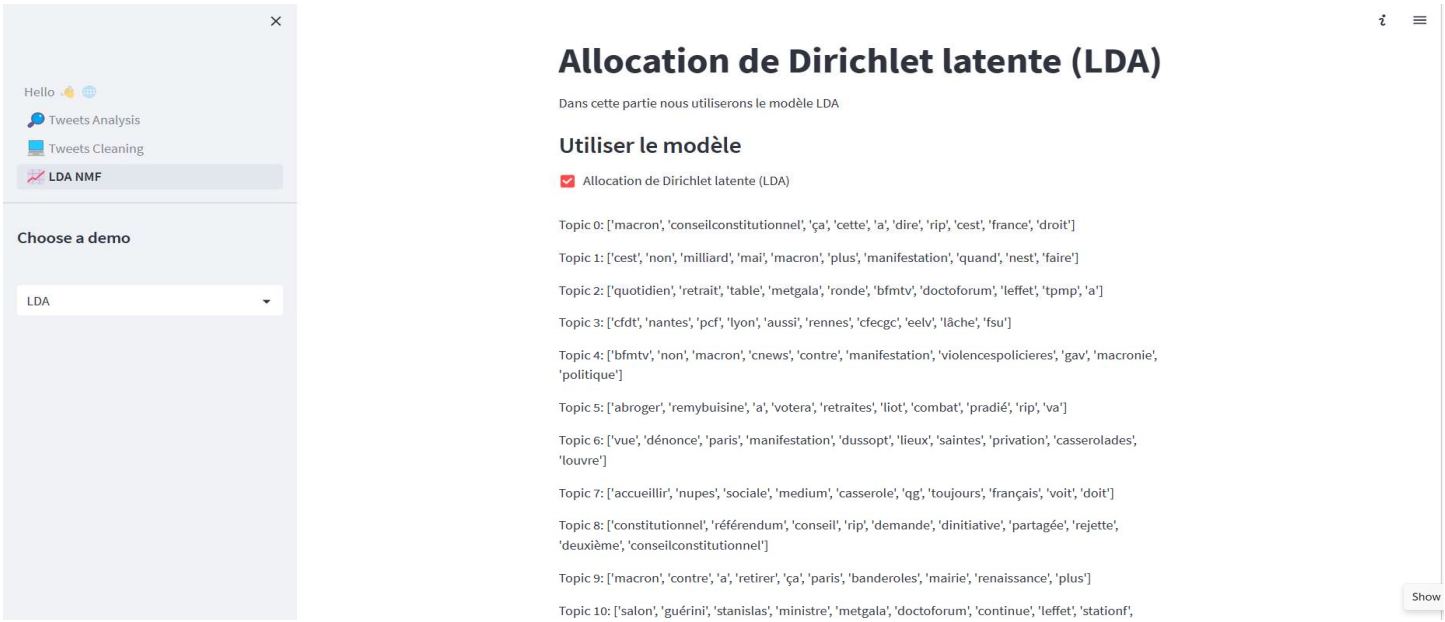
On a procédé comme suit :

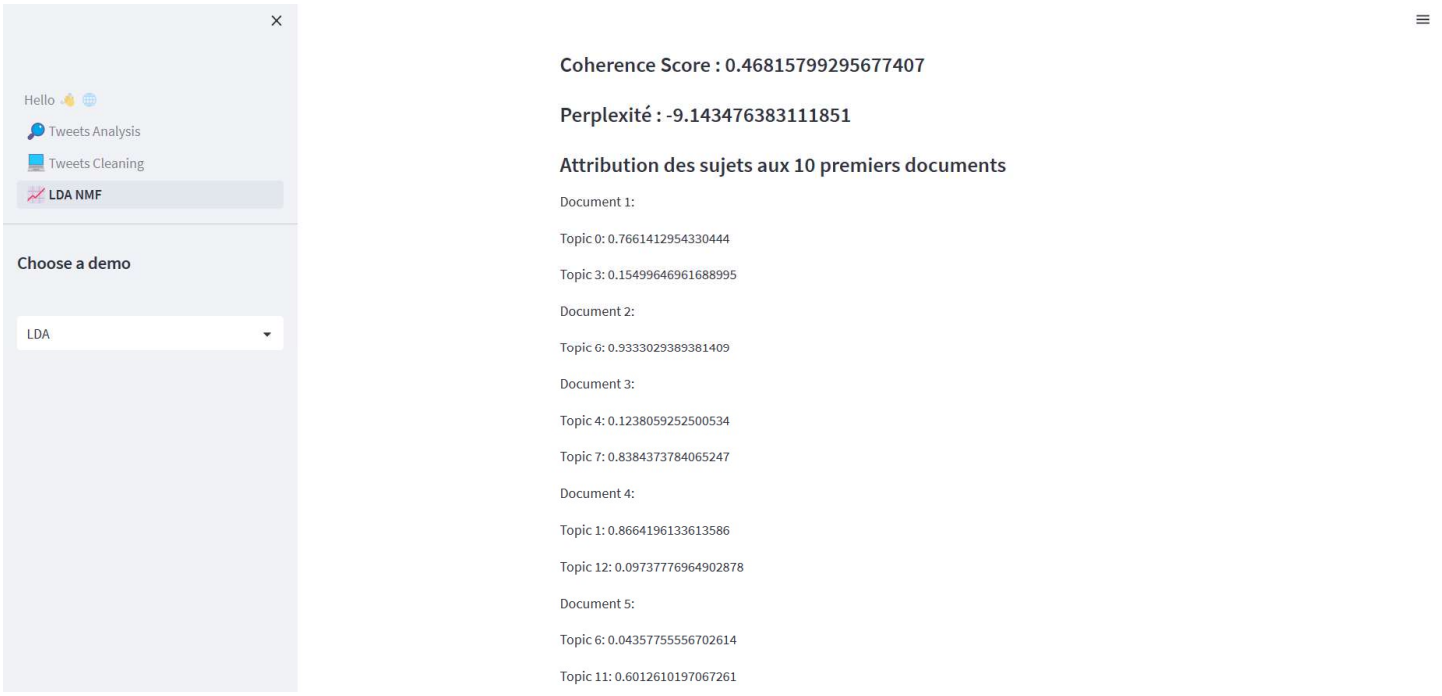
- Division de chaque tweet en une liste de mots séparés par des espaces
- Création d'un dictionnaire à partir des documents prétraités

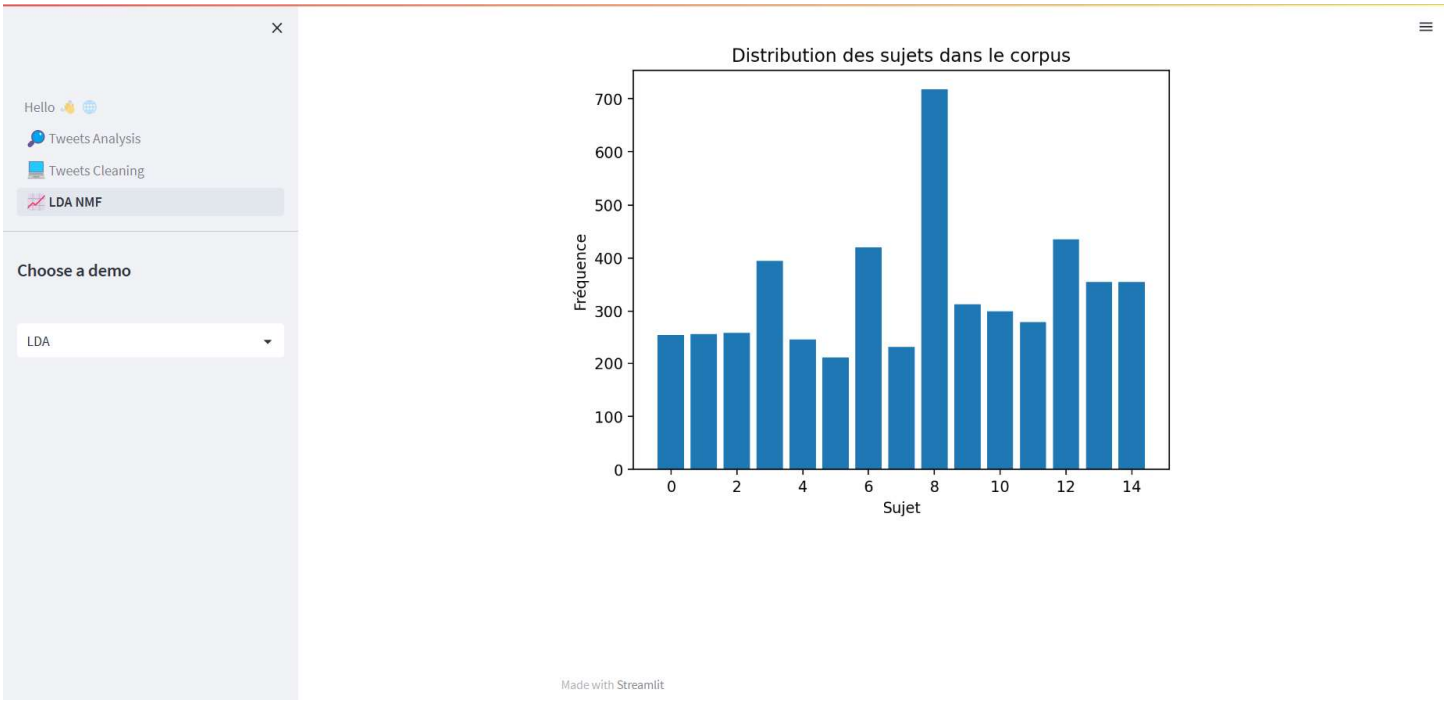
- Transformation de chaque document en un vecteur de fréquence de mots avec doc2bow ()
- Calcul de TF-IDF sur les documents transformés
- Entraînement du modèle LDA avec les hyper paramètres :
  corpus = corpus_tfidf,
  id2word = id2word,
  num_topics = 15,
  passes = 25,
  random_state = 100

- Affichage des sujets avec leurs mots clés
- Affichage du score de la cohérence et de la perplexité
- Attribution des sujets aux 10 premiers documents
- Distribution des sujets dans le corpus









- **NMF**

Dans ce sous menu, il ya une case à cocher "Factorisation matricielle non négative (NMF)", si l'utilisateur la coche, cela déclenchera l'entrainement du modèle NMF.

On a procédé comme suit :

- Vectorisation des données avec TfidfVectorizer () de sklearn.decomposition avec les hyper paramètres :
  max_df=0.95,
  min_df=2

- Entraînement du modèle NMF avec les hyper paramètres :
  n_components=10,
  random_state=4

- Affichage des sujets (Top 10)