CMPE 255: Program 1 Report

Rank: 22

F1- Score: 0.7730

Approach:

For this activity I have used the following approach.

Data Extraction (from given files) => Data Preprocessing => Train with the training data => Predict test data.

Methodology:

For **data extraction** step I converted the .dat files to .txt and read the files using file.readlines. After reading the file I have taken 2 different lists for the training data. In this first list are the numbers that are given in the train data. And in this second list at the corresponding index is the string of documents of the abstract.

And did the same thing for the test data. But in this case just one list which just test data abstracts.

In the next step, **Data Preprocessing,** I have performed various operations on the lists we received in the first step using CountVectorizer, TfidfTransformer, SGDClassifier. Here as it can be seen I have used Support Vector Machines **(SVM)** algorithm for classification.

But before this I tried the **Naïve Bayes** algorithm which code is in the **commented** part of my code submission. But with that I was getting only **0.43** of F1-Score. Thus, I took the decision to change the algorithm and after making a few changes I got the current F1-Score.

For the third step of **Train with the training data** is straightforward in python. Using the two lists found in the 2nd step, the list of all the abstracts from the train data and the numbers that indicate the one of the 5 classes, preprocessed. Also, in the second function I have used the pipeline function, which is used to combine all the preprocessing methods into one and the apply all the functions on the training data at once, thus there is no confusion about different variable names in the code.

Finally, for the **prediction** we use the trained data we found in step 3 and input the that predicator with the list we got in step one for the test data. Once the process is done we get a array of the predicated values for the test data and then using the *file.write* I have copied the data into a text file and then uploaded the file to obtain the above results.