**Project 2: Data Representations and Clustering**

**Part 1 - Clustering on Text Data**

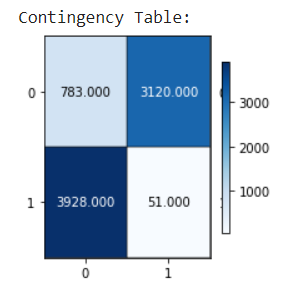
**Clustering with Sparse Text Representations**

**QUESTION 1:** Report the dimensions of the TF-IDF matrix you obtain.

**ANS:** The TF-IDF matrix has 7882 rows (documents) and 23522 columns (features/words).

**QUESTION 2:** Report the contingency table of your clustering result. You may use the provided plotmat.py to visualize the matrix. Does the contingency matrix have to be square-shaped?

**ANS:** The contingency table is reported below.

****

The contingency table does not have to be square-shaped. The shape depends on the number of unique class labels and the number of unique predicted cluster labels. If there are n unique class labels and m unique predicted cluster labels, the contingency table will have n rows and m columns and if n ≠ m, the contingency table will be rectangular.

**QUESTION 3:** Report the 5 clustering measures explained in the introduction for K-Means clustering.

**ANS:**

Clustering Measures:

\* Homogeneity: 0.564

\* Completeness: 0.580

\* V-Measure: 0.572

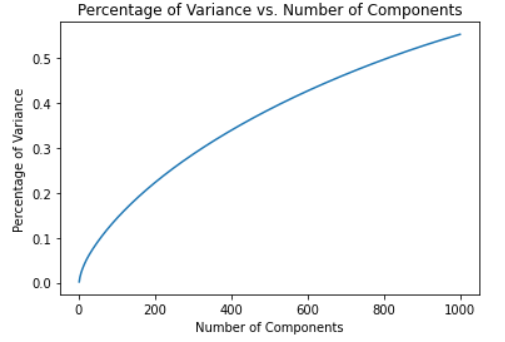
\* Adjusted Rand Index: 0.621

\* Adjusted Mutual Information: 0.572

**Clustering with Dense Text Representations**

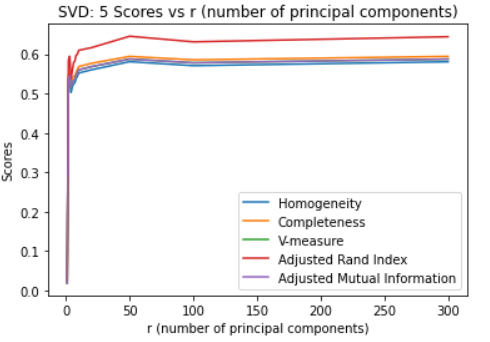
**QUESTION 4:** Report the plot of the percentage of variance that the top r principle components retain v.s. r, for r = 1 to 1000.

**ANS:**

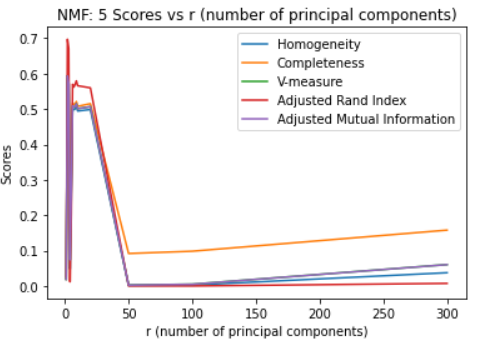
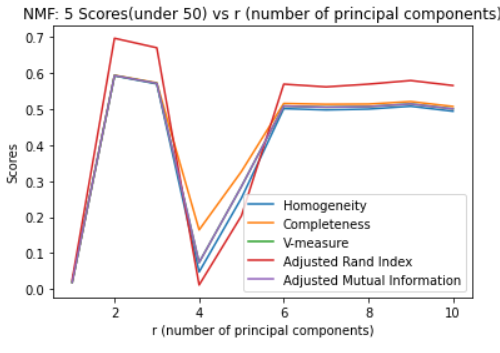
****

**QUESTION 5:** Let r be the dimension that we want to reduce the data to (i.e. n components). Try r = 1 − 10, 20, 50, 100, 300, and plot the 5 measure scores v.s. r for both SVD and NMF. Report a good choice of r for SVD and NMF respectively. Note: In the choice of r, there is a trade-off between the information preservation, and better performance of k-means in lower dimensions.

**ANS:** From the graph for SVD: 5 Scores vs r (number of principal components), the best choice for r is 300. Although, the graph seems to increase, if we increase r beyond 300, then it would increase till a particular point and then fall down.



From the graph for NMF: 5 Scores vs r (number of principal components), the best choice for r is 2 (magnified the graph on the right to identify r = 2)

**QUESTION 6:** How do you explain the non-monotonic behavior of the measures as r increases?

**ANS:** A non-monotonic behavior is observed in the graphs as r increases. The scores first increase, then fall and plateau off in the end as r increases. As the number of components increases, the dimensions from which k-means needs to perform clustering increases and k-means suffers due to the well known curse of dimensionality. It makes it difficult to cluster points as the euclidean distance is no longer a good metric because the ratio of euclidean distances between nearest and farthest point from the cluster center approaches 1. Thus, increasing the number of features beyond the elbow point does not add new information in the clustering task and the measures remain constant and eventually plateau.

**QUESTION 7:** Are these measures on average better than those computed in Question 3?

**ANS:** The results without data compression and after data compression using SVD and NMF with n\_components as 300 and 2 respectively are nearly the same. The observed trend has been summarized in the following points:

- The results of SVD are unique. The results of NMF are non-unique as it uses a stochastic algorithm for convergence.

- SVD has no restrictions on the data entries whereas in the case of NMF only positive entries are allowed in the reduced rank feature matrix. Thus, SVD is able to better represent the feature matrix as compared to NMF and the information loss is less in SVD. The results of SVD are more predictable when compared to NMF because the results are ordered by relevance on a geometric basis. Geometry is not considered by NMF in the feature space basis. Also, in SVD, increasing the value of r does not cause significant change in the scores because the most relevant features that are higher in the hierarchy are used in forming the feature matrix.

- In the graph of scores vs number of principle components for SVD, the scores saturate for r in the range 20 to 300. For NMF, the scores peak at r = 2.

Thus, the results of SVD are about the same and the results of NMF are worse.

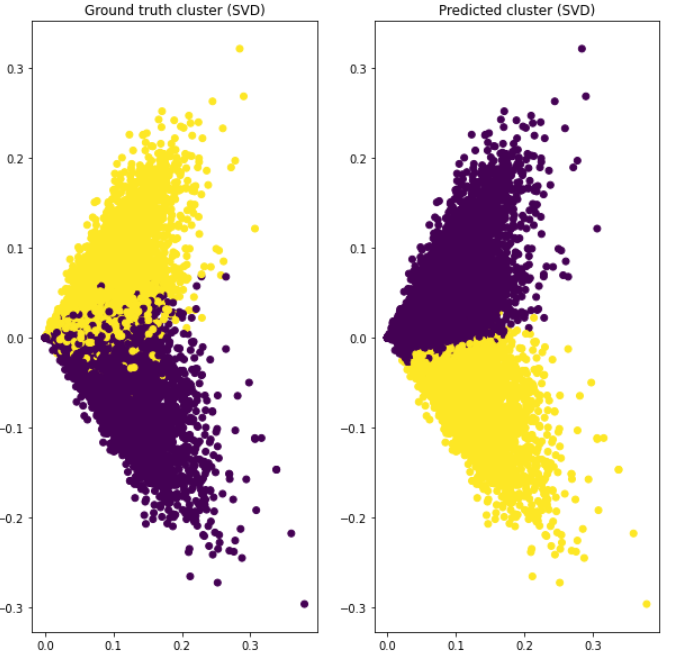
**QUESTION 8:** Visualize the clustering results for:

- SVD with your optimal choice of r for K-Means clustering;

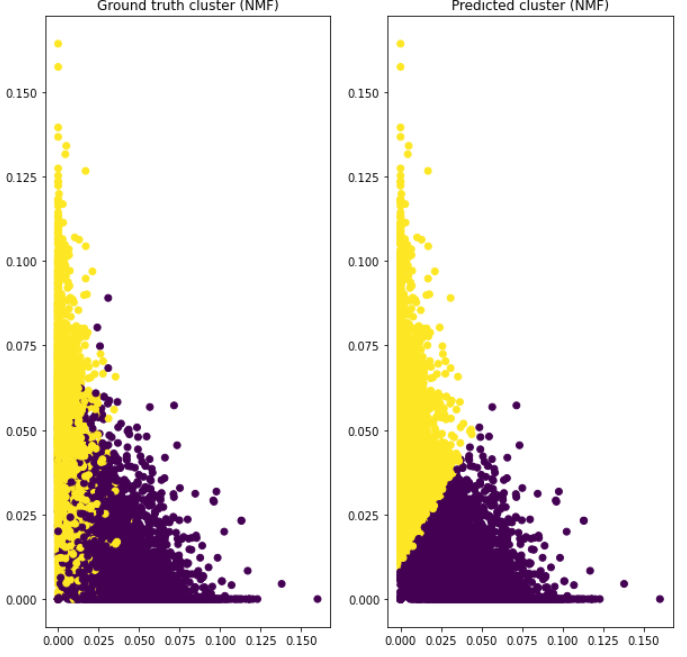
- NMF with your choice of r for K-Means clustering.

**ANS:** The 2 graphs are shown below.

- SVD with your optimal choice of r (300) for K-Means clustering



- NMF with your choice of r for K-Means clustering.



**QUESTION 9:** What do you observe in the visualization? How are the data points of the two classes distributed? Is distribution of the data ideal for K-Means clustering?

**ANS:** In both the visualizations, the data is linearly separable & the k-means model performs well on the given data. Yes the distribution is ideal for k-means clustering.