Documentație proiect ML Fundamentals
Task 1 - Twitter sentiment analysis

Tița Andrei-Petruț

Februarie 2021

Procurarea și prelucrarea datelor

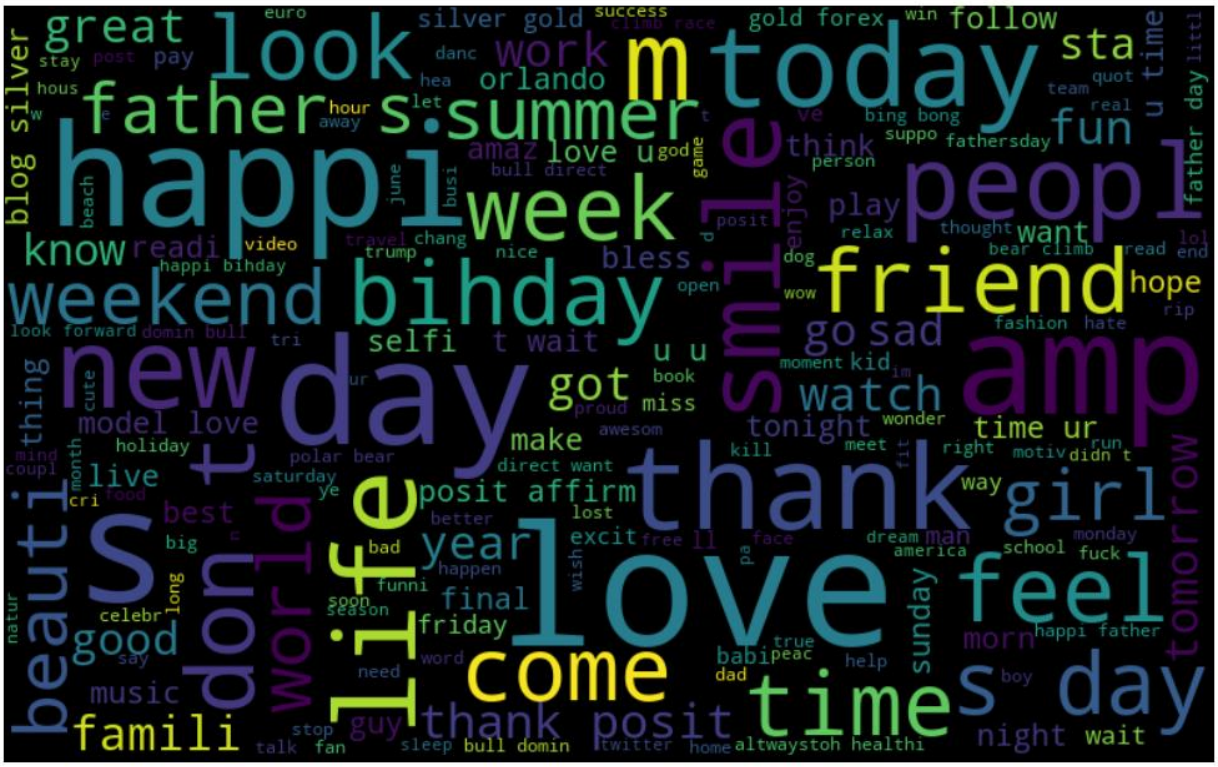
Primul pas este de a încărca datele din fișierele .csv. Pentru asta am făcut metoda get_data(filename). Am lucrat pe modelul de la detectorul de spam din cursul de NB.

Datele primite sunt deja împărțite în train și test însă datele din test nu vin însoțite de label-uri așa că m-am gândit ca prima dată să iau o parte din datele din train și să văd cât de bine va merge modelul pe ele pentru ca mai apoi să fie aplicat pe datele de test reale. Astfel pot avea o imagine despre modelul meu.

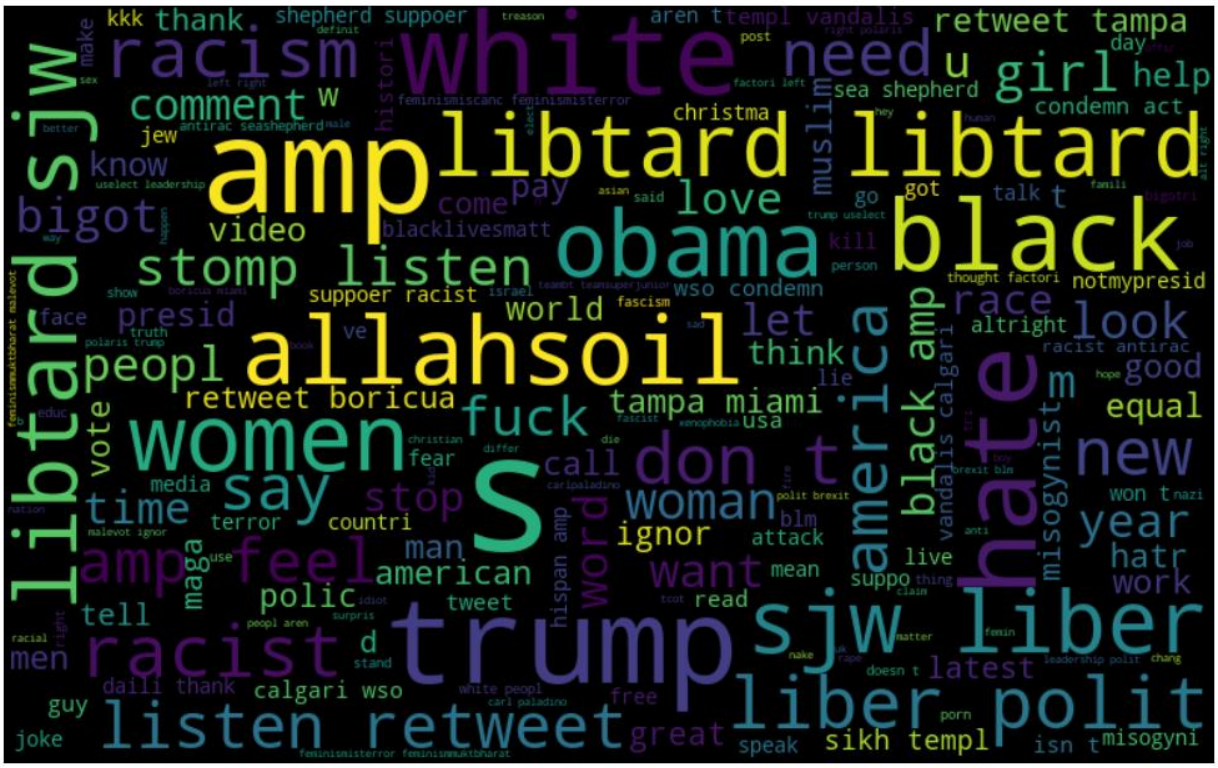
Am calculat numărul de exemple pozitive și negative și am obținut 93% exemple pozitive și 7% negative deci o împărțire a datelor întamplătoare nu ar fi prea bună pentru că setul de date este debalansat. Așa că voi folosi metoda de la Ham/Spam Detector de împărțire a datelor cu ShuffleStratifiedSplit.

Urmează să prelucrez datele citite. Inițial mă gândisem să folosesc Tweet tokenizer pentru prelucrarea datelor însă mi-am dat seama că nu mă ajută așa mult (era ok doar la scoaterea @user). Așa că am ales să iau fiecare tweet în parte și să aplic urmatoarele transformări pe el: ștergerea @user, aducerea la lowercase pentru că nu sunt relevante literele mari/mici, apoi am luat fiecare cuvânt prin split și am verificat să nu fie în lista de stopwords din spacy și am eliminat prefixele/sufixele folosind procedeul de steamming. In final am reunit cuvintele pentru a forma tweet-ul.

Apoi am vrut să văd care cuvinte se repetă cel mai mult în tweeturile normale și în cele cu hate. Petru 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 frecventele cuvintelor date și să le dau ca paramentru pentru antrenarea modelului multinomialNB.

Scorurile obtinute sunt:

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 foare bună, mai trebuie lucrat la scorurile pentru clasa ‘1’ adică tweeturile cu hate.