Group Name: Sedats

Name: Sedat CAN (one-member group)

Email: sedatcancan1@gmail.con

Country: Germany

Specialization: Data Science, NLP, Data Analyst

Problem Description

Hate speech is one of the biggest problems that social media users face on daily basis. Especially, Twitter users are harassed and discriminated against because of their religion, ethnicity, nationality, race, color, ancestry, sex, etc.

The hate speech detection model would evaluate and classify whether tweets include hate speech and filters these tweets to offer a better user experience.

Data

```
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      17197 non-null  int64
 1   tweet   17197 non-null  object
 2   label   17197 non-null  int64
dtypes: int64(2), object(1)
memory usage: 403.2+ KB
None
NULLS?
id       0
tweet    0
label    0
dtype: int64
```

The data has 10197*3 values. It does not have any null value. So it does not require the deletion of any data. The data has three columns: id, tweet and label.

1- 'id' is the number of row.

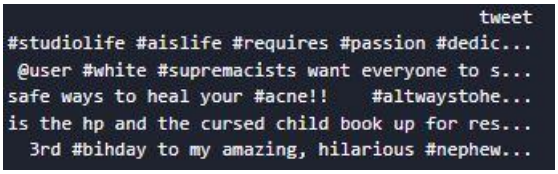
2- 'tweet' contains the user, text and hashtag. There are also some special characters such as '(' '&', which require a special cleaning process.

To clear tweet I use the code below.

```
df['cleartweet'] = df['tweet'].apply(lambda  tweet: re.sub("(@[A-Za-z0-
9]+)|(#)|(RT[\s]+)|(https?:\/\/\S+)|([^a-zA-Z0-9 -])", "", tweet))
```

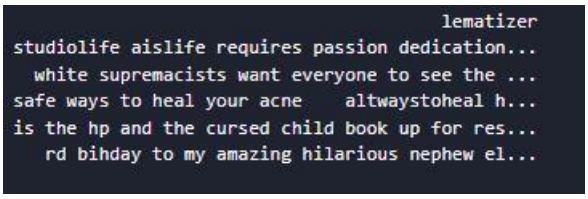
Yet twitter text also needs lemmatization which give an actual word of the language. I use "nltk.stem import WordNetLemmatizer".

Tweet before processed



	tweet
0	#studiolife #aislife #requires #passion #dedic...
1	@user #white #supremacists want everyone to s...
2	safe ways to heal your #acne!! #altwaystohe...
3	is the hp and the cursed child book up for res...
4	3rd #bihday to my amazing, hilarious #nephew...

Processed tweet data



	lematizer
0	studiolife aislife requires passion dedication...
1	white supremacists want everyone to see the ...
2	safe ways to heal your acne altwaystoheal h...
3	is the hp and the cursed child book up for res...
4	rd bihday to my amazing hilarious nephew el...

3-‘label’ the data has two label: ‘0’ and ‘1’. ‘0’ refer to tweet that contain hate speech.