project Data Mining ¶

Réalisé par : Nidhal Hazbri 3DNI2

Objectifs :

- Maitriser l'API de twitter pour l'extraction des tweets
- Maitriser la partie NLP (natural language processing) avec NLTK en Python
- Appliquer les principes de nettoyage des données
- Classer les tweets : regrouper ensemble les tweets qui sont similaires. C'est une étape qui peutêtre considérée comme une étape

Specifications

Imaginons que vous avez un compte Twitter, et que vous lez suivre les tweets sur ce reseau social. Vu le nombre colossal de Tweets, et faute de temps, vous n'avez pas la possibilite de les lire tous. Pour cela, vous avez besoin d'une application qui va jouer le role d'assistantet qui va vous effectuer un resume de toutes ces informations. Une des approches qu'on peut utiliser estde le classer sous former de groupes de sorte a ce qu'on presente a l'utilisateur un seul Tweet de chaque groupe. Pour cela, on doit proceder en trois grandes etapes :

Travail faire

On a Telecharger les tweets a partir de Twitter en utilisant l'API de twitter. Pour cela, vous devriez un compte « Twitter Developper ». Pour cela, vous devriez telecharger au moins 10 mille twwets. Pour la documentation de l'API de twitter, vous pouvez consulter les liens suivants :

In [31]:
```
import pandas as pd
import tweepy
consumer_key="LHZVzcEN30hfmN2cPBqkoB3wq"
consumer_secret="DGZ7gQFDlqXoPfmAUWHOsY2eMTA0qhgKVb3rbExcx8Vhav3x3a"
access_token="1325046107437752325-a2zNm36NnzJqTFBFkIagjzpkdCadjs"
access_token_secret="7ohQJ7WTf2DuHsr9NNwPkOPXq5zUkaycrzo2nPhPUoGLL"
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
```

In [32]:
```
twitter_data_analysis = pd.DataFrame(columns = ['text'])
i=0
```

In [33]:
```
tweets = tweepy.Cursor(api. user_timeline , id="twitter").items( 15000)
# Iterate and print tweets
for tweet in tweets:
    twitter_data_analysis.loc[i,"text"] = tweet.text
    i+=1
```

In [34]:
```
print(twitter_data_analysis.shape)
```

```
(3225, 1)
```

In [35]:
```
tweets = tweepy.Cursor(api. user_timeline , id="twitter").items( 15000)
# Iterate and print tweets
for tweet in tweets:
    twitter_data_analysis.loc[i,"text"] = tweet.text
    i+=1
```

In [36]:
```
print(twitter_data_analysis.shape)
```

```
(6450, 1)
```

In [37]:
```
tweets = tweepy.Cursor(api. user_timeline , id="twitter").items( 15000)
# Iterate and print tweets
for tweet in tweets:
    twitter_data_analysis.loc[i,"text"] = tweet.text
    i+=1
```

In [38]:
```
print(twitter_data_analysis.shape)
```

```
(9675, 1)
```

In [39]:
```
import csv
twitter_data_analysis.to_csv('twitter_data_analysis.csv',index = False)
twitter_data_analysis.head(10)
```

Out[39]:

	text
0	There’s more! We’ll also be testing sharing Tw...
1	Oh snap! \n\nSharing Tweets directly to your ...
2	@levantinepali a stamp of approval https://t.c...
3	2020 in one word
4	@Astro_AJC this is what cuffing season means t...
5	@un3asyy 2 is also cute
6	@DeePeeArts you're all amazing
7	RT @shesooosaddity: if you had a twitter befor...
8	@CloudNaii 40404
9	@issahairplug drink water replaced good morning

In [41]:
```
twitter_data_analysis.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9675 entries, 0 to 9674
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    9675 non-null   object
dtypes: object(1)
memory usage: 471.2+ KB
```

Pretraitement des tweets

Dans cette etape, l'objectif est d'eliminer le texte inutile des tweets tels que les #, les noms des utilisateurs, les url,emoji ...

In [42]:
```
import re
for index, row in twitter_data_analysis.iterrows():
    err = row['text']
    new0 = re.sub(r"http\S+", "", err)
    new1 = re.sub(r"#\S+", "", new0)
    new2 = re.sub(r"@\S+", "", new1)
    new3 = re.sub(r"\n+", "", new2)
    new4 = re.sub(r"RT+", "", new3)
    new5 = re.sub("hhh+", '', new4)
    emoji_pattern = re.compile("["
                               u"\U0001F600-\U0001F64F"  # emoticons
                               u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                               u"\U0001F680-\U0001F6FF"  # transport & map symbol
                               u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                               u"\U00002500-\U00002BEF"  # chinese char
                               u"\U00002702-\U000027B0"
                               u"\U00002702-\U000027B0"
                               u"\U000024C2-\U0001F251"
                               u"\U0001f926-\U0001f937"
                               u"\U00010000-\U0010ffff"
                               u"\u2640-\u2642"
                               u"\u2600-\u2B55"
                               u"\u200d"
                               u"\u23cf"
                               u"\u23e9"
                               u"\u231a"
                               u"\ufe0f"  # dingbats
                               u"\u3030"
                               "]+", flags=re.UNICODE)
    new6 = re.sub(emoji_pattern, "", new5)
    twitter_data_analysis.loc[index,'text'] =new6
```

In [43]:
```
twitter_data_analysis.head(40)
```

Out[43]:

	text
0	There’s more! We’ll also be testing sharing Tw...
1	Oh snap! Sharing Tweets directly to your Snapc...
2	a stamp of approval
3	2020 in one word
4	this is what cuffing season means to us
5	2 is also cute
6	you're all amazing
7	if you had a twitter before 2020 rt this
8	40404
9	drink water replaced good morning
10	we're taking oomf to the Fleets
11	remember "I dedicate my 500th Tweet to:____"
12	they're tourists
13	proof you're doing it right
14	some of you hating...but we see you Fleeting
15	That thing you didn’t Tweet but wanted to but ...
16	this is art
17	aren't we all six feet
18	this Tweet just graduated with honors
19	saw it, love it, can't wait for the wedding p...
20	
21	breathe
22	apology accepted
23	H2
24	THIRSTY
25	looking hydrated
26	the moon will share
27	bark among the stars
28	rubber ducky knew all along
29	If the moon can hydrate so can you
30	Reading an article before Retweeting it? That’...
31	Hey everyone, we made a temporary change to th...
32	Me seeing my Twitter friends I’ve never met ...
33	

	text
34	dedication
35	not a single person on this app
36	but was it a good Tweet?
37	checks out
38	you forgot one:
39	mutual acknowledgment of good Tweets is frien...

In [44]:
```
twitter_data_analysis.to_csv('cleaning_twitter_data_analysis.csv',index = False)
```

In [45]:
```
import nltk
nltk.download('stopwords' )
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\nidhal\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[45]:
```
True
```

Traitement des tweets: NLP (Natural LanguageProcessing)

On doit proceder a !'analyse du tweet en respectant les differentes etapes du NLP (Natural LanguageProcessing). La bibliotheque a utiliser est NLTK en Python.

In [46]:
```
from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords
ps = PorterStemmer()
stemed_dataset=[]
for i in range(0,twitter_data_analysis.shape[0]):
    stemmed_array=twitter_data_analysis['text'][i].split()
    stemmed=[ps.stem(word) for word in stemmed_array if not word in set(stopwords
    stemmed=' '.join(stemmed)
    stemed_dataset.append(stemmed)
print(stemed_dataset[0:10])
```

```
['there’ more! we’ll also test share tweet IG stori small % keep eye', 'Oh sna
p! share tweet directli snapchat stori easier ever. roll today ios!', 'stamp ap
prov', '2020 one word', 'cuf season mean us', '2 also cute', 'amaz', 'twitter 2
020 rt', '40404', 'drink water replac good morn']
```

In [47]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X=cv.fit_transform(stemed_dataset)
print(X)
```

```
  (0, 2430)     1
  (0, 1581)     1
  (0, 2688)     1
  (0, 1426)     1
  (0, 130)      1
  (0, 2412)     1
  (0, 2134)     1
  (0, 2560)     1
  (0, 1233)     1
  (0, 2307)     1
  (0, 2196)     1
  (0, 1327)     1
  (0, 864)      1
  (1, 2134)     1
  (1, 2560)     1
  (1, 2307)     1
  (1, 1690)     1
  (1, 2206)     1
  (1, 700)      1
  (1, 2207)     1
  (1, 763)      1
  (1, 828)      1
  (1, 2022)     1
  (1, 2480)     1
  (1, 1278)     1
  :     :
  (9669, 1426)  1
  (9669, 222)   1
  (9669, 1691)  1
  (9669, 516)   1
  (9670, 1327)  1
  (9670, 349)   1
  (9670, 2440)  1
  (9670, 181)   1
  (9671, 2688)  1
  (9671, 1426)  1
  (9671, 959)   1
  (9671, 2575)  1
  (9671, 121)   1
  (9671, 2726)  1
  (9671, 515)   1
  (9672, 1055)  1
  (9672, 1162)  1
  (9672, 884)   1
  (9673, 429)   1
  (9674, 2560)  1
  (9674, 1048)  2
  (9674, 1582)  1
  (9674, 1642)  1
  (9674, 142)   1
  (9674, 1089)  1
```

Classification des tweets

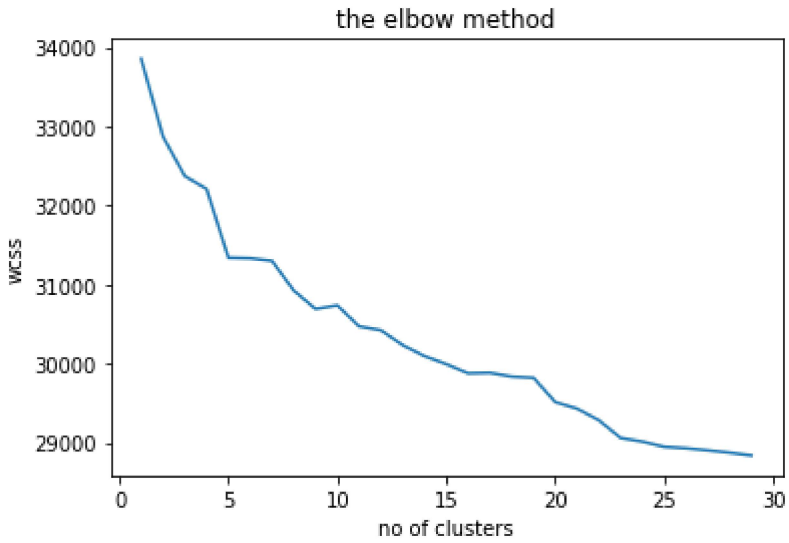
Etant donne un ensemble de tweets, l'objectif est de les resumer sous formes de groupes de sorte a ce que les Tweets qui sont dans le meme groupe soient similaires. Ainsi, l'utilisateur pourra par la suite lire juste un Tweet de chaque groupe (le Tweet qui est le centro"ide de groupes). on a Utiliser l'algorithme K-Means pour classer les Tweets en k classes ,valeurs de k allant de 1 a30 par exemple).

In [48]:
```
from sklearn.cluster import KMeans
wcss=[]
```

In [49]:
```
for i in range(1,30):
    Kmeans=KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10,random_sta
    Kmeans.fit(X)
    wcss.append(Kmeans.inertia_)
```

```
Iteration  3, inertia 33113.887
Iteration  4, inertia 33111.250
Converged at iteration 4: center shift 0.000000e+00 within tolerance 1.242630
e-07
Initialization complete
Iteration  0, inertia 42360.000
Iteration  1, inertia 32947.882
Iteration  2, inertia 32801.420
Iteration  3, inertia 32799.624
Converged at iteration 3: center shift 0.000000e+00 within tolerance 1.242630
e-07
Initialization complete
Iteration  0, inertia 51819.000
Iteration  1, inertia 32580.176
Iteration  2, inertia 32502.427
Iteration  3, inertia 32452.494
Converged at iteration 3: center shift 0.000000e+00 within tolerance 1.242630
e-07
Initialization complete
```

In [50]:
```
import matplotlib.pyplot as plt
plt.plot(range(1,30),wcss)
plt.title('the elbow method')
plt.xlabel('no of clusters')
plt.ylabel('wcss')
plt.show()
```



In [51]:
```
true_k=30
Kmeans=KMeans(n_clusters=true_k,init='k-means++',n_init=1)
Kmeans.fit(X)
```

Out[51]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=30, n_init=1, n_jobs=None, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

La cellule suivante contient les mots cles de chaque cluster

In [52]:

```
print("Top terms per cluster:")
order_centroids = Kmeans.cluster_centers_.argsort()[:, ::-1]
terms = cv.get_feature_names()
for i in range(true_k):
    print("Cluster %d:" % i)
    for ind in order_centroids[i, :10]:
        print(' %s' % terms[ind])
    print()
print("\n")
```

```
Top terms per cluster:
Cluster 0:
 thu
 far
 person
 best
 some
 tweet
 news
 thi
 fix
 flex

Cluster 1:
 twitter
 get
 friend
 follow
 know
```

On a choisir un Tweet par classe comme representant. Les tweets choisis seront les resumes de toutes les informations contenues dans les tweets.

In [53]:
```
result_final_twitter_data_analysis = pd.DataFrame(columns = [ 'text'])
i=0
j=0
while i<30:
    while True:
        Y=cv.transform([stemed_dataset[j]])
        prediction=Kmeans.predict(Y)
        if i == prediction:
            print("tweet of cluster "+str(prediction)+twitter_data_analysis.loc[j
            result_final_twitter_data_analysis.loc[i,"text"] =twitter_data_analys
            j=0
            break
        j+=1
    i+=1
result_final_twitter_data_analysis.to_csv('result_final_twitter_data_analysis.csv
```

```
tweet of cluster [0] This is the best "some personal news" Tweet thus far
tweet of cluster [1]  if you had a twitter before 2020 rt this
tweet of cluster [2]There’s more! We’ll also be testing sharing Tweets to IG St
ories to a small % so keep an eye out
tweet of cluster [3] Our sources say no.
tweet of cluster [4]Hot takes? Always. Hot topics? Brand new.Now you can follow
specific topics to discover the Tweets you care about.
tweet of cluster [5] love the hustle
tweet of cluster [6] Thank you so much for all that you do, we’d love to share
our gratitude. Follow us so we can DM you.
tweet of cluster [7]Reading an article before Retweeting it? That’s growth.Befo
re you Retweet an article, we’ll remind you to read it…
tweet of cluster [8] drink water replaced good morning
tweet of cluster [9] You're about to teach your teachers something
tweet of cluster [10]10 days. 40 cities. Virtual  are happening around the worl
d. Join us and follow the convo!…
tweet of cluster [11]You can Tweet a Tweet. But now you can Tweet your voice!Ro
lling out today on iOS, you can now record and Tweet wi…
tweet of cluster [12] still waiting for someone to DM us the plan
tweet of cluster [13] We're sco pa tu manaaing sco pa tu manaa now? It's over
tweet of cluster [14] we-       we-         we-   we-  we-       we-        we-
tweet of cluster [15]THANK YOU — wow I owe youThank you! — literally, thanksTy
— whateverThx — for nothing
tweet of cluster [16] She loves your photog skills AND your Twitter. Living leg
end.
tweet of cluster [17] *cher voice* IF WE COULD SPEED UP TIME IF WE COULD FIND A
WAY
tweet of cluster [18]  There's power in healing.There's power in our voices.The
re's power in our platforms.We've curated "Thread of Threads:
tweet of cluster [19] chicken noodle soup, chicken noodle soup, chicken noodle
soup with a Tweet on the side
tweet of cluster [20] went from candy corn to candy canes real quick
tweet of cluster [21]Hey everyone, we made a temporary change to the Retweet fu
nction. When you hit the Retweet button, you can either…
tweet of cluster [22] saw it, love it, can't wait for the wedding pics
tweet of cluster [23]The only thing we want going viral is this Tweet
tweet of cluster [24]Give the gift of GIFs. You can now upload your iOS Live Ph
otos as GIFs anywhere you upload photos on Twitter.
tweet of cluster [25] one thought about dogs leads to another thought about dog
s leads to another thought about dogs leads to another thought about dogs
tweet of cluster [26]  “shared a tweet with you” &amp;gt;&amp;gt;&amp;gt;&amp;gt;&amp;gt;
```

```
tweet of cluster [27] Nope! Sending you sweet treats &amp;amp; a DM!
tweet of cluster [28]Oh snap! Sharing Tweets directly to your Snapchat Stories
is now easier than ever. Rolling out today on iOS!
tweet of cluster [29] Reply hazy, try again.
```

conclusion :

on a charge les tweets d'apres l'api de twitter, on les a mis dans le fichier csv twitter_data_analysis. puis on a fait le data cleaning et on a mis le resultat dans le fichier cleaning_twitter_data_analysis.csv.Et enfin on a mis un tweet de chaque cluster dans le fichier result_final_twitter_data_analysis.

lien github (https://github.com/hazbri/projectDataMining/)

In [ ]:

In [ ]: