

Documentație proiect ML Fundamentals

Task 1 - Twitter sentiment analysis

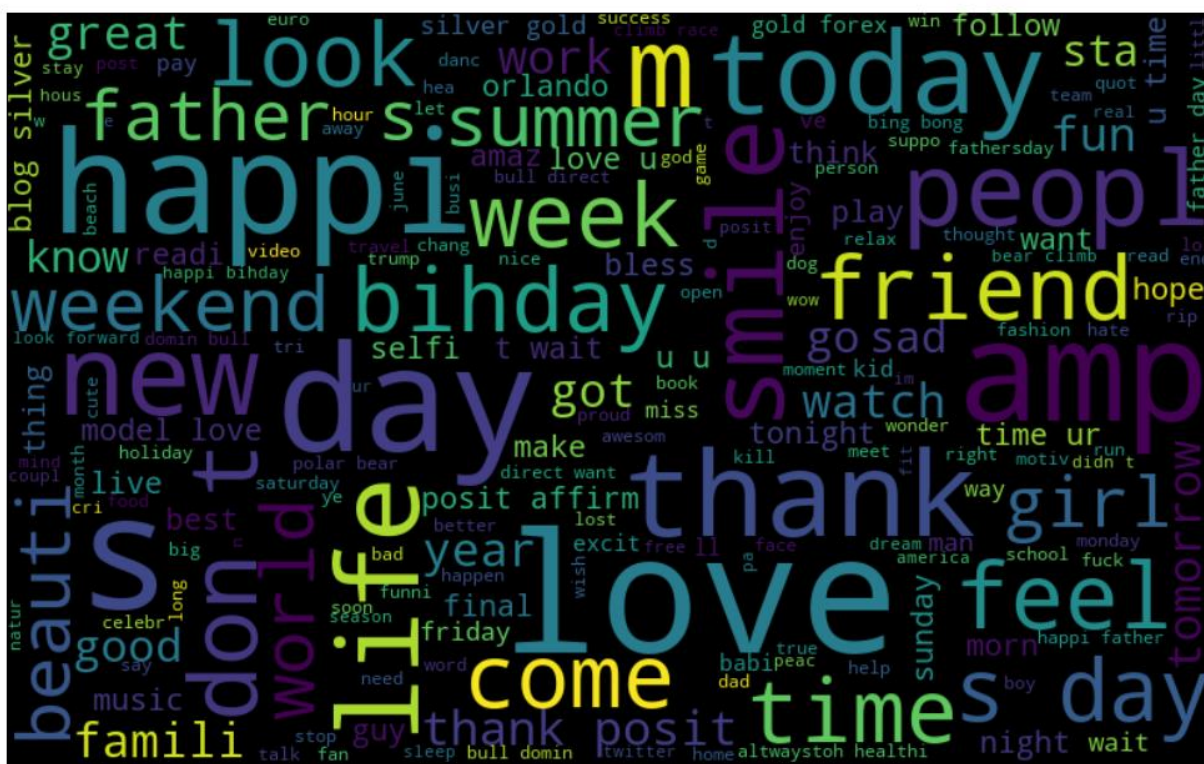
Tița Andrei-Petruț

Februarie 2021

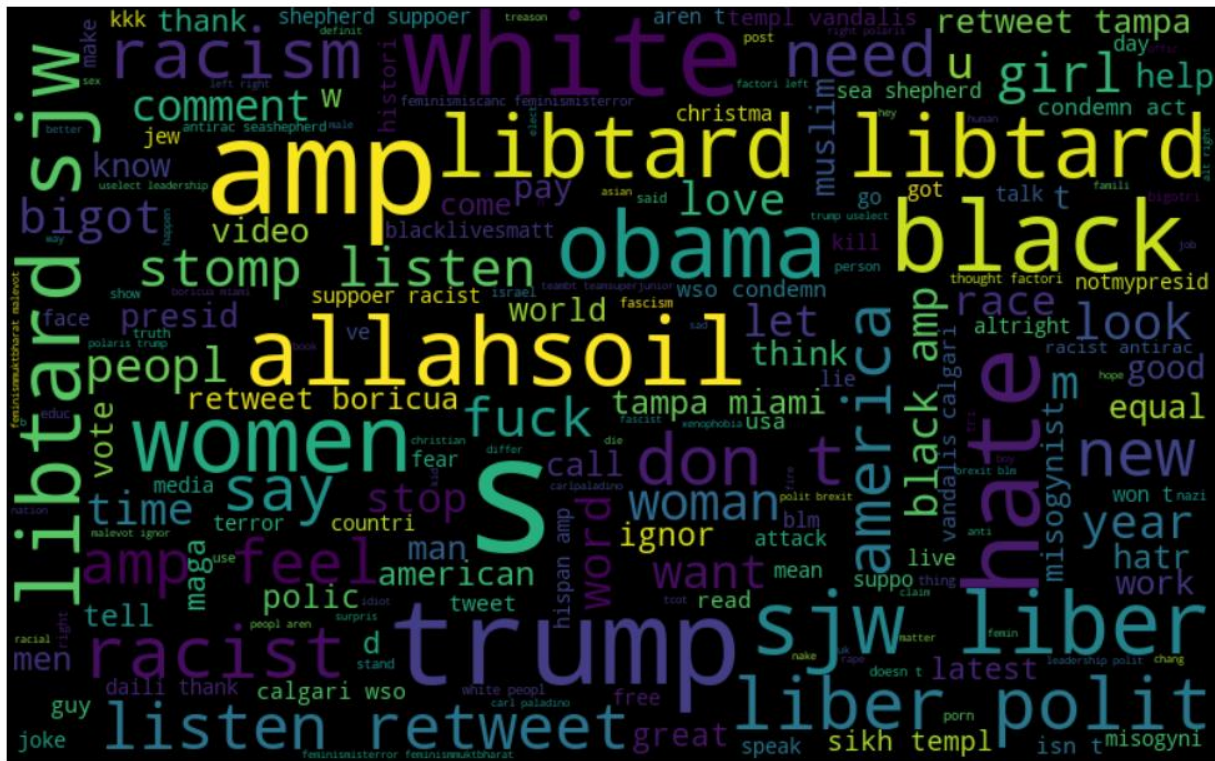
Primul pas este de a încărca datele din fișierele .csv. Pentru asta am făcut metoda `get_data(filename)`. Am lucrat pe modelul de la detectorul de spam din cursul de NB.

Am calculat numărul de exemple pozitive și negative și am obținut 93% exemple pozitive și 7% negative deci o împărțire a datelor întâmplătoare nu ar fi prea bună pentru că setul de date este debalansat. Așa că voi folosi metoda de la Ham/Spam Detector de împărțire a datelor cu ShuffleStratifiedSplit.

Apoi am vrut să văd care cuvinte se repetă cel mai mult în tweeturile normale și în cele cu hate. Pentru asta am folosit WordCloud și am obținut, pentru tweeturile normale:



Iar pentru cele de cu hate:



Observăm ca aceste cuvinte sunt destul de diferite deci probabil modelul o să reușească să prezică destul de bine în funcție de frecvența acestora.

Acum a ramas doar să iau un CountVectorizer care calculează frecvențele cuvintelor date și să le dau ca parametru pentru antrenarea modelului multinomialNB.

Scorurile obtinute sunt (pentru $\alpha=1$):

```

Accuracy score of test 0.9551071484436102
      precision    recall  f1-score   support

     0       0.96      0.99      0.98       5945
     1       0.79      0.49      0.61        448

 accuracy          0.96       6393
 macro avg       0.87      0.74      0.79       6393
weighted avg       0.95      0.96      0.95       6393

```

Acuratețea este foarte bună, mai trebuie lucrat la scorurile pentru clasa '1' adică tweeturile cu hate. Singura variabila pe care o puteam modifica era alpha de la MultinomialNB așa că am dat valori în mai multe încercări și am reușit să mai cresc recall-ul de la clasa 1:

```
Accuracy score of test 0.9601126231816048
```

	precision	recall	f1-score	support
0	0.98	0.98	0.98	5945
1	0.72	0.71	0.72	448
accuracy			0.96	6393
macro avg	0.85	0.85	0.85	6393
weighted avg	0.96	0.96	0.96	6393

Am considerat destul de bune rezultatele și m-am oprit aici cu acest task.