Yichen Wu 504294181
Siyuan Chen 405024391

Read README.md
Or go to https://github.com/popo0293/EE219Project2

1. Stop words and punctuations were removed from the documents in order to get meaningful data. Words were not stemmed. With CountVectorizer, we tokenized the documents into words. By setting the parameter min_df as 3, we get our result as:
   With min_df = 3, (training documents, terms extracted): (4732, 20297)
2. K-means was utilized for clustering in this task. We generated 2 centroids in order to get 2 clusters. Clustering result is shown in the contingency table: 3383 points were in correct clusters, and 1349 points were in wrong clusters. By examination of the homogeneity score, the completeness score, the V-measure, the adjusted Rand score and the adjusted mutual info score, we could also find that the clustering result was not ideal: the homogeneity score showed that only 24.5% of the points in the clusters were from a single class, the completeness score showed that 31.9% of the points from the same class were put into the same clusters, Rand Index was 0.185 which showed a low accuracy, and the adjusted mutual information score showed that 24.5% of points shared mutual information, which was consistent with homogeneity score since we have only two clusters.
3. The graph shows that as the number of components increased, the percent of the variance increased. In addition, we can see that the rate of growth of the percent of variance was getting smaller as the number of components increased.
   From contingency table and the graph of scores vs r value, we can see that the performance of clustering algorithm was not monotonically changing as the r value of LSI/NMF increased, and the scores showed that when r=2 we can get the best clustering result. This can be explained by the fact that the information a low dimension (say 1) matrix carried was not enough, while if there were more dimensions than needed, distances between points would be close and dimension curse would hinder the clustering performance.
4. By visualization and the measure of five scores that we used before, we can see that normalization would hinder the clustering performance significantly when LSI was used for dimension reduction, while normalization does not have a significant effect on the performance of clustering when NMF was used.
   Logarithm transformation in this case gave us better results for NMF on the performance of clustering. This can be explained by the property of logarithm: it is a function with positive derivative and negative second derivative. In this case, logarithm transformation changes the distance between points and separated the points better.
5. When dataset was expanded into 20 categories, the best r value for truncated SVD and NMF was 10. With this r value, we tried different methods of transformations. The results show that in this case, neither normalization nor logarithm would have significant effect on the performance of clustering.