ΕΡΓΑΣΙΑ 1

Καλλιοπη Μυρσιλιδη
ΑΜ: 1115201400122

Ποταμιας Σωκρατης
ΑΜ: 1115201400166

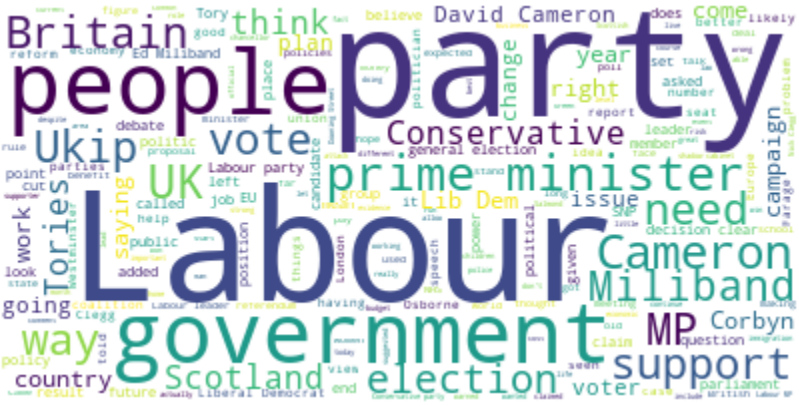
WORDCLOUD

Αχικά για το wordcloud μαζευουμε όλα τα κείμενα σε ένα string και το περναμε μέσα στο
wordcloud.Χρησιμοποιήσαμε μερικά δικά μας words που καναμε union μετα stop words, διότι εμφανίζονταν σε πολλα κειμενα χωρίς να προσφέρουν κάποια ιδιατερη σημασία στο wordcloud.

Business:

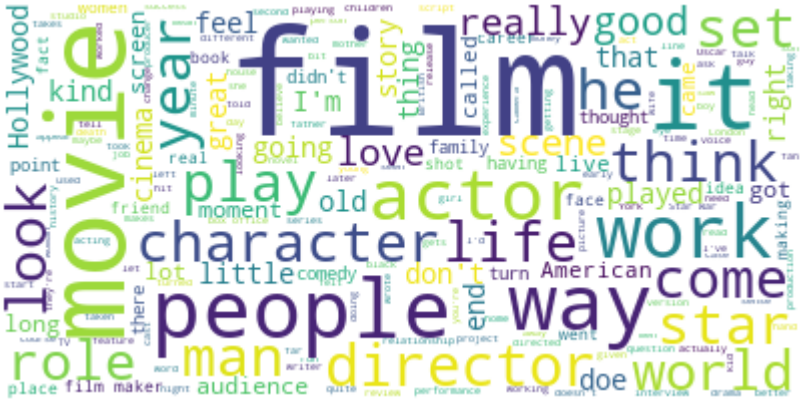
Technology:

Politics:



Football:

Film:



Classification

Υλοποιησαμε τους SVM, Nayve, RandomForest και τον KNN.

Nayve:

Γενικα ητανε ο πιο αποτελεμσατικός αλγόριθμος, τα metrics του ήταν καλυτερα σε σχεση με αυτά των αλλων δυο. Χρησιμοποιησαμε τον MULTINOMIALNB εφόσον δεν υπήρχαν αρνητικές τιμες στο vector.

Statistic_Metrics	Naive Bayes
Accuracy	0.9572799877
Presicion	0.9529742967
Recall	0.9545273306
F_Measure	0.9536337423

SVM:

Παρατηρησαμε οτι καλυτερος σε γενικες γραμμες ητανε ο linear γι'αυτό και τον χρησιμοποιησαμε.

Statistic_Metrics	SVM
Accuracy	0.9532033461
Presicion	0.949810154
Recall	0.9490754563
F_Measure	0.9492863842

RandomForest:

Ήταν ενας μετριος classifier, κατωτερος απο τον SVM(linear) με 8-10 estimators πηγαινε καλα. (metrics εχουνε το macro σαν average ημασταν αναμεσα σε αυτο και το weighted)

Statistic_Metrics	Random Forest
Accuracy	0.9326591336
Presicion	0.9291214515
Recall	0.9239418083
F_Measure	0.9260794253

KNN:

Προσπαθήσαμε να τον υλοποιήσουμε δοκιμάσαμε αρκετούς αλγορίθμους αλλα δε μας ετρεξαν.Στον κώδικα υπάρχει ο KNN αλλά δεν βγάζει αποτελέσματα.

Prediction του test_set(πρωτα 40)

ID	Predicted_Category
2	Politics
10	Technology
25	Technology
28	Business
29	Business
33	Business
34	Business
37	Technology
39	Technology
40	Technology
44	Business
56	Business
66	Technology
75	Business
77	Business
85	Business
89	Business
92	Technology
107	Technology
112	Technology
115	Technology
139	Business
140	Business
152	Business
156	Technology
162	Technology
167	Business
170	Business
171	Technology
173	Politics
182	Technology
186	Football
193	Technology
196	Business
210	Business
218	Football
224	Business
227	Technology
230	Politics