EE 219 PROJECT 4

TEAM MEMBERS

**PART 1**

Tf - idf (*term frequency-inverse document frequency*) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus

For this problem,

* The documents in the data set are turned into numerical feature vectors.
* The words here are tokenized and unwanted tokens like punctuation marks, stop words,

stemming words etc are removed for preprocessing the data.

* After this a TFxIDF vector representation is created and its terms are reported as

**Total number of terms are 54433**

**PART 2**

In this problem, k-means clustering was applied to the data points with k = 2.

To find the correct ground truth cluster labels, we assigned the cluster value 0 to labels starting with ‘com.’ if most data points with belonging to these labels were clustered in cluster 0. And similarly for cluster 1. The algorithm was run 10 times and error percentage, homogeneity score, completeness score, adjusted rand score. The adjusted mutual info score were found and the following results were obtained-