

A decorative graphic on the right side of the page. It features three concentric blue circles of different sizes, with the largest one at the top right and the smallest one in the middle. Thin blue lines intersect these circles and extend across the page, creating a geometric design.

# 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 #reformesdesretraites 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

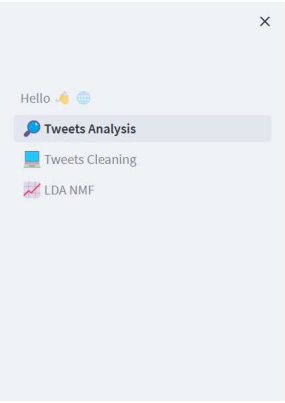
Ce menu est pour l'analyse descriptive des tweets.

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



i ≡

# Tweets Analysis

Dans cette partie nous ferons un peu d'analyse descriptive des tweets.

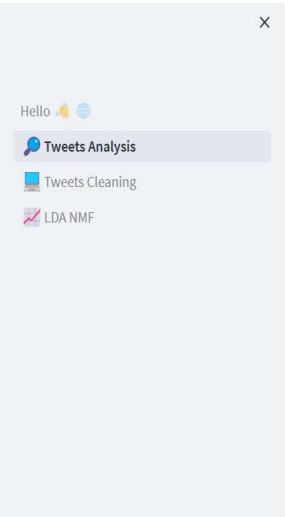
Choisissez un fichier CSV

Drag and drop file here  
Limit 200MB per file • CSV

Browse files

tweets.csv 0.6MB

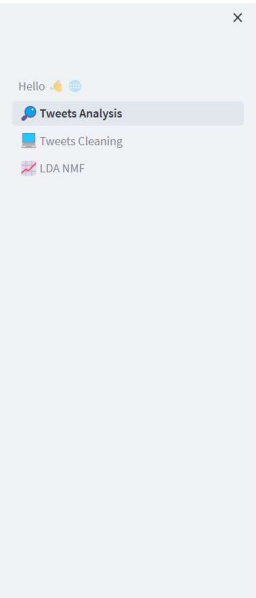
Nombre de tweets : 2000



i ≡

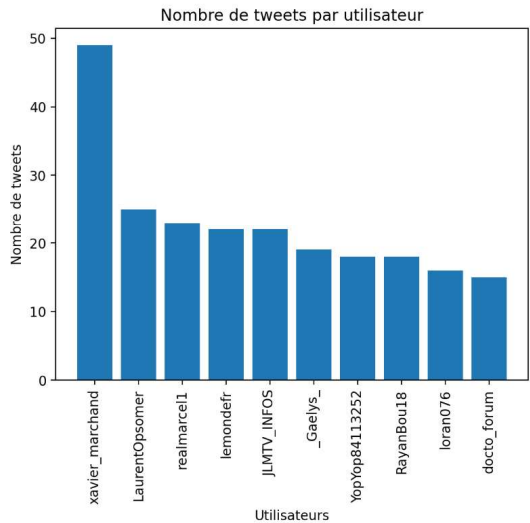
## Nombre de message par users (top 10)

username	count
xavier_marchand	49
LaurentOpsomer	25
realmarcel1	23
lemondefr	22
JLMTV_INFOS	22
_Gaelys_	19
YopYop84113252	18
RayanBou18	18
loran076	16
docto_forum	15

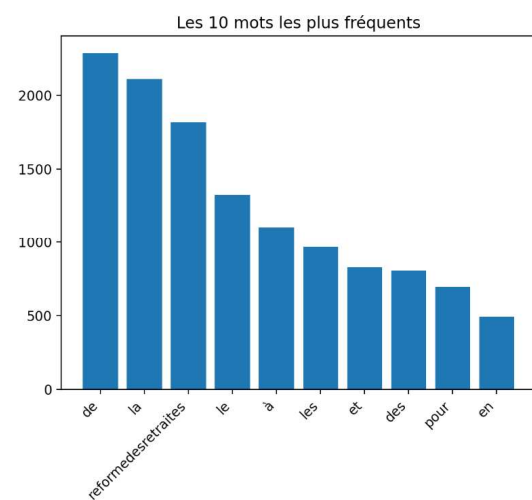


i ≡

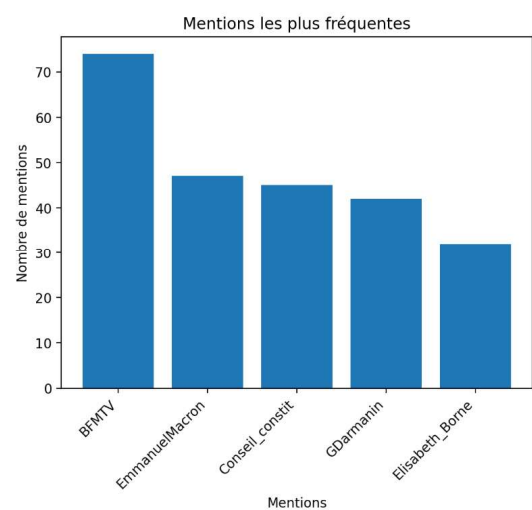
## Users ayant envoyé 5 messages ou plus(top 10)



A screenshot of a mobile application interface. At the top right is a close button (X). Below it, the text "Hello" is followed by a yellow smiley face emoji and a blue globe icon. A horizontal separator line follows. Below the separator is a list of items. The first item is "Tweets Analysis" with a blue magnifying glass icon. The second item is "Tweets Cleaning" with a blue document icon. The third item is "LDA NMF" with a pink document icon. The background is a light gray gradient.

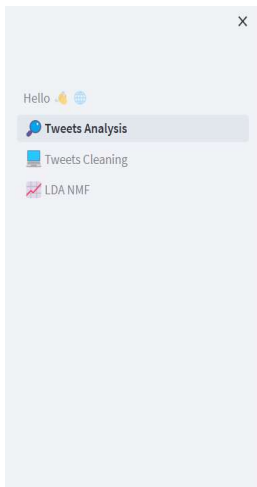


A screenshot of a mobile application interface. At the top, there is a header bar with a light blue background. On the left, it says "Hello" followed by a yellow smiley face icon and a blue speech bubble icon. On the right, there is a close button represented by a black "X" icon. Below the header, there is a list of tweets. The first tweet is highlighted with a light blue background. It features a blue speech bubble icon, the text "Tweets Analysis", and a small blue speech bubble icon. The second tweet features a blue speech bubble icon, the text "Tweets Cleaning", and a small blue speech bubble icon. The third tweet features a blue speech bubble icon, the text "LDA NMF", and a small blue speech bubble icon. At the bottom of the screen, there is a navigation bar with a light blue background. It contains four icons: a blue speech bubble icon, a blue speech bubble icon, a blue speech bubble icon, and a blue speech bubble icon.



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



Les 10 mots les plus fréquents dans les tweets (Sans les caractères spéciaux) :

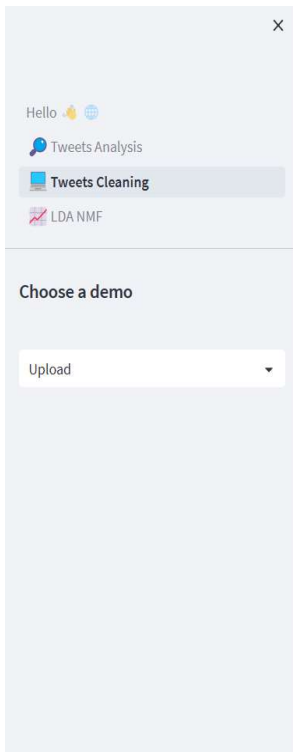
	mot	compte
0	de	2,286
1	la	2,109
2	reform	1,817
3	le	1,323
4	à	1,103
5	les	965
6	et	826
7	des	803
8	pour	692
9	en	493

### 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.



## Tweets Cleaning

Dans cette partie nous utiliserons des méthodes de nettoyage de tweets. La suppression des caractères spéciaux, des ponctuations, des émojis, des liens, des hashtags et mentions dans le texte des tweets. La stemming, la lemmatisation et tokenisation.

Recherche

	id	date	text
0	1,654,269,630,129,528,800	2023-05-04 23:39:49+00:00	La CGT 13 sur la pente pro-Russe ? Les liaisons dangi
1	1,654,263,036,591,440,000	2023-05-04 23:13:37+00:00	@EmmanuelMacron MM Macron et Pap Ndiaye accu
2	1,654,261,083,060,805,600	2023-05-04 23:05:51+00:00	#Mafia_D_état: les #français roulent au #carburant le
3	1,654,258,660,900,585,500	2023-05-04 22:56:14+00:00	@NathalieOziol Par contre l'Extrême-Gauche Populi:
4	1,654,257,773,322,829,800	2023-05-04 22:52:42+00:00	Si je manifeste, ce nest pas seulement contre la #Ref

## ▪ 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 apparaître 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

×

Hello 🌍

Tweets Analysis

**Tweets Cleaning**

LDA NMF

Choose a demo

Preprocessing

## Tweets Cleaning

Dans cette partie nous utiliserons des méthodes de nettoyage de tweets. La suppression des caractères spéciaux, des ponctuations, des émojis, des liens, des hashtags et mentions dans le texte des tweets. La stemming, la lemmatisation et tokenisation.

### Options de nettoyage

- ☐ Supprimer les liens
- ☐ Supprimer les caractères spéciaux
- ☐ Supprimer les émojis
- ☐ Supprimer les hashtags
- ☐ Supprimer les mentions

Made with Streamlit

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

X

Hello 🙋🌐

Tweets Analysis

Tweets Cleaning

LDA NMF

Choose a demo

Preprocessing ▾

### Options de nettoyage

- ☒ Supprimer les liens
- ☐ Supprimer les caractères spéciaux
- ☐ Supprimer les émojis
- ☐ Supprimer les hashtags
- ☐ Supprimer les mentions

### Téléchargement du fichier prétraité

[Télécharger les données \(CSV\)](#)

	text	text_clean
0	La CGT 13 sur la pente pro-Russe ? Les liaisons dangereuses du syndicat.	La CGT 13 sur la pente pr
1	@EmmanuelMacron MM Macron et Pap Ndiaye accueillis au son des casseroles 🥳	@EmmanuelMacron MM
2	#Mafia_D_état: les #français roulent au #carburant le plus raffiné d'Europe 🇪🇺 ça ex	#Mafia_D_état: les #franc
3	@NathalieZioliol Par contre l'Extrême-Gauche Populiste s'accroche. La #ReformeDesR	@NathalieZioliol Par cont
4	Si je manifeste, ce nest pas seulement contre la #ReformeDesRetraites	Si je manifeste, ce nest p

- **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.

X

Hello 🌤️🌐

Tweets Analysis

Tweets Cleaning

LDA NMF

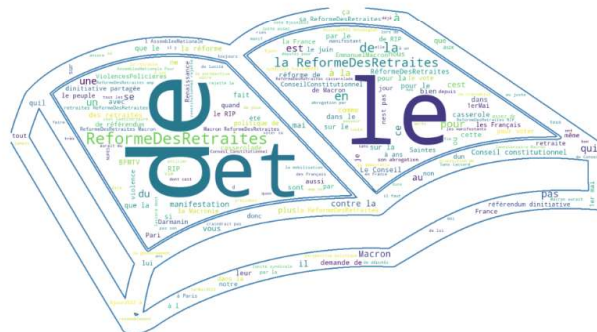
Choose a demo

Word Processor ▼

## Tweets Cleaning

Dans cette partie nous utiliserons des méthodes de nettoyage de tweets. La suppression des caractères spéciaux, des ponctuations, des émojis, des liens, des hashtags et mentions dans le texte des tweets. La stemming, la lemmatisation et tokenisation.

WordCloud avant Stemming/Lemmatization/Supp  
StopWords





A screenshot of the Twitter application interface. At the top right is a close button (X). Below it, a greeting "Hello" is followed by a sun icon and a globe icon. The main content area shows two tweet cards. The first card is titled "Tweets Analysis" and features a blue speech bubble icon. The second card is titled "Tweets Cleaning" and features a blue document icon with a checkmark. Below the "Tweets Cleaning" card is a link "LDA NMF" with a red and blue line graph icon. A horizontal separator line divides the content from the bottom section. The bottom section is titled "Choose a demo" and contains a dropdown menu with "Word Processor" selected and a downward arrow.

### Options de nettoyage

### Téléchargement du fichier traité

	text	text_clean
0	La CGT 13 sur la pente pro-Russe ? Les liaisons dangereuses du syndicat.	cg_t pente liaison dangere
1	@EmmanuelMacron MM Macron et Pap Ndiaye accueillis au son des casseroles 🥳	emmanuelmacron mm n
2	#Mafia_D_état: les #français roulent au #carburant le plus raffiné d'Europe 🇪🇺 ça ex	français roulent carburan
3	@NathalieOziol Par contre l'Extrême-Gauche Populiste s'accroche. La #ReformeDesR	nathalieoziol contre pop
4	Si je manifeste, ce nest pas seulement contre la #ReformeDesRetraites	si manifeste nest seulem

[illegible]

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

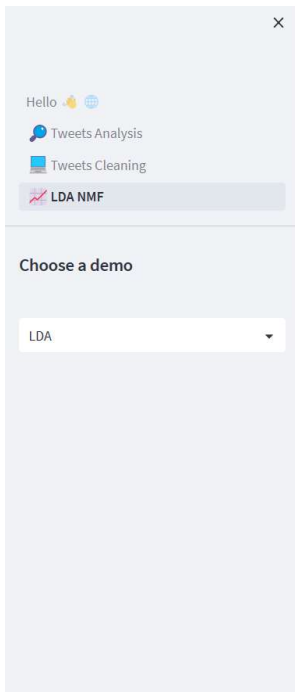
- LDA

On a procédé comme suit :

- ```
corpus = corpus_tfidf,
id2word = id2word,
num_topics = 15,
passes = 25,
random_state = 100
```

## Allocation de Dirichlet latente (LDA)

## Allocation de Dirichlet latente (LDA)



Coherence Score : 0.46815799295677407

Perplexité : -9.143476383111851

#### Attribution des sujets aux 10 premiers documents

Document 1:

Topic 0: 0.7661412954330444

Topic 3: 0.15499646961688995

Document 2:

Topic 6: 0.9333029389381409

Document 3:

Topic 4: 0.1238059252500534

Topic 7: 0.8384373784065247

Document 4:

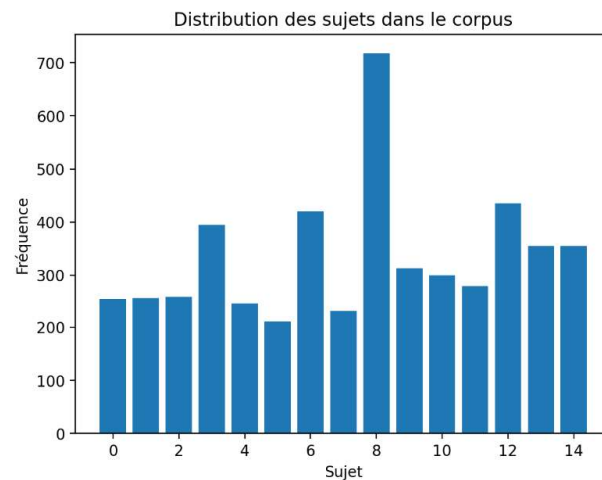
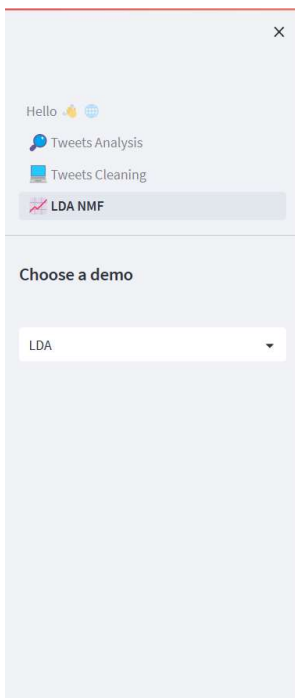
Topic 1: 0.8664196133613586

Topic 12: 0.09737776964902878

Document 5:

Topic 6: 0.0435775556702614

Topic 11: 0.6012610197067261



Made with Streamlit

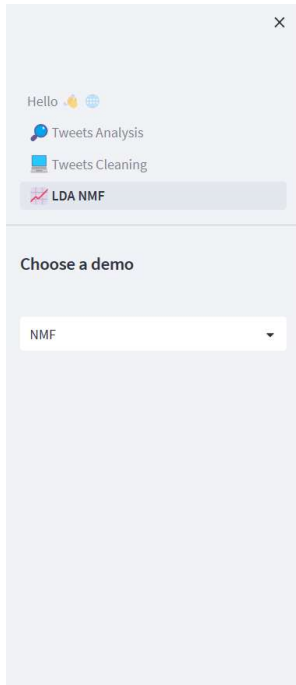
## ■ NMF

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



## Factorisation matricielle non négative (NMF)

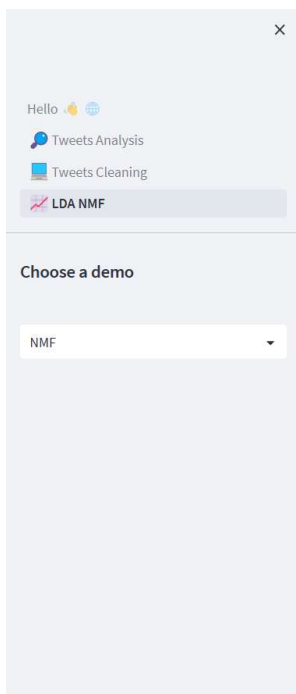
Dans cette partie nous utiliserons le modèle NMF

### Utiliser le modèle

☐ Factorisation matricielle non négative (NMF)

Made with Streamlit

Monday, May



Dans cette partie nous entraînerons le modèle NMF.

### Utiliser le modèle

☒ Factorisation matricielle non négative (NMF)

Topic #0: craindrait, onvagner, lunité, lanniversaire, perspective, laccord, assembleenationale, abrogation, assez, avecclanupes

Topic #1: référendum, conseil, constitutionnel, dinitiative, demande, partagée, rejetée, deuxième, retraites, réforme

Topic #2: reformedesretraites, contre, retraites, france, cest, ça, français, plus, réforme, cette

Topic #3: pcf, lâche, unsa, cfdt, cfecgc, devise, lycee, fsu, rennes, eelv

Topic #4: non, reformedesretraites, bfmtv, olivierdusopt, brunolemaire, elisabethborne, cnews, toujours, cest, auroreberge

Topic #5: manifestation, paris, mai, vue, lors, gardes, police, dénonce, liberté, lieux

Topic #6: macron, saintes, casserolades, reformedesretraites, casserole, macrondémission, dusopt, intervallesmacron, casseroladegenerale, lycée

Topic #7: rip, conseilconstitutionnel, décision, reformedesretraites, conseilconstit, sage, démocratie, constitutionnel, conseil, demande

Topic #8: an, retraite, fabius, touché, mediapart, laurent, conseil, président, constitutionnel, révèle

Topic #9: cgt, tout, va, défense, énergie, noir, faire, continues, chez, si

Monday, May