Lecture 4: Bayesian Learning

Project 5

Να φορτώσετε το dataset **fetch_20newsgroups** και να εφαρμόσετε τη μέθοδο **Naive Bayes** (συνιστάται η χρήση της βιβλιοθήκης **scikit** για Python) για την ταξινόμηση κειμένου σε κατηγορίες.

Να χρησιμοποιηθεί ο αλγόριθμος Naive Bayes με την υπόθεση ότι τα δεδομένα ακολουθούν πολυωνυμική κατανομή (και συνεπώς να δοκιμαστούν οι διάφορες τιμές της παραμέτρου *a*). Να υπολογιστούν οι τιμές των μετρικών **Recall**, **Precision** και **F1**, και να δημιουργηθεί το γράφημα θερμότητας (heatmap) του πίνακα σύγχυσης, όπως φαίνεται στο συνημμένο παράδειγμα. Τα αποτελέσματα των μετρικών να φαίνονται στον τίτλο του γραφήματος.

Σημείωση 1: Ένας σημαντικός (αλλά προαιρετικός) σκοπός είναι να επιτευχθεί τιμή για την **F1** κάτω από **70%** (ή περίπου εκεί)!

Σημείωση 2: Κατά την υποβολή της εργασίας, είναι απαραίτητο να συμπεριληφθεί και ο κώδικας που χρησιμοποιήθηκε, πέραν του γραφήματος. Για διευκόλυνση, επισυνάπτεται το αρχείο **NB_Template.py** μέσα στο οποίο μπορεί να συμπληρωθεί ο απαραίτητος κώδικας σε Python.