2η Προγραμματιστική εργασία

Μάθημα: Τεχνητή Νοημοσύνη

Ακαδημαϊκό έτος: 2020–21

ΓΕΩΡΓΙΟΣ ΚΟΤΣΙΦΟΣ-3190093

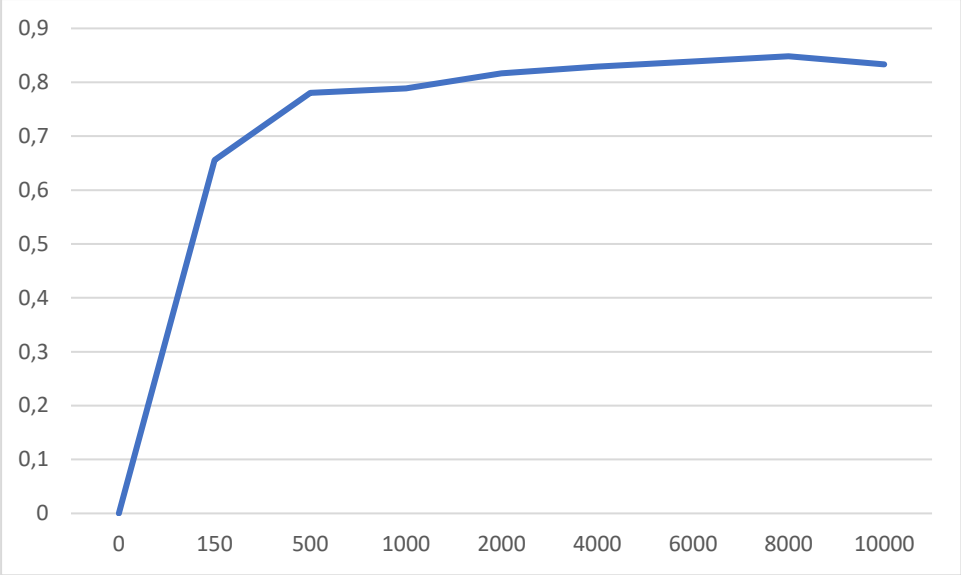
ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΑΡΚΟ-3190112

Έπειτα από αρκετές δοκιμές καταλήξαμε ότι και οι δύο αλγόριθμοι (Naive Bayes, Logistic Regression) τρέχουν σε πολύ καλό βαθμό με υπερπαραμέτρους: m=1000 και n=50 δηλαδή παίρνοντας τις 1000 συχνότερες λέξεις και από αυτές παραλείποντας τις 50 πιο συχνές. Οι δοκιμές αυτές έγιναν κυρίως διαισθητικά αφού π.χ. με λίγο μεγαλύτερο m ο Naive Bayes έβγαλε ελάχιστα καλύτερο αποτέλεσμα, αλλά η διαφορά είναι τόσο αμελητέα που αποφασίσαμε να κρατήσουμε «χοντρικά» αυτές τις υπερπαραμέτρους και για τους δύο αλγορίθμους. Επίσης χρησιμοποιήσαμε την train_test_split για να χωρίσουμε τα δεδομένα μας σε αντίστοιχα train και test data με test_size=0.2 δηλαδή 1/5 των δεδομένων. Για τον Logistic Regression μας φάνηκε καλύτερο το lr=0.05 και iters=1000. Τέλος όλες οι μεταβλητές και οι παράμετροι κάθε αλγορίθμου καθώς και αυτές που χρησιμοποιούνται κατά την φόρτωση των δεδομένων μπορούν να αλλαχθούν. P.S υπάρχει πιθανότητα εν βάλετε πολύ λίγα data (<100) να έχει θέμα το πρόγραμμα και να κάνει σε κάποιο σημείο division με 0.

Αλγόριθμος Naïve Bayes:

- Πίνακας Accuracy:

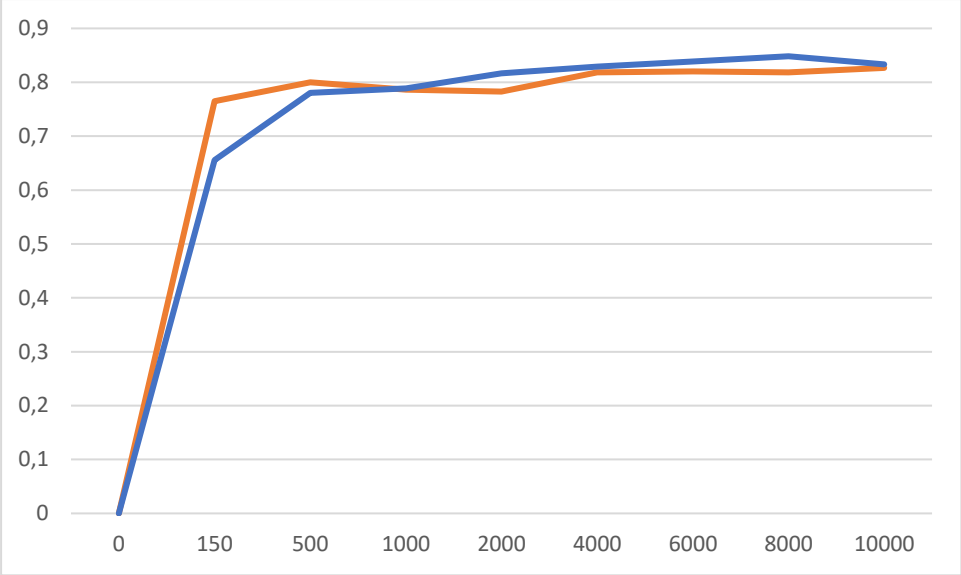
Data Size	**Accuracy**
150	0.7
500	0.8
1000	0.795
2000	0.8125
4000	0.8025
6000	0.840833
8000	0.83
10000	0.8185



- Πίνακας Precision:

Data Size	Positive Precision	Negative Precision
150	0.8333333333333334	0.6666666666666666
500	0.7884615384615384	0.8125
1000	0.8125	0.7727272727272727
2000	0.7810945273631841	0.8442211055276382
4000	0.7910447761194029	0.8140703517587939
6000	0.839344262295082	0.8423728813559322
8000	0.8109339407744874	0.853185595567867
10000	0.8077651515151515	0.8305084745762712

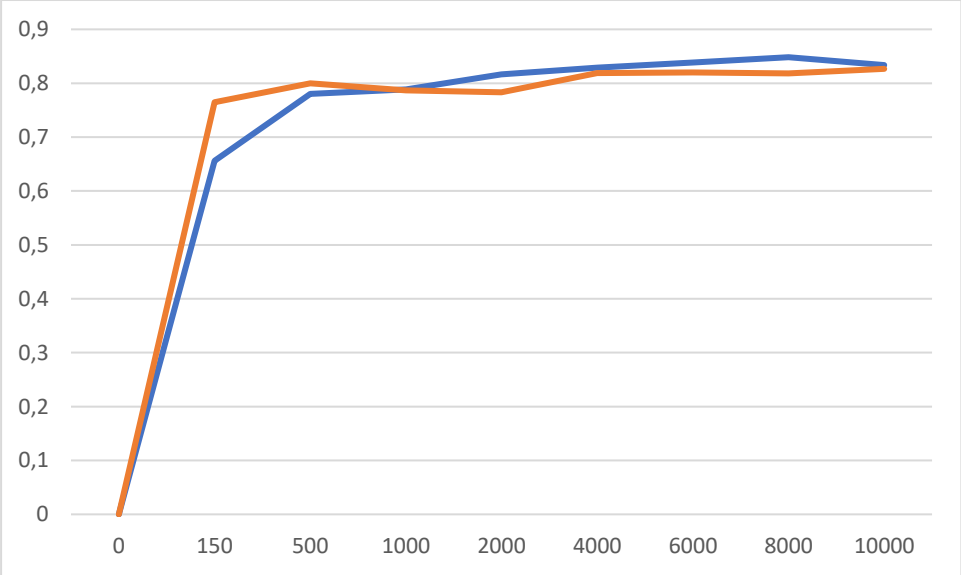
Ο μπλε είναι ο Positive, ο πορτοκαλί είναι ο Negative:



- Πίνακας Recall

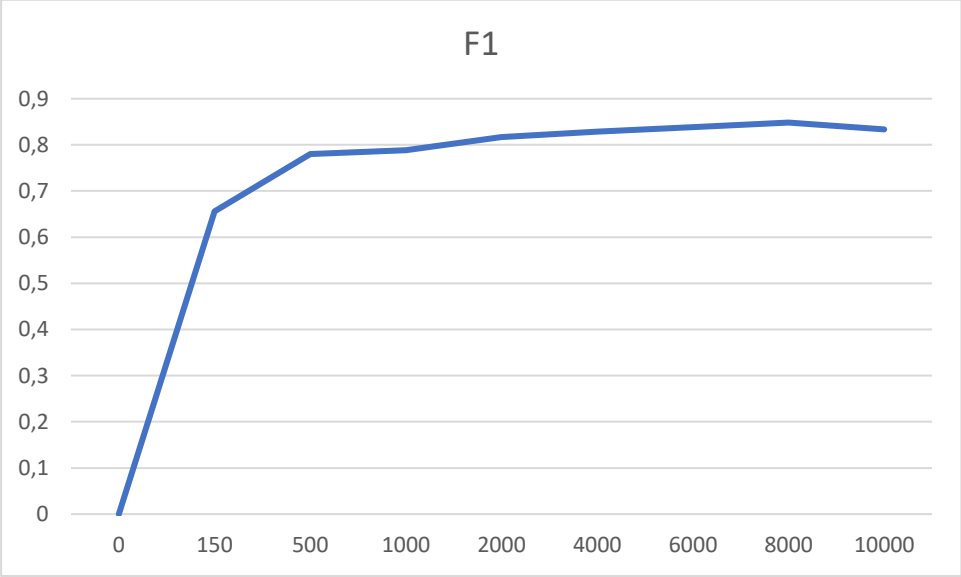
Data Size	Positive Recall	Negative Recall
150	0.38461538461538464	0.9411764705882353
500	0.82	0.78
1000	0.8198198198198198	0.7640449438202247
2000	0.8351063829787234	0.7924528301886793
4000	0.8112244897959183	0.7941176470588235
6000	0.8462809917355372	0.8352941176470589
8000	0.8704156479217604	0.7877237851662404
10000	0.8420533070088845	0.7943262411347518

Ο μπλε είναι ο Positive, ο πορτοκαλί είναι ο Negative:



- Πίνακας F1

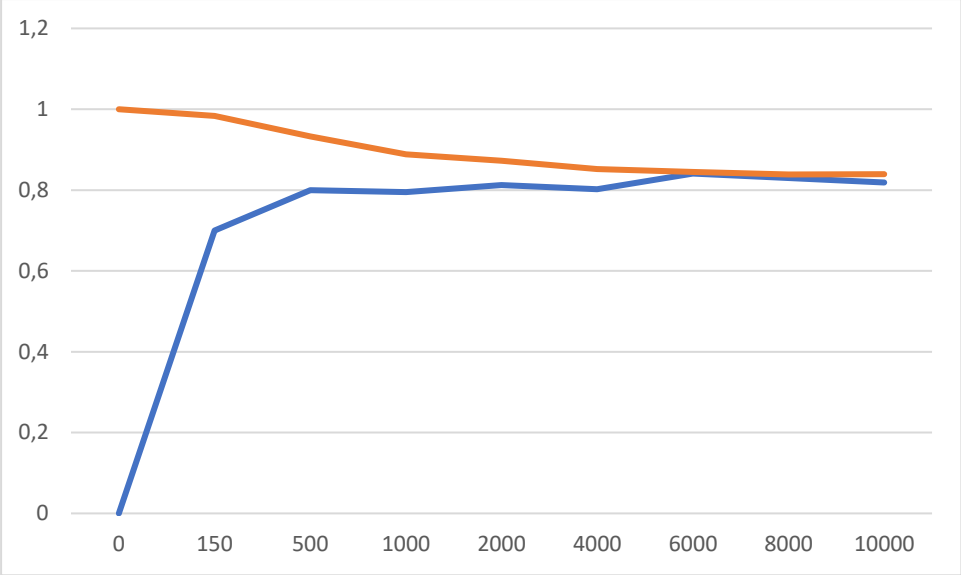
Data Size	**F1**
150	0.7037630104083268
500	0.8002403124061279
1000	0.7922728626436518
2000	0.8132183246527267
4000	0.802614312170327
6000	0.84082306175885
8000	0.8305620512944251
10000	0.8186630196724108



- Train και Test Curves:

Data Size	Train Accuracy
150	0.9833333333333333
500	0.9325
1000	0.88875
2000	0.8725
4000	0.851875
6000	0.8445833333333334
8000	0.83859375
10000	0.839125

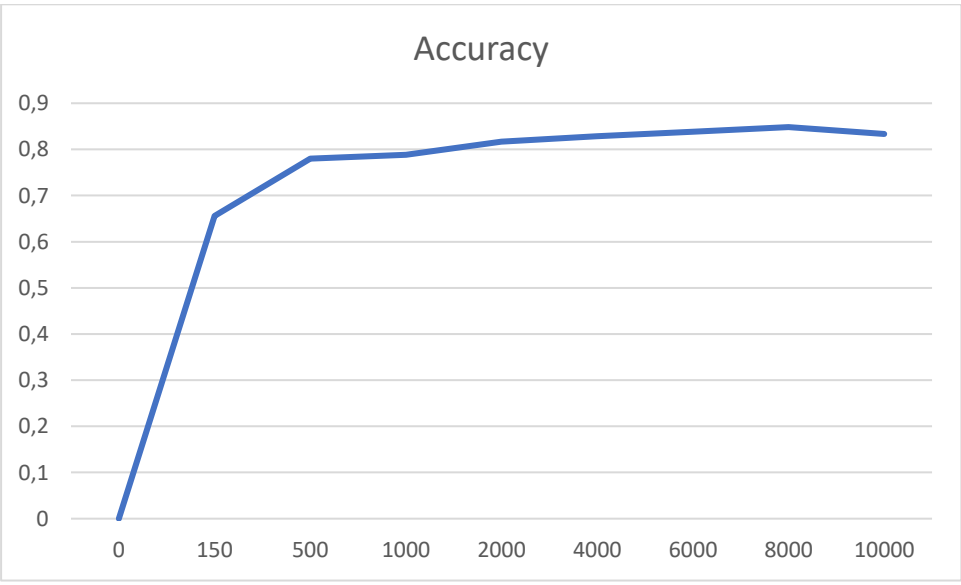
Ο μπλε είναι ο Test, ο πορτοκαλί είναι ο Train:



Αλγόριθμος Logistic Regression:

- Πίνακας Accuracy:

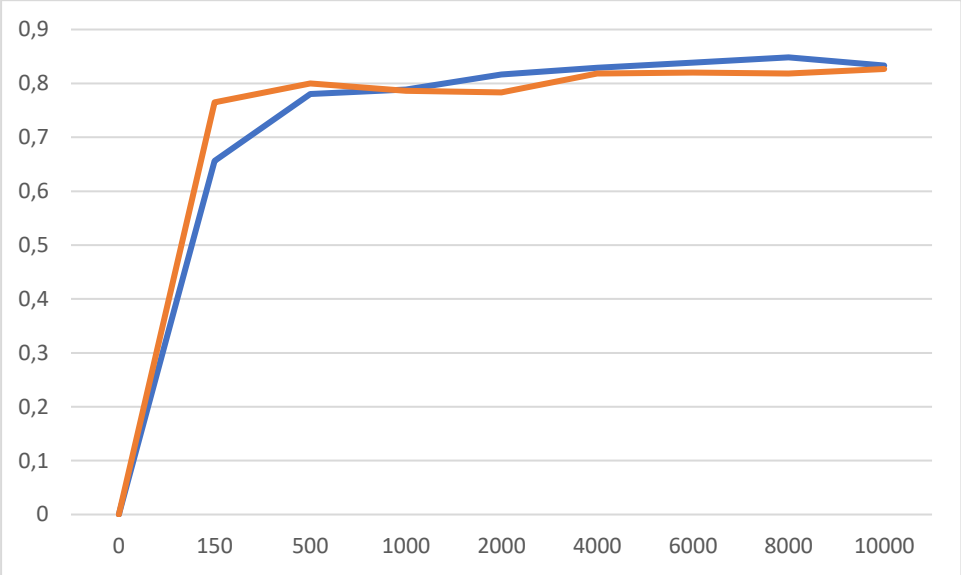
Size Data	Accuracy
150	0.6666666666666666
500	0.78
1000	0.79
2000	0.815
4000	0.82875
6000	0.8383333333333334
8000	0.848125
10000	0.8335



- Πίνακας Precision:

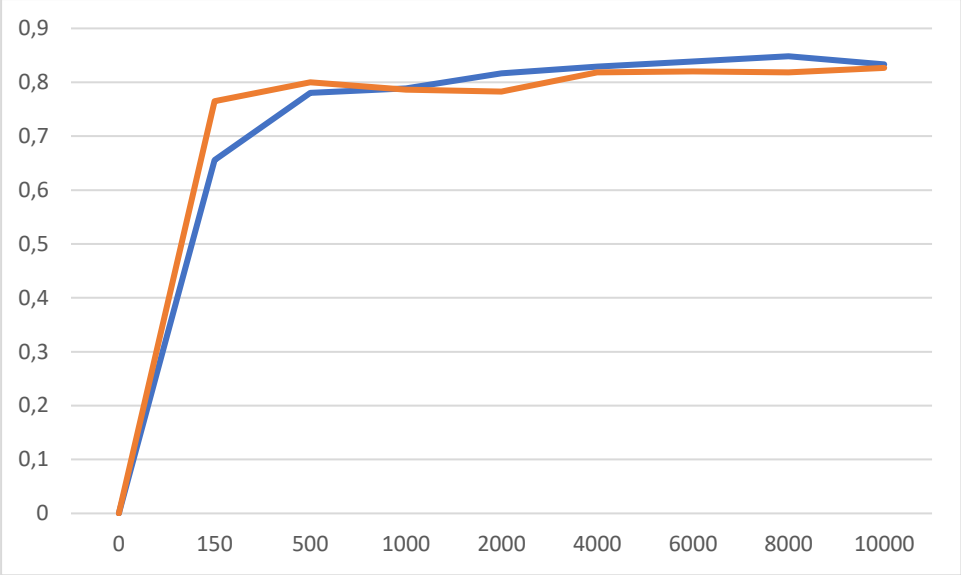
Data Size	Positive Precision	Negative Precision
150	0.6363636363636364	0.6842105263157895
500	0.7916666666666666	0.7692307692307693
1000	0.822429906542056	0.7526881720430108
2000	0.7766990291262136	0.8556701030927835
4000	0.8163771712158809	0.8413098236775819
6000	0.8288	0.8486956521739131
8000	0.8346915017462165	0.863697705802969
10000	0.8326810176125244	0.8343558282208589

Ο μπλε είναι ο Positive, ο πορτοκαλί είναι ο Negative:



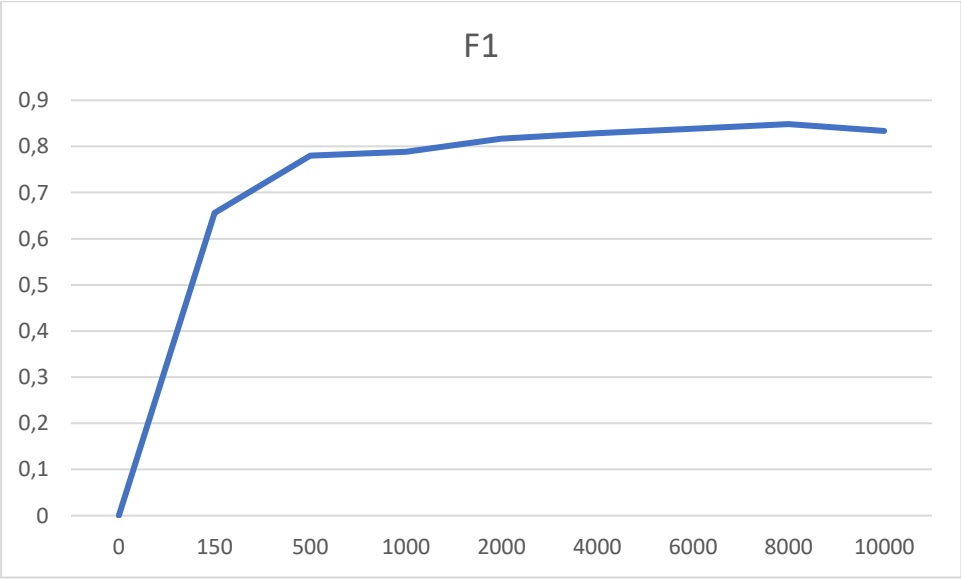
- Πίνακας Recall

Data Size	Positive Recall	Negative Recall
150	0.5384615384615384	0.7647058823529411
500	0.76	0.8
1000	0.7927927927927928	0.7865168539325843
2000	0.851063829787234	0.7830188679245284
4000	0.8392857142857143	0.8186274509803921
6000	0.856198347107438	0.8201680672268907
8000	0.8765281173594132	0.8184143222506394
10000	0.8400789733464955	0.8267477203647416



- Πίνακας F1

Data Size	F1
150	0.6559065253985542
500	0.7802242944583659
1000	0.7886055389004415
2000	0.8166127327505607
4000	0.8289000361828954
6000	0.8384654215740981
8000	0.8483320365283804
10000	0.833465881574388



- Train και Test curves:

Data Size	Train Accuracy
150	1
500	0.995
1000	0.9675
2000	0.914375
4000	0.914375
6000	0.8702083333333334
8000	0.86375
10000-	0.86025

Ο μπλε είναι ο Test, ο πορτοκαλί είναι ο Train: