MASTER INGENIERIE DE DONNEES ET DEVELOPPEMENT LOGICIEL – TA

MINI PROJET
Sentiment Analysis & Text Generation
« NATURAL LANGUAGE PROCESSING»

REALISE PAR :

AMEZIANE MOHAMED

ET-TABTI YOUNESS

ENCADRE PAR :

MAHMOUDI ABDELHAK

Année Universitaire : 2020/2021

PLAN

PREMIÈRE PARTIE : PROBLEMATIQUE

DEUXIÈME PARTIE : CONFIGURATION DU MODELE

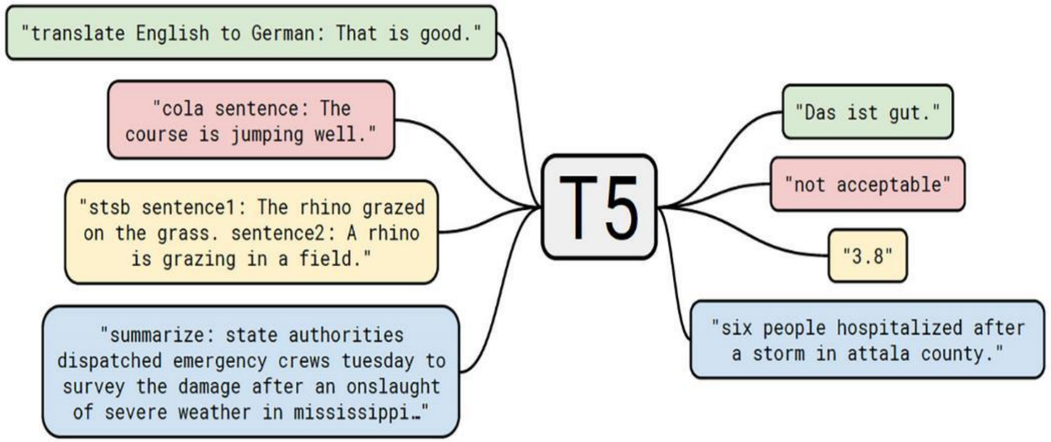
TROISIEME PARTIE : RESULTATS

PREMIÈRE PARTIE :

PROBLEMATIQUE

PROBLÉMATIQUE

Le **NLP** pour **N**atural **L**anguage **P**rocessing ou Traitement Numérique du Langage est une discipline qui porte essentiellement sur la compréhension, la manipulation et la génération du langage naturel par les machines. Ainsi, le NLP est réellement à l'interface entre la science informatique et la linguistique. Il porte donc sur la capacité de la machine à interagir directement avec l'humain.



PROBLÉMATIQUE

À propos de ce projet

Ce mini projet présente deux partie pour capturer les sentiments des tweets lors d'événements significatifs. Dans ce cas particulier, nous avons travaillé sur la langue espagnole (espagnol colombien). nous nous concentrerons sur le jour des élections colombiennes 2019 et avons divisé notre analyse en deux parties :

- **La première partie** : on a utilises des techniques d'apprentissage automatique "Machine Learning" pour classer les tweets en fonction de leur sentiment comme positif ou négatif.

- **La deuxième partie** : La génération de texte à l'aide de réseau de neurones artificiels "Deep LSTM". Les mêmes tweets ont été introduits dans le réseau afin de produire un texte qui résume/capture le contexte autour des mots-clés contenus dans les tweets.

PROBLÉMATIQUE

L’utilisation de ce projet

Nous pensons qu'il existe un besoin croissant pour les entreprises de médias de capturer et de partager en temps réel le sentiment en ligne concernant des sujets spécifiques.
Pour ce projet particulier, nous nous concentrons sur l'arène politique étant donné que le **Maroc** traverse actuellement ses élections présidentielles et que le principal cas d'utilisation que nous imaginons est la couverture en direct des sentiments en ligne lors des débats présidentiels.

Nous espérons qu'un outil comme celui-ci (et après avoir résolu d'autres types de problèmes, comme les bot-armées des réseaux sociaux), permettra plus de transparence et une meilleure représentation des opinions du public.

- **Classification des sentiments** étiquetés L'attitude du public actuel est-elle positive ou négative ?
- **Génération de texte** Donnerait une idée de ce que le public associe à un certain mot clé.

DEUXIÈME PARTIE :

CONFIGURATION DU MODELE

CONFIGURATION DU MODELE

Chargement du modèle :
Tout d'abord, nous avions importé les packages requis :

```
import pandas as pd
import numpy as np

#NLTK
import nltk
from nltk import word_tokenize, WordPunctTokenizer, regexp_tokenize
from nltk import word_tokenize, WordPunctTokenizer, regexp_tokenize

#Plotting
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS

#Keras
from numpy import array
from keras.preprocessing.text import Tokenizer
from tensorflow.keras.utils import to_categorical
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers import Embedding
```

CONFIGURATION DU MODELE

Ensuite, nous avions chargé les données et le traitée et l'afficher

```
import xml.etree.ElementTree as ET
tree = ET.parse('DATA/general-Tweet.xml')
root = tree.getroot()
```

```
train_set = pd.DataFrame({'tweet_id':[],'tweetText':[],'polarity_value':[],'polarity_type':[],'topic':[]})
row=0
for tweet in root:
    tweet_id = 'ID:'+tweet.find('tweetid').text
    #user = tweet.find('user').text
    tweetText = tweet.find('content').text
    lang = tweet.find('lang').text
    polarity_value = tweet.find('sentiments').find('polarity').find('value').text
    polarity_type = tweet.find('sentiments').find('polarity').find('type').text
    topic = tweet.find('topics').find('topic').text

    if lang == 'es':
        train_set.loc[row] = [tweet_id,tweetText,polarity_value,polarity_type,topic]
        row+=1
```

```
train_set['set'] = 'train'
train_set.head(10)
```

```
              tweet_id                                          tweetText  \
0  ID:142389495503925248        Salgo de #VeoTV , que día más largoooooo...
1  ID:142389933619945473  @PauladeLasHeras No te libraras de ayudar me/n...
2  ID:142391947707940864                            @marodriguezb Gracias MAR
```

CONFIGURATION DU MODELE

Colombian Tweets mentioning candidates

```
tweets = pd.read_csv('tweets_mentioning_candidates.csv')
tweets['set'] = 'test'
tweets['polarity_value'] = np.NaN
tweets.shape
```

```
(30000, 18)
```

```
processed_tweets= pd.concat([pd.DataFrame({'tweetID':tweets.tweetID, 'tweetText':tweets.tweetText,'polarity_value':twee
                             pd.DataFrame({'tweetID':train_set.tweet_id, 'tweetText':train_set.tweetText, 'polarity_va
processed_tweets['processed_tweet'] = processed_tweets.tweetText
processed_tweets.sample(4)
```

```
                      tweetID  \
30922   ID:151378281772490753
7087    ID:1008500006595252225
32734   ID:167750882333696000
29985   ID:1008486856529448962

                                          tweetText polarity_value  \
30922      Intentemos disfrutar, al margen de todo esto.              P
7087   Espero que calle bocas el candidato de centro ...            NaN
32734  No se donde tengo las llaves, ni la agenda ni ...              N
```

CONFIGURATION DU MODELE

Le changement des Tags et les hashtags et les liens

```
import re
hash_regex = re.compile(r"#(\w+)")
hstgs = []
def hash_repl(match):
    _ = '__HASH_'+match.group(1).upper()
    hstgs.append(_)
    return _
```

Change Usernames to _

```
user_regex = re.compile(r"@(\w+)")
usr_names = []
def user_repl(match):
    _ = '__user_'+match.group(1).upper()
    usr_names.append(_)
    return _
```

Change URLs to _

```
url_regex = re.compile(r"(http|https|ftp)://[a-zA-Z0-9\./]+")
def url_repl(match):
    return '__URL_'
```

CONFIGURATION DU MODELE

Remplacer les emoticons avec ca unicode valeur

```
# Emoticons
emoticons = \
    [   # For __EMOT_SMILEY
        (' __emoji: U+1F601',   [':-)', ':)', '(:', '(-:', ] )  ,\
        # for __EMOT_LAUGH
        (' __emoji: U+1F923',       [':-D', ':D', 'X-D', 'XD', 'xD', ] )    ,\
        # For __EMOT_LOVE
        (' __emoji: U+2764',        ['<3', ':\*', ] )   ,\
        # For __EMOT_WINK
        ('__emoji: U+1F609',        [';-)', ';)', ';-D', ';D', '(;', '(-;', ] ) ,\
        # For __EMOT_FROWN
        (' __emoji: U+2639',        [':-(', ':(', '(:', '(-:', ] )  ,\
        # For __EMOT_CRY
        (' __emoji: U+1F622',       [':,(', ':\'(', ':"(', ':(('] ) ,\
    ]
def escape_paren(arr):
    return [text.replace(')', '[)}\]]').replace('(', '[({\[') for text in arr]
def regex_union(arr):
    return '(' + '|'.join( arr ) + ')'
emoticons_regex = [ (repl, re.compile(regex_union(escape_paren(regx))) ) for (repl, regx) in emoticons ]
```

```
# Test
text = "This is a text with one emoticon :) and another :("
for (repl, regx) in emoticons_regex :
    text = re.sub(regx, ' '+repl+' ', text)

print(text)
```

```
This is a text with one emoticon   __emoji: U+1F601  and another   __emoji: U+2639
```

TROISIEME PARTIE :

RESULTATS

RESULTATS FINALS

Afficher les tweet positive

Positive Tweets:

```
positives = NB_results.loc[NB_results.sentiment == 'P']
sample_size = 10
for tweet in positives.tweet.sample(sample_size):
    print(tweet)
```

MagicPython

```
('@kienyke @A_OrdonezM @IvanDuque Gracias Dr. Ordoñez, usted fue fundamental para esta victoria.', nan)
('"Una sociedad libre de miedo, es una sociedad en verdadera paz 🕊"@IvanDuque', nan)
('@IvanDuque ¡Felicitaciones @IvanDuque! ¡Hoy gana Colombia! ¡Hoy gana la Democracia! Estirpando ese cáncer llamado Castro-Chavismo. Colombia hacia el progreso...éxitos.', nan)
('#OjaláDuque adhiera a su plan de gobierno las mejores propuestas de @sergio_fajardo @petrogustavo y las priorice. Energías limpias ,educación, regulación a la banca y lucha por la Paz!', nan)
('@petrogustavo 😢estoy triste pero quiero seguir pensando en Colombia Humana como alternativa gracias por la ilusión de cambiar esta sociedad', nan)
('@Darojasti @merchoblack @mariolopez1959 @petrogustavo @IvanDuque Preparece para lo que viene, yo me voy del país', nan)
('@liliantintori @IvanDuque Grande Lilian!!', nan)
('Cumplimos con la democrácia\nY ganó Colombia co. @IvanDuque 👏🏻', nan)
('Si cayó Petro, también  lo puede hacer el Peje...Felicitaciones #Colombia y a su nuevo presidente @IvanDuque.', nan)
('@IvanDuque @JuanManSantos Empieza despues de que Uribe le escriba el guión de todo lo que tiene que hacer', nan)
```

RESULTATS FINALS

Afficher les tweet Négative

Negative Tweets:

```
negatives = NB_results.loc[NB_results.sentiment == 'N']
sample_size = 10
for tweet in negatives.tweet.sample(sample_size):
    print(tweet)
```

MagicPython

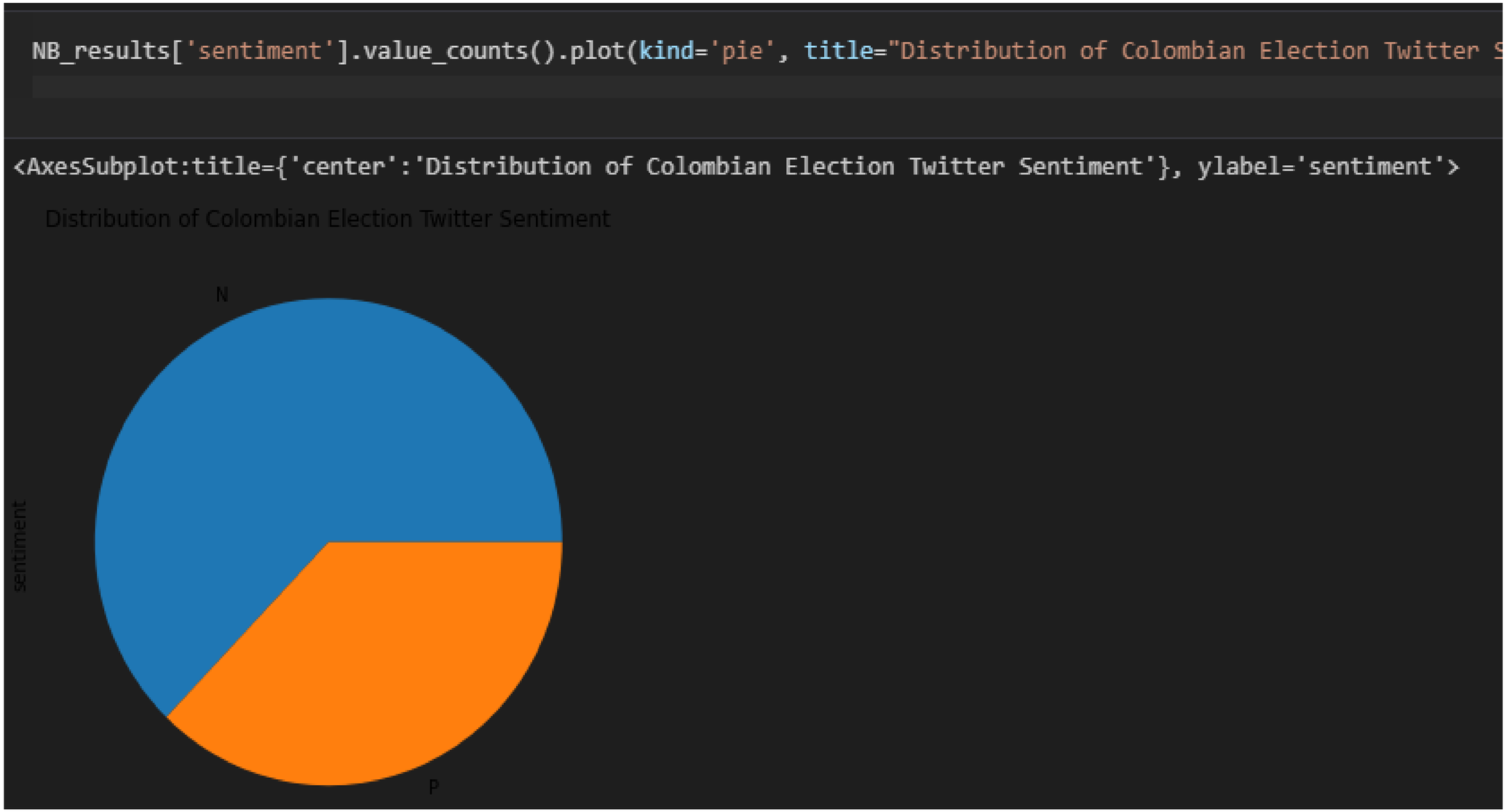
```
('@petrogustavo perdón en nombre de los 10 Millones de gente que le tiene miedo al cambio. La Colombia Humana sigue en pie.', nan)
('@DeLaCalleHum @IvanDuque @petrogustavo Me decepcionó cucho.', nan)
('@NoticiasCaracol  no veo a @ClaudiaLopez al lado de @petrogustavo en la transmisión? ya se voltio?', nan)
('@EPN @IvanDuque Y yo Felicito cordialmente al sr Cuauhtemoc Blanco por haberte dado tu madriza en un ejemplar descuido de su parte .... ah no.. pense que eras Faitelson.. pero si mereces tu madriza', nan)
('@OliverLopezCano @VandaloNic @petrogustavo No seas estúpido que aquí no hay socialismo, lo que hay un Estado Fascista, estúpido el que cree eso no jodan, en Nicaragua lo que hay es sistema capitalista de mierda.', nan)
('@paula_g_g @IronHinds @CamiloS12979723 @petrogustavo Si ya que le gusta los gobiernos socialista venga a Venezuela y vivan en el socialismo puro hambre y miseria.', nan)
('Hace algunos días atrás lo había dicho, si duque ganaba significará para mí la total negacia de tener un presidente #DuqueNoEsMiPresidente @IvanDuque  sos lo peor que le puede pasar a #Colombia', nan)
('@liliantintori @IvanDuque Señora, estudie historia de Colombia, @IvanDuque significa muerte y corrupción!', nan)
('8 millones q creemos q hay continuar apoyando la #paz https://t.co/6SyAg58Uie', nan)
('@ARENABURSATIL @petrogustavo Ehhh pero que terquedad, que Antioquia fue determinante por su gran cantidad de votantes. https://t.co/Xoi9daIb2o', nan)
```

RESULTATS FINALS

Barplot pour afficher les pourcentage des tweet positive et négative

```
NB_results['sentiment'].value_counts().plot(kind='pie', title="Distribution of Colombian Election Twitter S
```

```
<AxesSubplot:title={'center':'Distribution of Colombian Election Twitter Sentiment'}, ylabel='sentiment'>
```



RESULTATS FINALS

Et pour générer un texte en doit saisir une mot dans input et cliquer sur entrer
et voila le txt et générer

Generate my text !

```
allWordExceptStopDist.most_common(10)
```

```
[('userivanduque', 7),
 ('mejor', 1),
 ('discurso', 1),
 ('cuan', 1),
 ('usermanuelrosalesg', 1),
 ('usermluciaram', 1),
 ('usergloserna', 1),
 ('userfransupelano', 1),
 ('usericolombianoo', 1),
 ('userelpatriota', 1)]
```

```
string = input("What do you want to test? ")
print(generate_seq(model, tokenizer, string , 5)) #give me 5 words in a sequence
```