ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Χειμερινό εξάμηνο 2021 -2022

2η Εργασία

|  |  |
| --- | --- |
| Καθηγητής | Ίων Ανδρουτσόπουλος |
| Συμμετέχον 1ος | Δημήτρης Καββαδάς |
| Αριθμός Μητρώου | 3190064 |
| Συμμετέχον 2ος | Αλέξης Φιλιππακόπουλος |
| Αριθμός Μητρώου | 3190212 |

# Contents

[Naive Bayes 2](#_bookmark0)

[Logistic Regression 5](#_bookmark2)

Naive Bayes

**Ζητούμενο**

Για την κατάταξη κειμένων στις 2 κατηγορίες(θετικές, αρνητικές) υλοποιήσαμε τον Naive Bayes αλγόριθμο σε πολυωνυμική μορφή με εκτιμήτρια Laplace κατά τον υπολογισμό τον prior πιθανοτήτων. Τις class conditional πιθανότητες τις υπολογίσαμε με την Gauss formula.

**Επιλογή δεδομένων**

Καθώς τα δεδομένα έχουν γίνει random shuffle, η επιλογή των train δεδομένων είναι τυχαία, καθώς από τα 25000 για την αναπαράσταση των καμπύλων έχω πάρει τα 12500 train και όλα τα test δεδομένα.

**Υπερπαράμετροι**

Οι υπερπαράμετροι χρησιμοποιούνται καθώς αναπαριστώ όλα τα δεδομένα σε 0,1 (encrypt\_reviews(file\_path,dict,m,n)). Μετά από πολλές δοκιμές του αλγορίθμου.

Για το m που είναι η επιλογή των συχνότερων λέξεων δοκιμάσαμε τις τιμές: 500,1000,1500,3000

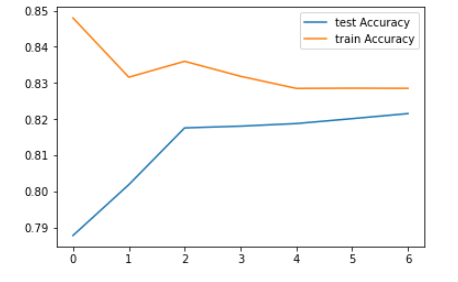
Και για το n που είναι η παράλειψη των n συχνότερων λέξεων δοκιμάσαμε τις τιμές: 1,10,15,20,30

Μεγιστοποίηση του accuracy των test δεδομένων γίνεται με:

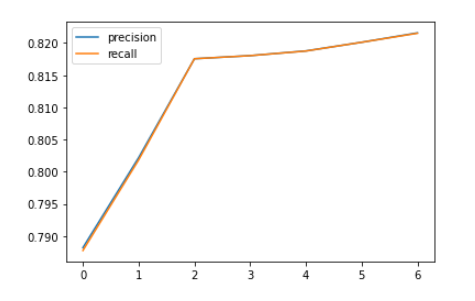
m=1000 και n=15

**Καμπύλες Μάθησης**

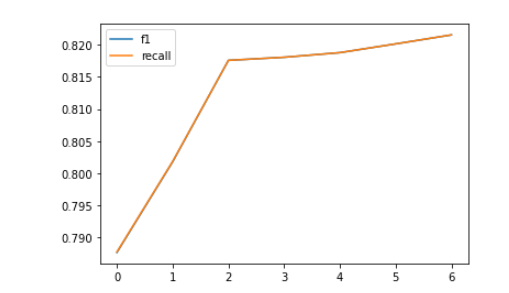
Παρατηρούμε ότι αυξάνοντας το πλήθος των train δεδομένων το accuracy για αυτά που χρησιμοποιήθηκαν μειώνεται , καθώς το αντίθετο γίνεται στα test δεδομένα



**Καμπύλη Precision-Recall**

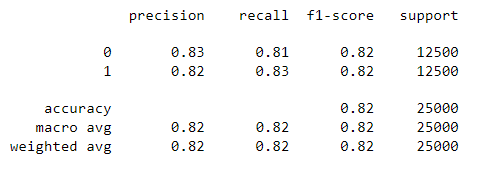
****

**Καμπύλη F1-Recall**

****

**Classification Report**

Εφαρμόζουμε το καλύτερο μοντέλο στο οποίο το accuracy μεγιστοποιείται με τις κατάλληλες υπερπαραμέτρους.



Logistic Regression

**Ζητούμενο**

Για την κατάταξη κειμένων στις 2 κατηγορίες(θετικές, αρνητικές) υλοποιήσαμε τον Logistic Regression αλγόριθμο

προσθέτοντας όρο κανονικοποίησης (λ) στην αντικειμενική συνάρτηση.

**Επιλογή δεδομένων**

Καθώς τα δεδομένα έχουν γίνει random shuffle, η επιλογή των train δεδομένων είναι τυχαία, καθώς από τα 25000 για την αναπαράσταση των καμπύλων έχω πάρει τα 12500 train και όλα τα test δεδομένα.

**Υπερπαράμετροι**

Οι υπερπαράμετροι χρησιμοποιούνται καθώς αναπαριστώ όλα τα δεδομένα σε 0,1 (encrypt\_reviews(file\_path,dict,m,n)). Μετά από πολλές δοκιμές του αλγορίθμου.

Για το m που είναι η επιλογή των συχνότερων λέξεων δοκιμάσαμε τις τιμές: 500,1000,1500,3000

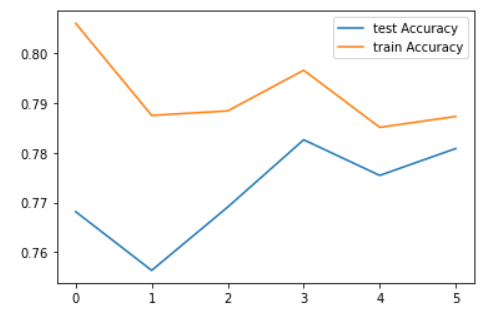
Και για το n που είναι η παράλειψη των n συχνότερων λέξεων δοκιμάσαμε τις τιμές: 1,10,15,20,30

Μεγιστοποίηση του accuracy των test δεδομένων γίνεται με:

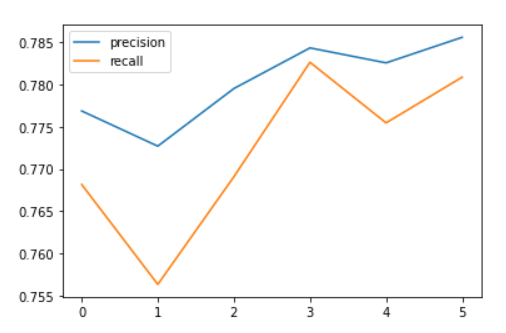
m=1000 και n=15

Tέλος ο όρος κανονικοποιήσης (λ) παίρνει την τιμή 0,001.

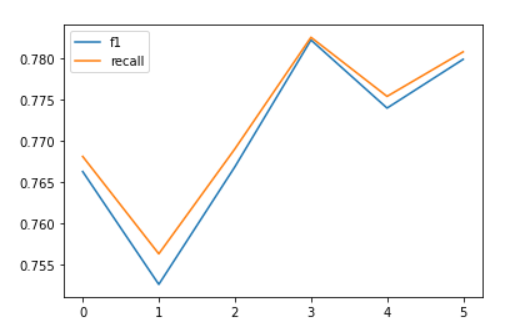
**Καμπύλες Μάθησης**



**Καμπύλη Precision-Recall**



**Καμπύλη F1-Recall**



**Classification Report**

Εφαρμόζουμε το μοντέλο με τις ίδιες παραμέτρους που βάλαμε και στον πρώτο αλγόριθμο

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Όλες οι δοκιμές των αλγορίθμων γίνανε με τις ίδιες τιμές στις υπερπαραμέτρους καθώς κάθε φορά για τις αναπαραστάσεις των καμπύλων χρησιμοποιήσαμε και πάλι των ίδιο αριθμό train δεδομένων και όλα τα test δεδομένα.