2η Εργασία Τεχνητής Νοημοσύνης

Φοιτητές:

Βιτάλης Ιωάννης (ΑΜ: 3150011)

Το πρόγραμμα δοκιμάστηκε σε υπολογιστή με τα εξής τεχνικά χαρακτηριστικά:

OS: Windows 10 Pro(x64)

CPU: Ryzen 5 3600

RAM: 2X8GB @3200MHz

Storage Type: SATAIII SSD

IDE: IntelliJ IDEA

Εισαγωγή:

Για την κατάταξη κειμένων σε δύο κατηγορίες(θετικές/αρνητικής) υλοποίησα σε Java τον αλγόριθμο Αφελή ταξινομητή Bayes(πολυωνυμική μορφή) με εκτιμήτρια Laplace κατά τον υπολογισμό των πιθανοτήτων.

Στην πολυωνυμική μορφή μας ενδιαφέρει το πόσες φορές εμφανίστηκε η λέξη στο κείμενο άρα δεν μπορούσα να αναπαραστήσω το κείμενο σε ένα διάστημα ιδιοτήτων με τιμές 0 ή 1, αντιθέτως πρόσθετα κάθε φορά το πόσες φορές εμφανίζεται η κάθε λέξη.

Διαχωρισμός δεδομένων:

Το σύνολο δεδομένων «IMDB dataset», περιέχει συνολικά 25.000 train δεδομένα χωρισμένα σε 12.500 positive reviews και 12.500 negative reviews, απο τα οποία επέλεξα το 10% (1.250 positive και 1.250 negative) ως Dev δεδομένα, τα οποία θα χρησιμοποιηθούν για την απεικόνηση των καμπύλων μάθησης.

Τα 25.000 test δεδομένα χρησιμοποιήθηκαν εξ'ολοκλήρου για την εξαγωγή συμπερασμάτων.

Υπερπαράμετροι:

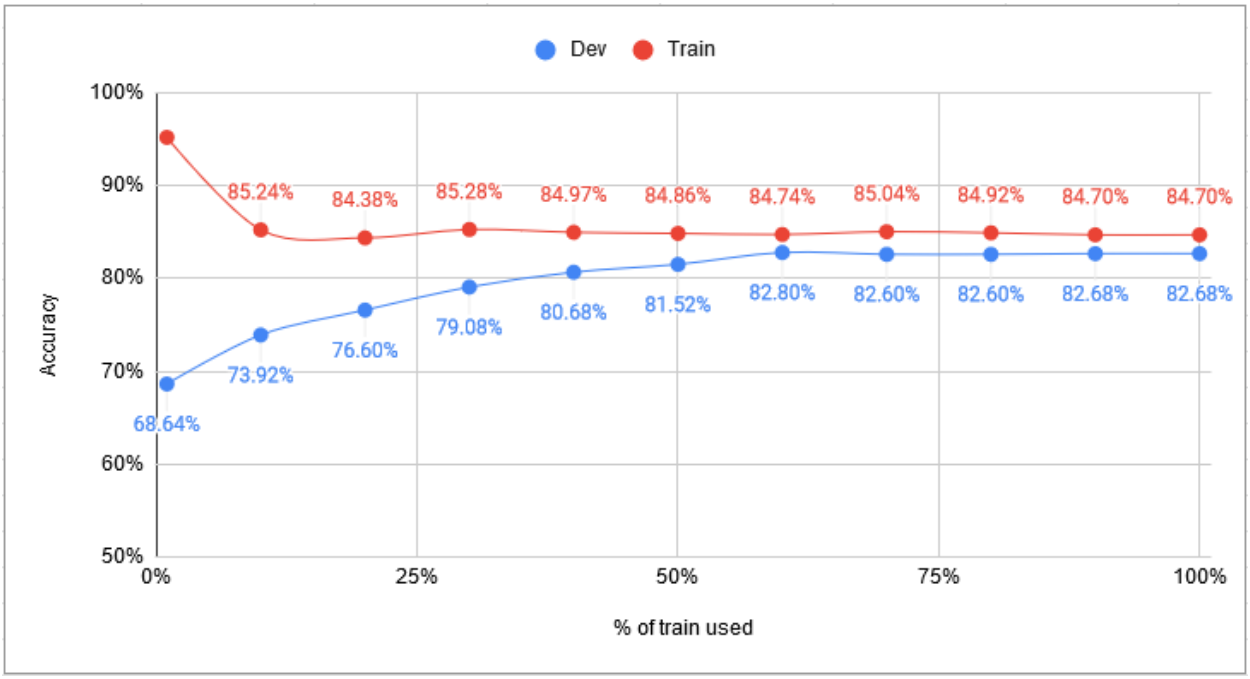
Η επιλογή υπερπαραμέτρων έγινε με brute-forcing ελέγχοντας τον αλγοριθμό στα dev δεδομένα για διαφορετικα n,m, καλώντας την μέθοδο removeUninformative(n,m) της κλάσσης TrainD.

Δοκίμασα την παράλειψη των 50,60,70,80,90 και 100 πιο σύχνων λέξεων (n) και παρατήρησα οτι για n=75 το Accuracy λάμβανε την μέγιστη τιμή του.

Με όμοιο τρόπο έγινε και η επιλογή των m συχνότερων λέξεων που θα περιλαμβάνονταν στο λεξιλόγιο, δοκιμάστηκαν οι 500,1000,1500,2000,2500,3000 συχνότερες λέξεις με το Accuracy στα dev δεδομένα να μεγιστοποιείται για m=4000.

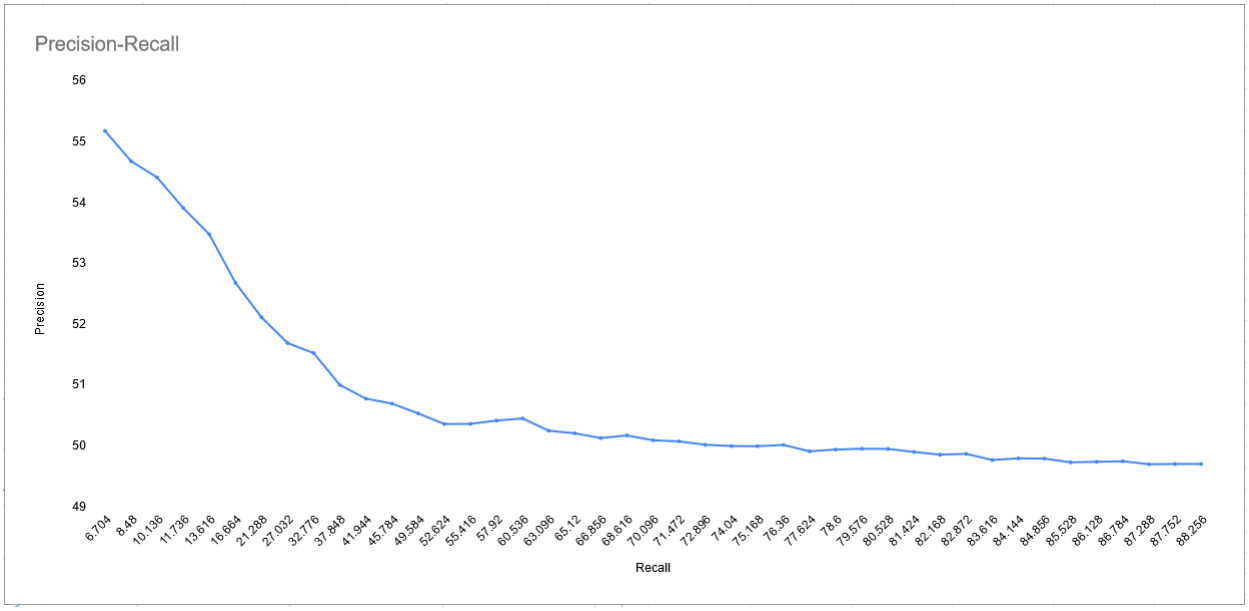
Καμπύλες Μάθησης:

Αυξάνοντας το πλήθος των παραδειγμάτων εκπαίδευσης το Accuracy στα Train data που χρησιμοποιήθηκαν αυξάνεται, ενώ το Accuracy στα Dev data αυξάνεται και προκύπτει η παρακάτω καμπύλη μάθησης.

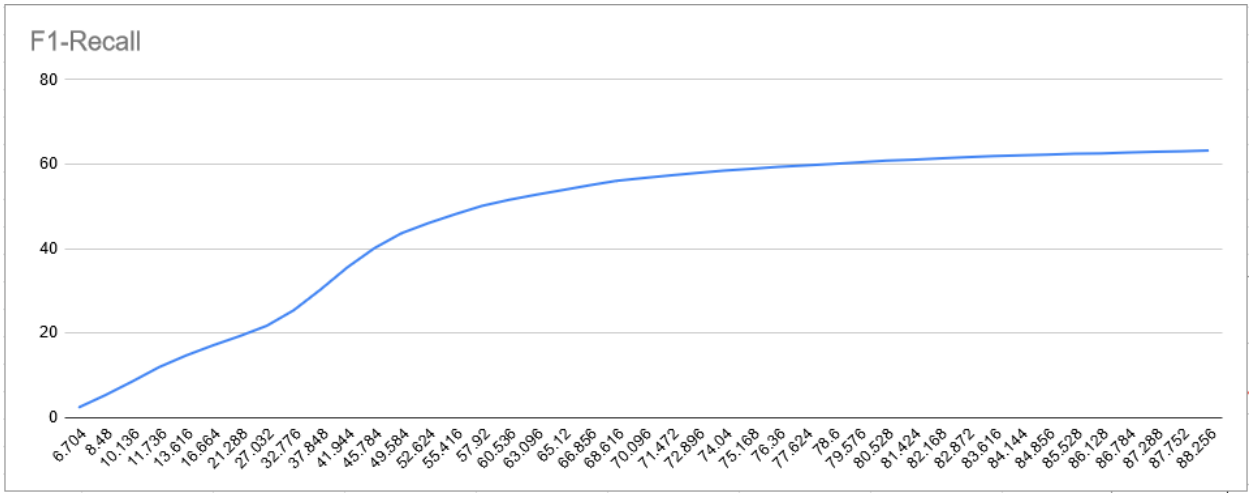


Παρατηρούμε οτι το Accuracy των Dev data μεγιστοποίειτε όταν χρησιμοποιούμε το 60% των συνολικών training data.

Καμπύλη Precision-Recall:



Καμπύλη F1-Recall:



Test Data:

Εφαρμόζουμε το καλύτερο μοντέλο που προέκυψε απο την καμπύλη μάθησης στο σύνολο των test data και προκύπτουν οι παρακάτω τιμές.

```
Acc: 83.284%
Precision: 82.23444%
Recall: 84.912%
F1 Score: 83.55177%
```