In the following report, i am going to describe the results of the experiments that take place as an assingment of the 'Programming

for language technology II'. First of all, the feature which were used are the N-grams (from 2 to 6) with the support of a Logistic

Regression model. Moreover, the feature selection method which was applied is the TfidfVectorizer with max\_features as a parameter

that build a vocabulary that only consider the top max\_features ordered by term frequency across the corpus. As a result, the features

which are more important are those which in experiment of 10000 made Average accuracy score 0.686329738863644 (['100 different',

'13 years','13 years emilio', '13 years emilio the', '13 years emilio the godfather', '13 years emilio the godfather has', '25 which',

'25 which would', '25 which would expect', '25 which would expect at']). The next are those which in experiment of 100 made Average

accuracy score 0.682529996336231 (['all the', 'and it', 'and service', 'and the', 'and the food', 'and the service', 'and they',

'and you', 'at the', 'but not']). Last but not least are those which in experiment of 1000 made Average accuracy score 0.6793527715150409

(['about sushi', 'about sushi and', 'about sushi and were', 'about sushi and were both', 'about sushi and were both please', 'about the',

'about this', 'above average', 'all in', 'all in all']).