Bayesian Learning

Project 5

Καλτσίδης Μιχάλης

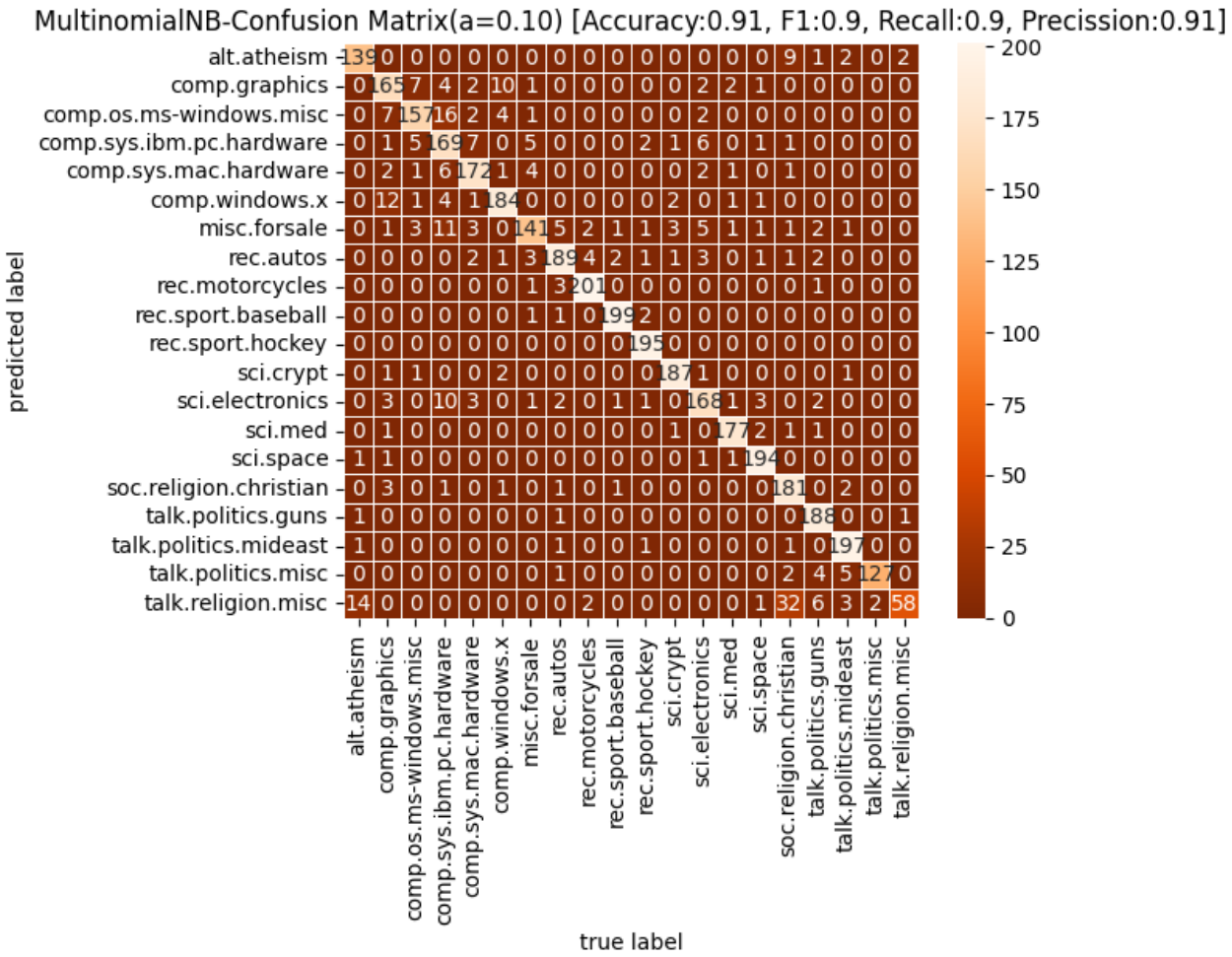
AEM:108

Στην παρούσα εργασία υλοποιήθηκε ο αλγόριθμος Naive Bayes με τον πολυωνυμικό ταξινομητή Multinomial Naive Bayes, ο οποίος είναι κατάλληλος για ταξινόμηση με διακριτά χαρακτηριστικά (π.χ. πλήθος λέξεων για ταξινόμηση κειμένου). Η πολυωνυμική κατανομή απαιτεί κανονικά πλήθος χαρακτηριστικών ακεραίων. Ωστόσο, στην πράξη, οι κλασματικές μετρήσεις όπως το tf-idf μπορεί επίσης να λειτουργήσουν.

Επιλέχθηκε το σύνολο δεδομένων fetch_20newsgroups το οποίο περιέχει 11314 διαφορετικά κείμενα διαφορετικού περιεχομένου και 20 διαφορετικές θεματικές ενότητες ι οποίες αποτελούν τις κατηγορίες του συνόλου των δεδομένων, το target ουσιαστικά.

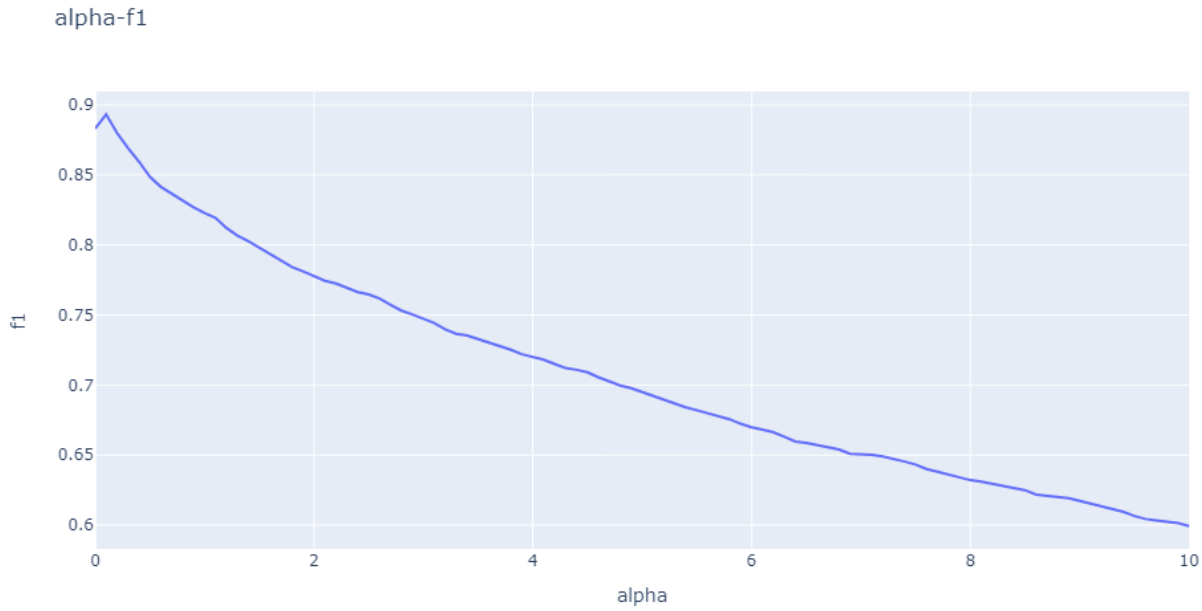
Φυσικά επειδή μιλάμε για ένα απλό κείμενο το οποίο δεν έχει προεπεξεργαστεί, θα πρέπει να γίνουν κάποιες τροποποιήσεις. Όπως για παράδειγμα η αφαίρεση των αποσιοποιητικών ή των αριθμών στο κείμενο διότι δεν αποτελούν στοιχεία τα οποία θα βοηθήσουν στην υλοποίηση κάποιου αλγορίθμου.

Μετά από όλη την επεξεργασία δημιουργήθηκε και ο πίνακας συσχέτισης μεταξή των 20 αυτών κατηγοριών και μετά δημιουργήθηκε ένας heatmap στον οποίο διαφένεται αν η προβλεπόμενη τιμή που πρόβλεψε ο αλγόριθμος ταυτίζεται με την πραγματική τιμή.



Παρατηρείται ότι όταν η προβλεπόμενη τιμή του αλγορίθμου δείνει talk.religion.misc, μόνο οι 58 από τις 108 είναι όντως η talk.religion.misc. 32 από τις 108 δίνουν προβλεπόμενη τιμή talk.religion.misc ενώ η πραγματική τιμή είναι soc.religion.christian. Ο παραπάνω πίνακας δημιουργήθηκε για alpha=0.10. Είναι ενδιαφέρον να υλοποιηθεί ο αλγόριθμος για διαφορετικά a. Για τον λόγο αυτόν δημιουργήθηκε πίνακας ο οποίο περιέχει τις τιμές του a και κάποια μετρική (accuracy, f1, recall, precission).

Στην παρούσα εργασία ζητείται να επιτευχθεί ο στόχος του f1 να είναι γύρω στο 70%. Το γράφημα που με αρκετή ευκολία θα δώσει την απάντηση είναι το παρακάτω:



Εδώ παρατηρείται η εκθετική μείωση μείωση της f1 μετρικής συναρτήση του a και ο στόχος του 0.70 για την μετιρκή f1 επιτυγχάνεται περίπου για alpha=4.4.