**Project 2: Data Representations and Clustering**

**Part 1 - Clustering on Text Data**

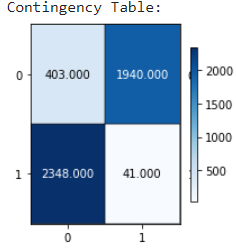
**Clustering with Sparse Text Representations**

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

**ANS:** The TF-IDF matrix has 4732 rows (documents) and 17131 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: 0.590

\* Completeness: 0.601

\* V-Measure: 0.595

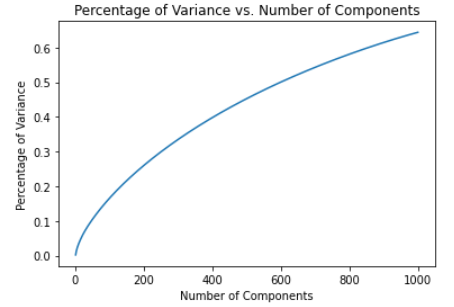
\* Adjusted Rand Index: 0.660

\* Adjusted Mutual Information: 0.595

**Clustering with Dense Text Representations**

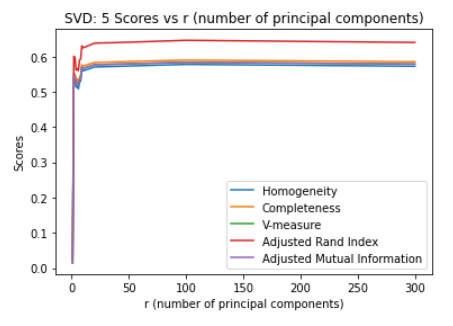
**QUESTION 4:** Report the plot of the percentage of variance that the top r principle components retain v.s. r, for r = 1 to 1000.

**ANS:**

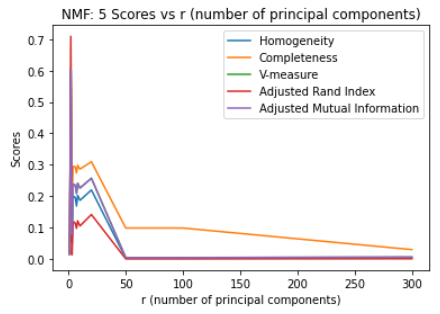
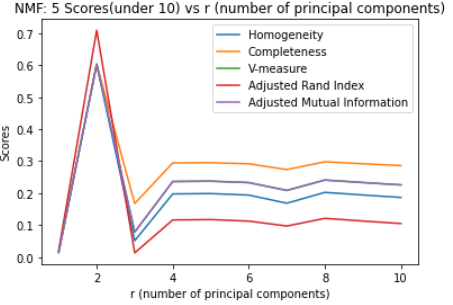
****

**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.

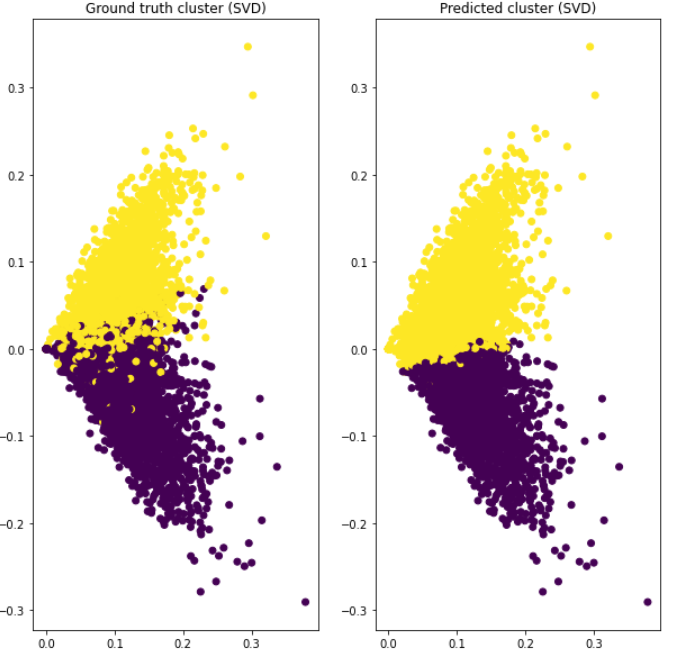
**QUESTION 8:** Visualize the clustering results for:

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

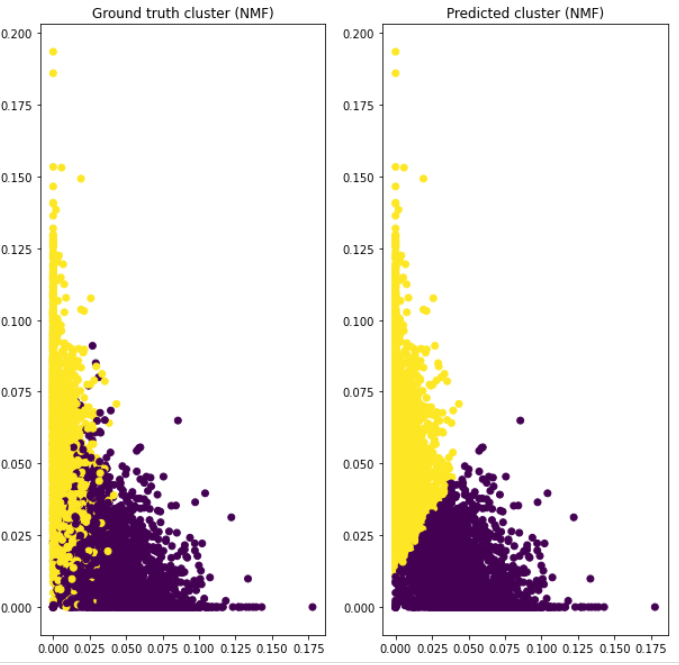
- NMF with your choice of r for K-Means clustering.

**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?

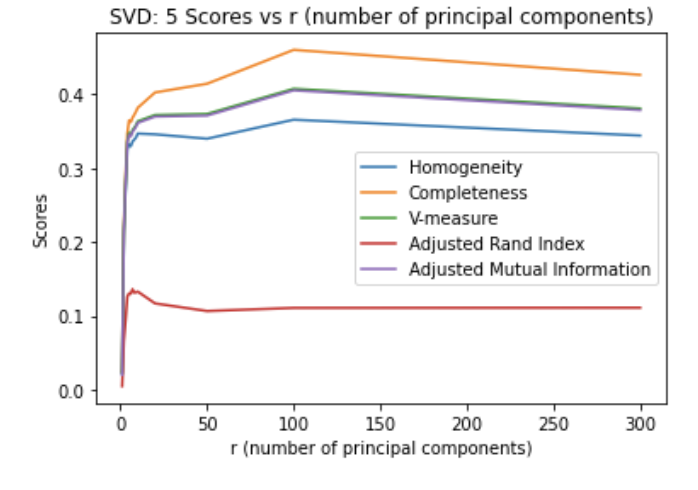
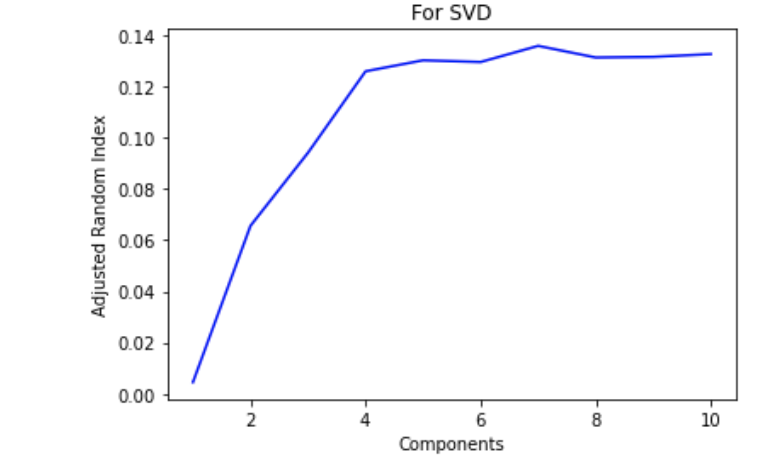
**ANS:** In both the visualizations, the data is linearly separable & the k-means model performs well on the given data. Yes the distribution is ideal for k-means clustering.

**3. Clustering of the Entire 20 Classes**

**QUESTION 10:** Load documents with the same configuration as in Question 1, but for ALL 20 categories. Construct the TF-IDF matrix, reduce its dimensionality using BOTH NMF and SVD (specify settings you choose and why), and perform K-Means clustering with k=20. Visualize the contingency matrix and report the five clustering metrics (DO BOTH NMF AND SVD).

**ANS:**

**For SVD:**From the graph for SVD: 5 Scores vs r (number of principal components), the best choice for r is 100. The cluster scores are first increasing and then decreasing after r = 100. The Adjusted Random index takes its best value at r = 7 but the value is low for other scores.

** **

5 clustering and contingency matrix for r = 100 is mentioned below

Clustering Measures:

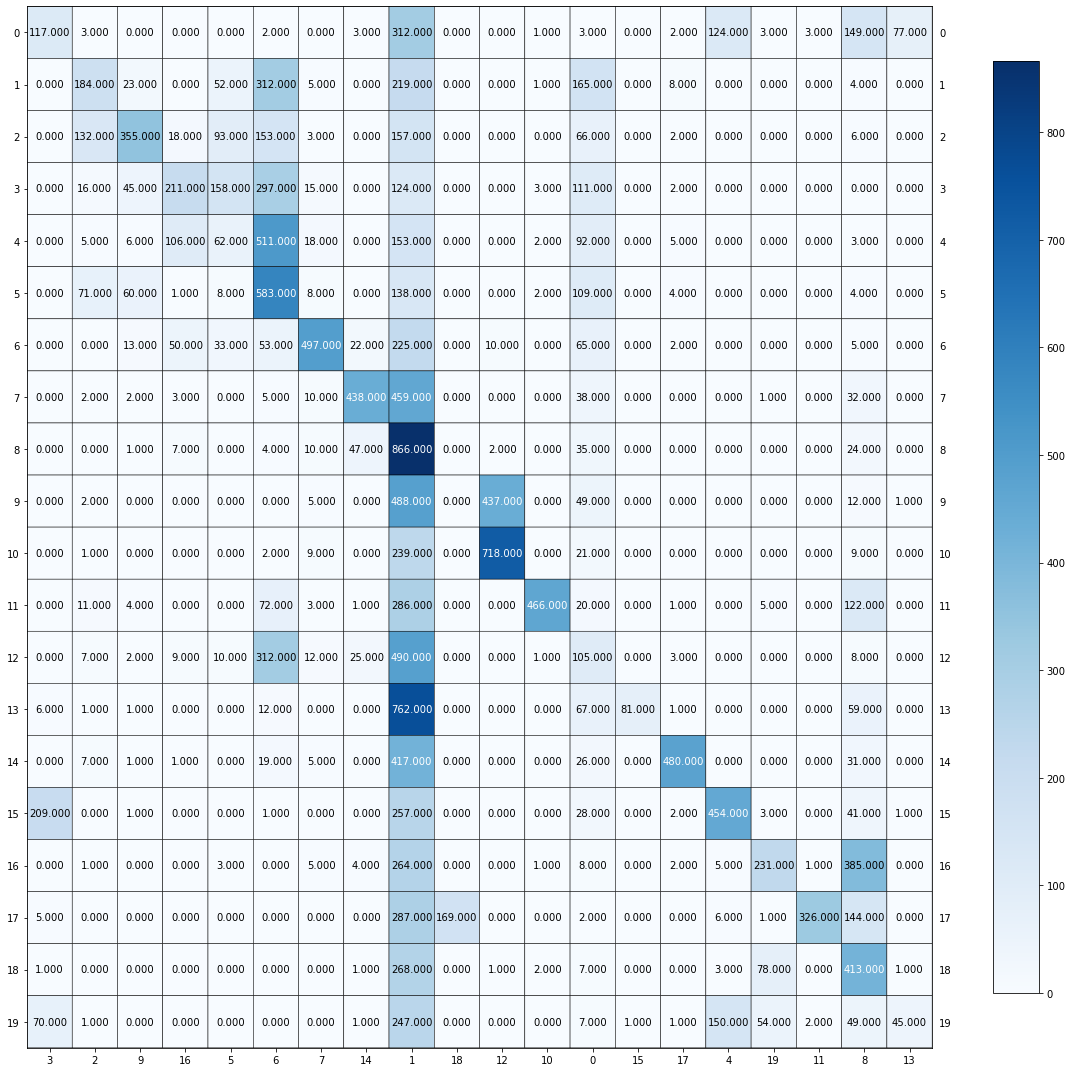
\* Homogeneity: 0.3657

\* Completeness: 0.4604

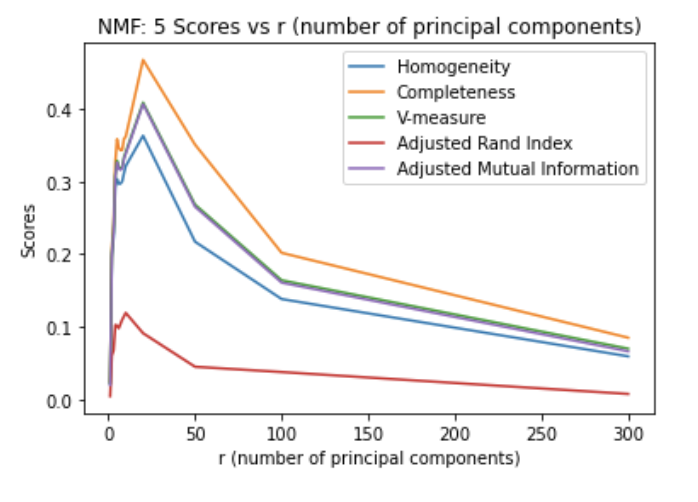
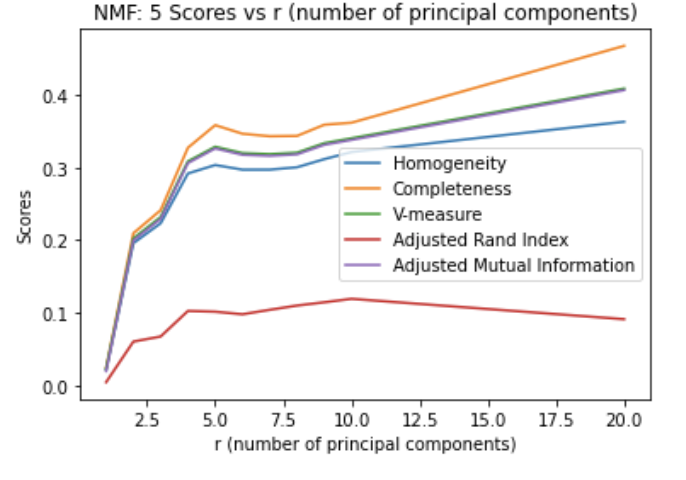
\* V-Measure: 0.4076

\* Adjusted Rand Index: 0.110

\* Adjusted Mutual Information: 0.4055

****

**For NMF:** 5 Scores vs r (number of principal components), the best choice for r is 20 (magnified the graph on the right to identify r = 20)

5 clustering and contingency matrix for r = 20 is mentioned below

Clustering Measures:

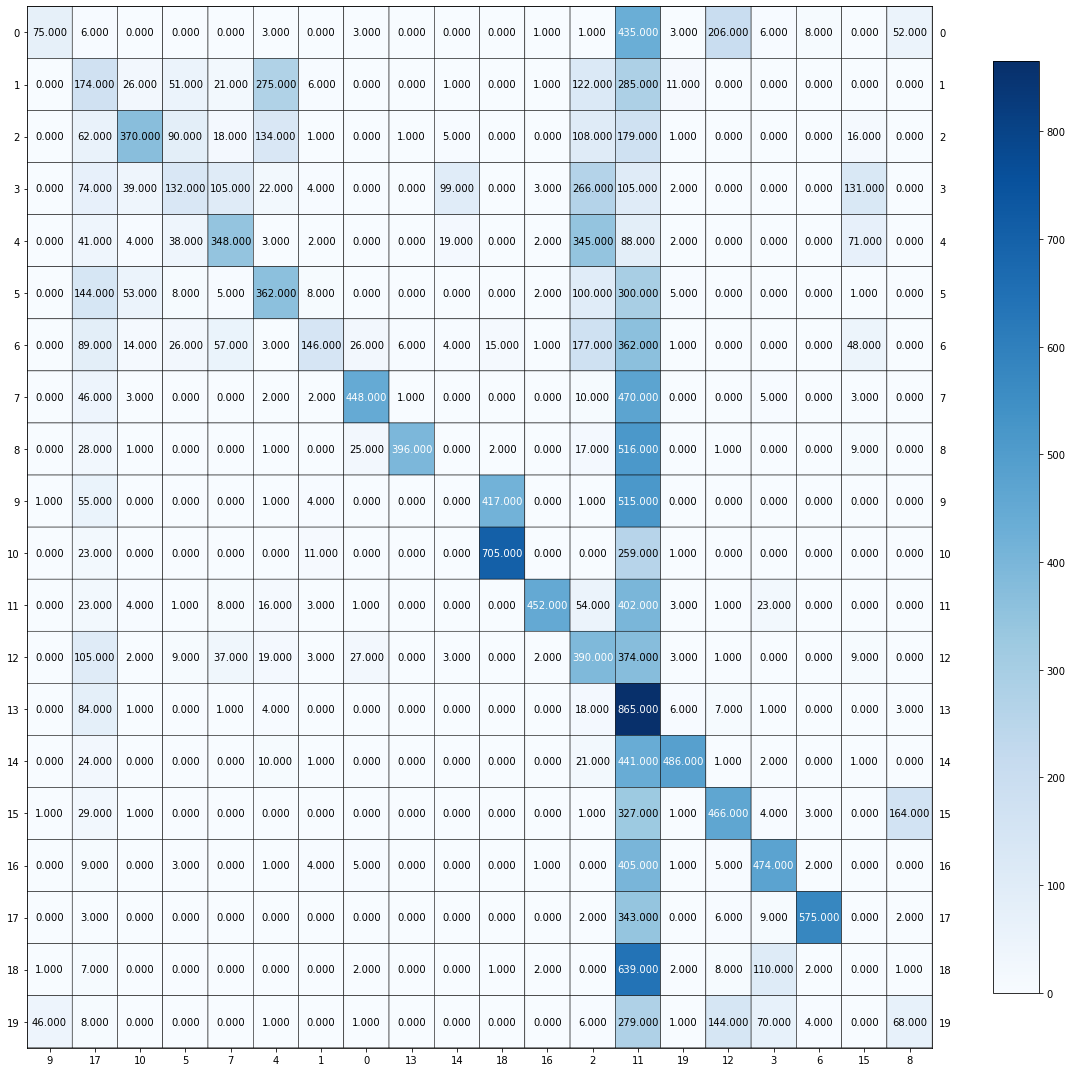
\* Homogeneity: 0.362

\* Completeness: 0.4673

\* V-Measure: 0.408

\* Adjusted Rand Index: 0.0911

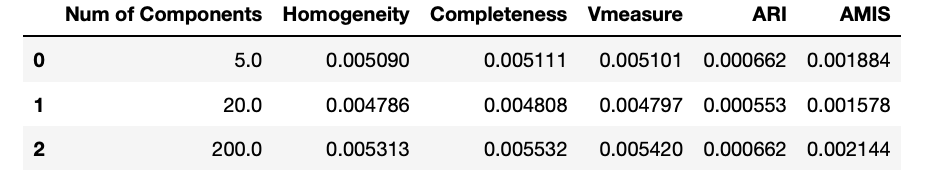
\* Adjusted Mutual Information: 0.4063



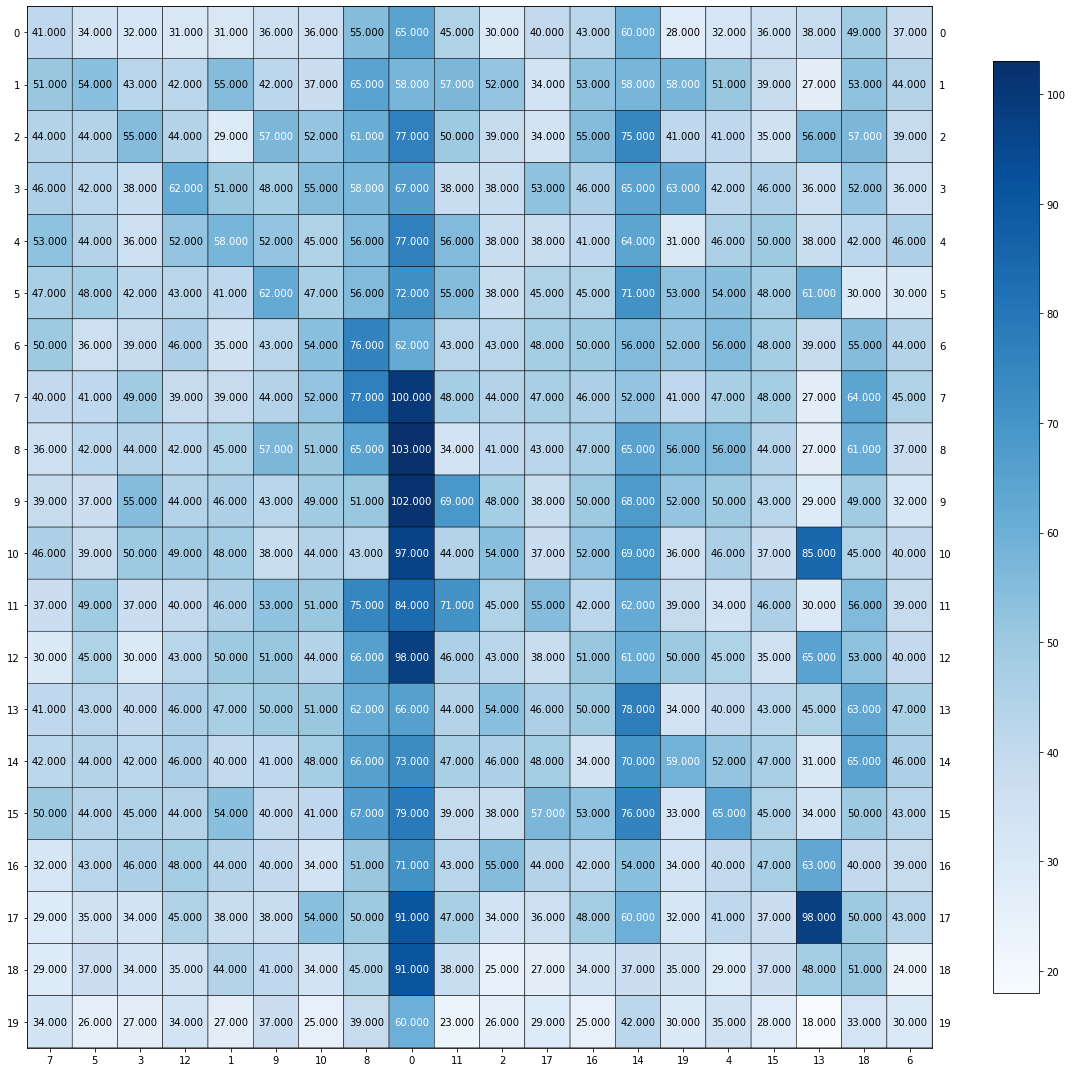
**4. UMAP**

**QUESTION 11:** Reduce the dimension of your dataset with UMAP. Consider the following settings: n components = [5, 20, 200], metric = ”cosine” vs. ”euclidean”. If ”cosine” metric fails, please look at the FAQ at the end of this spec.Report the permuted contingency matrix and the five clustering evaluation metrics for the different combinations (6 combinations).

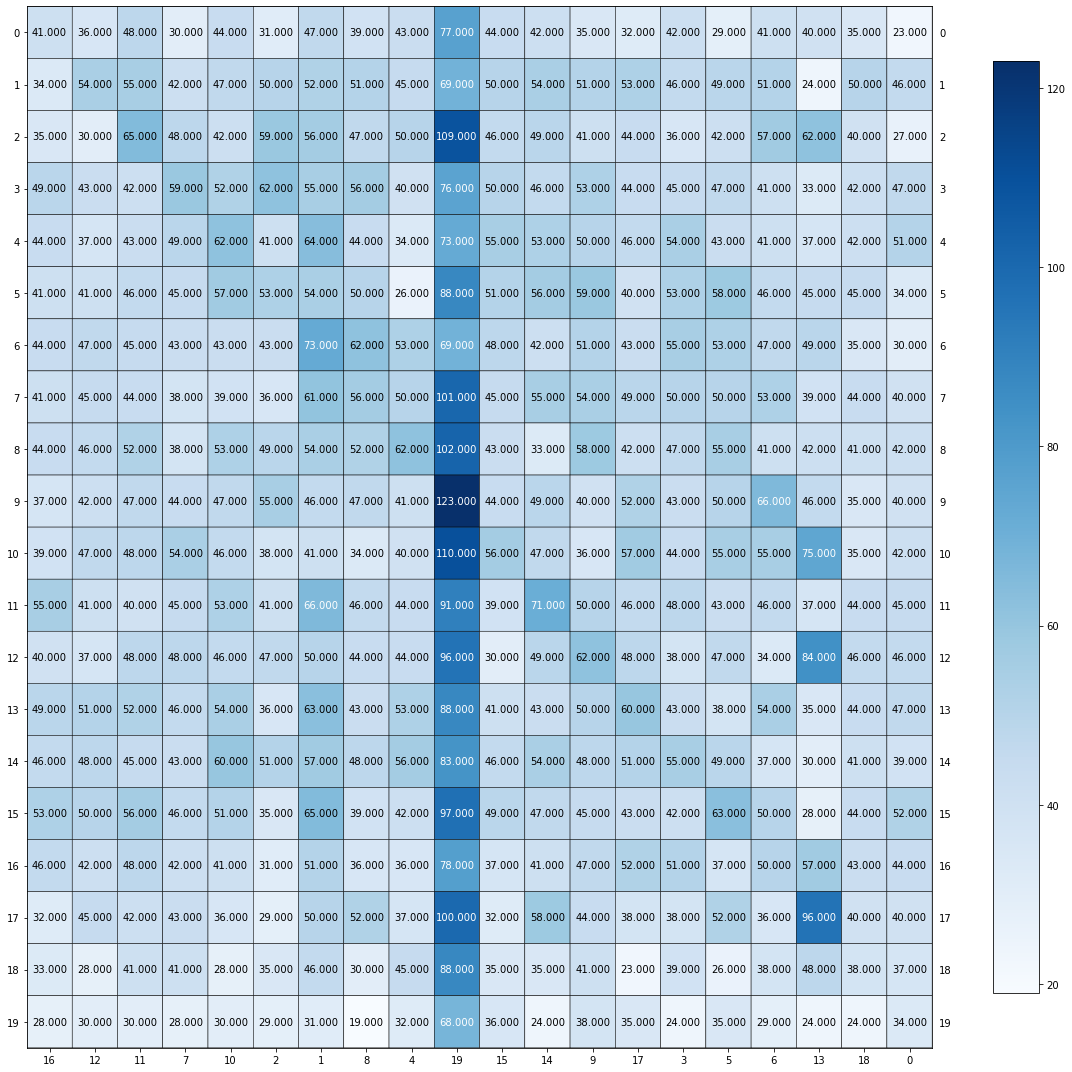
**ANS: For Euclidean:** 5 Scores for Umap with Euclidean metric for different number of components



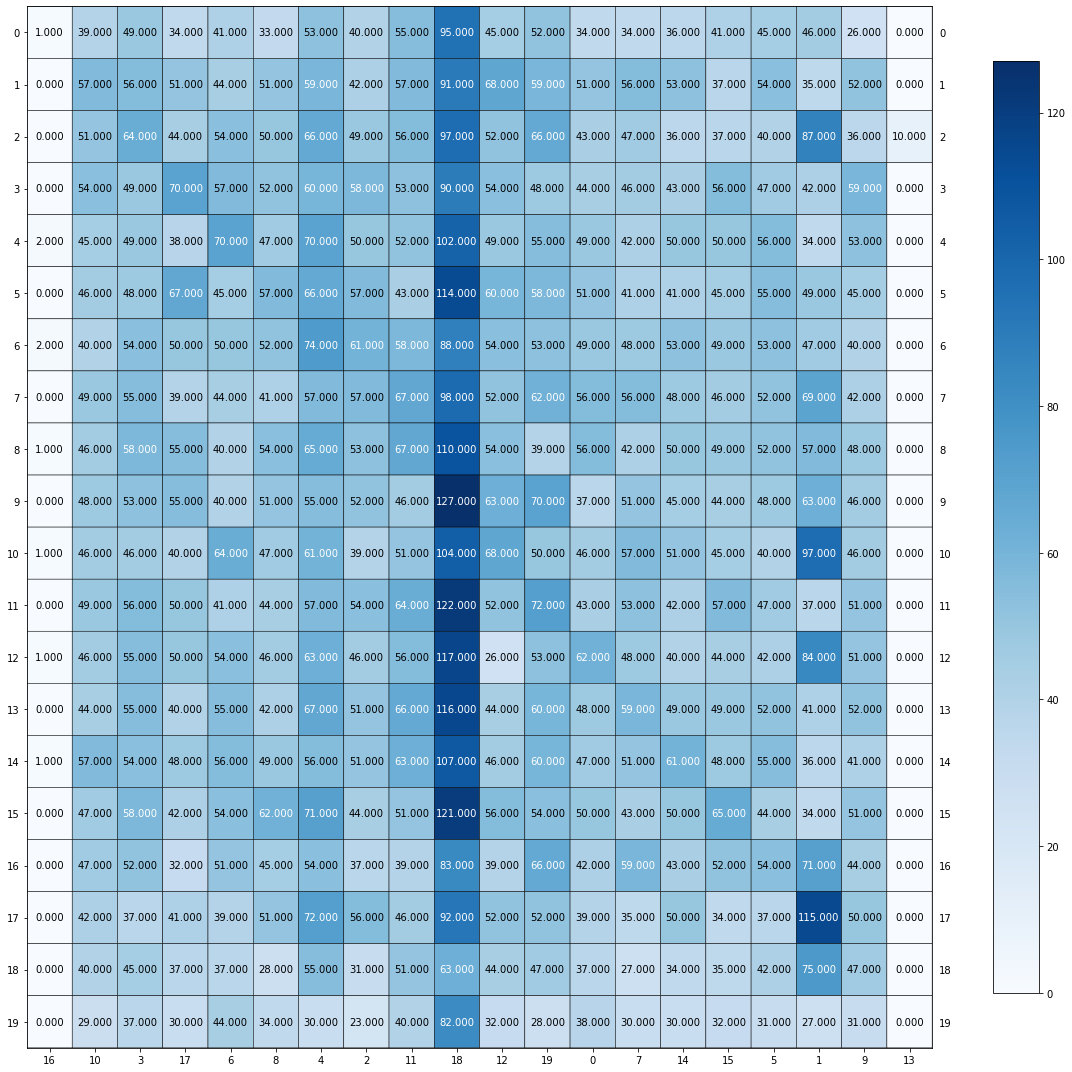
**CM for # Comp = 5:**



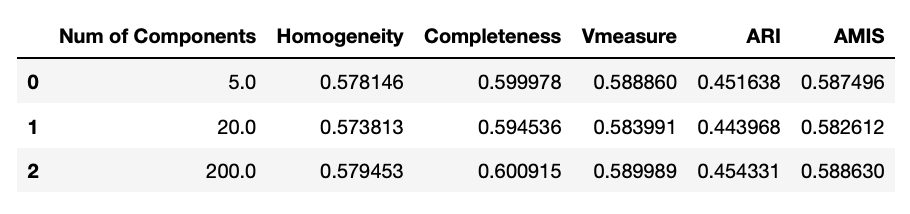
**CM for # Comp = 20:**

****

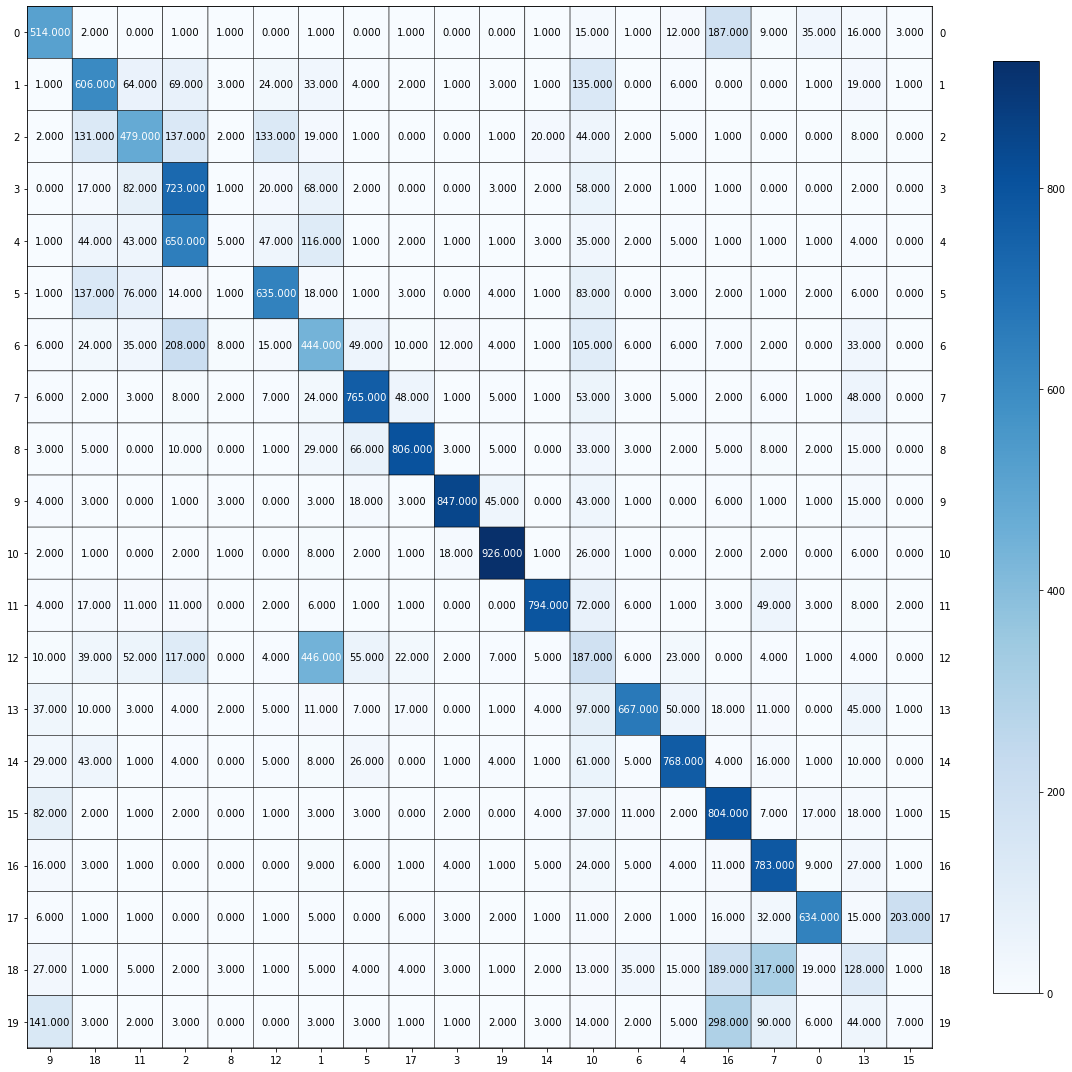
**CM for # Comp = 200:**

****

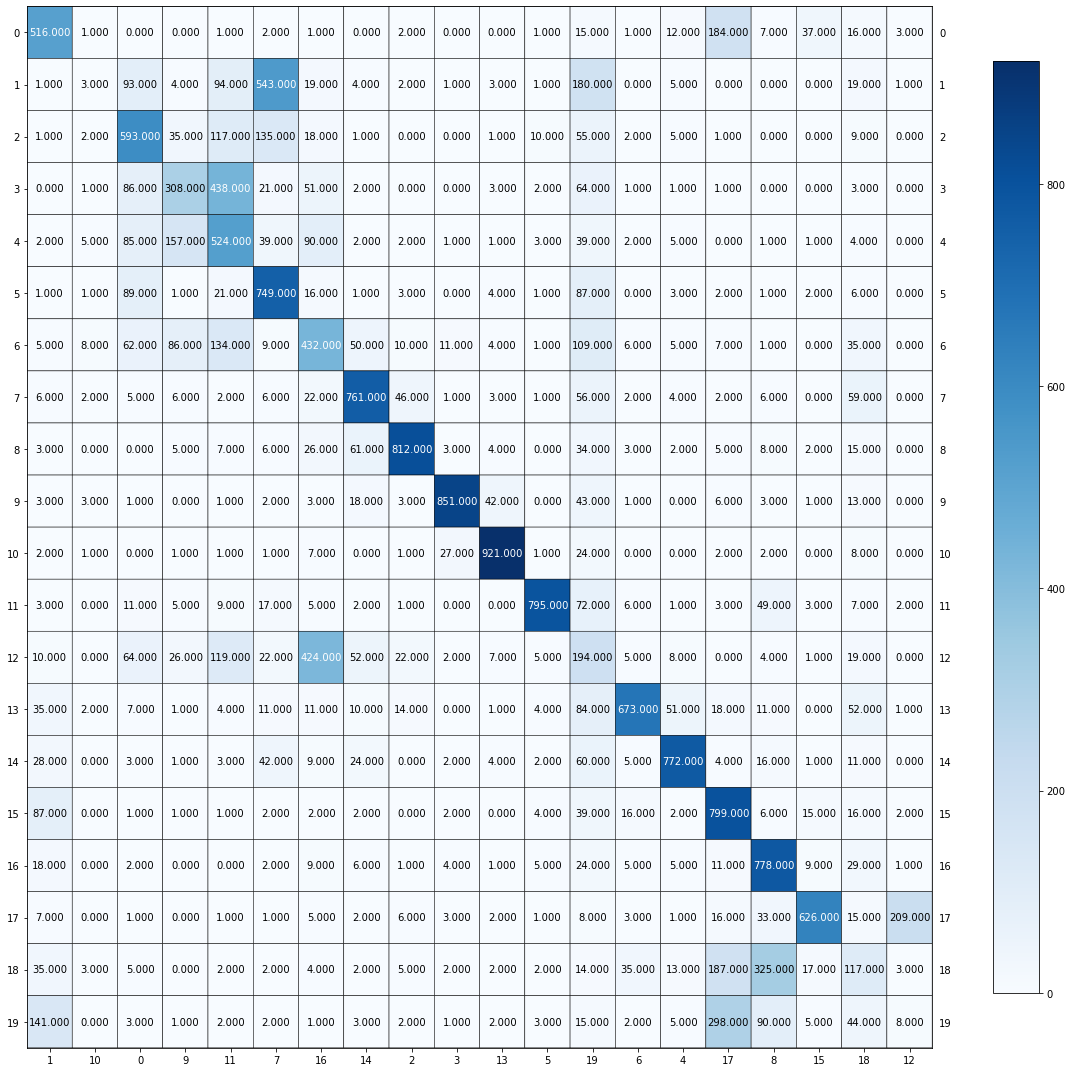
**For Cosine Metric:** 5 Scores for Umap with Cosine Metric for different number of components

****

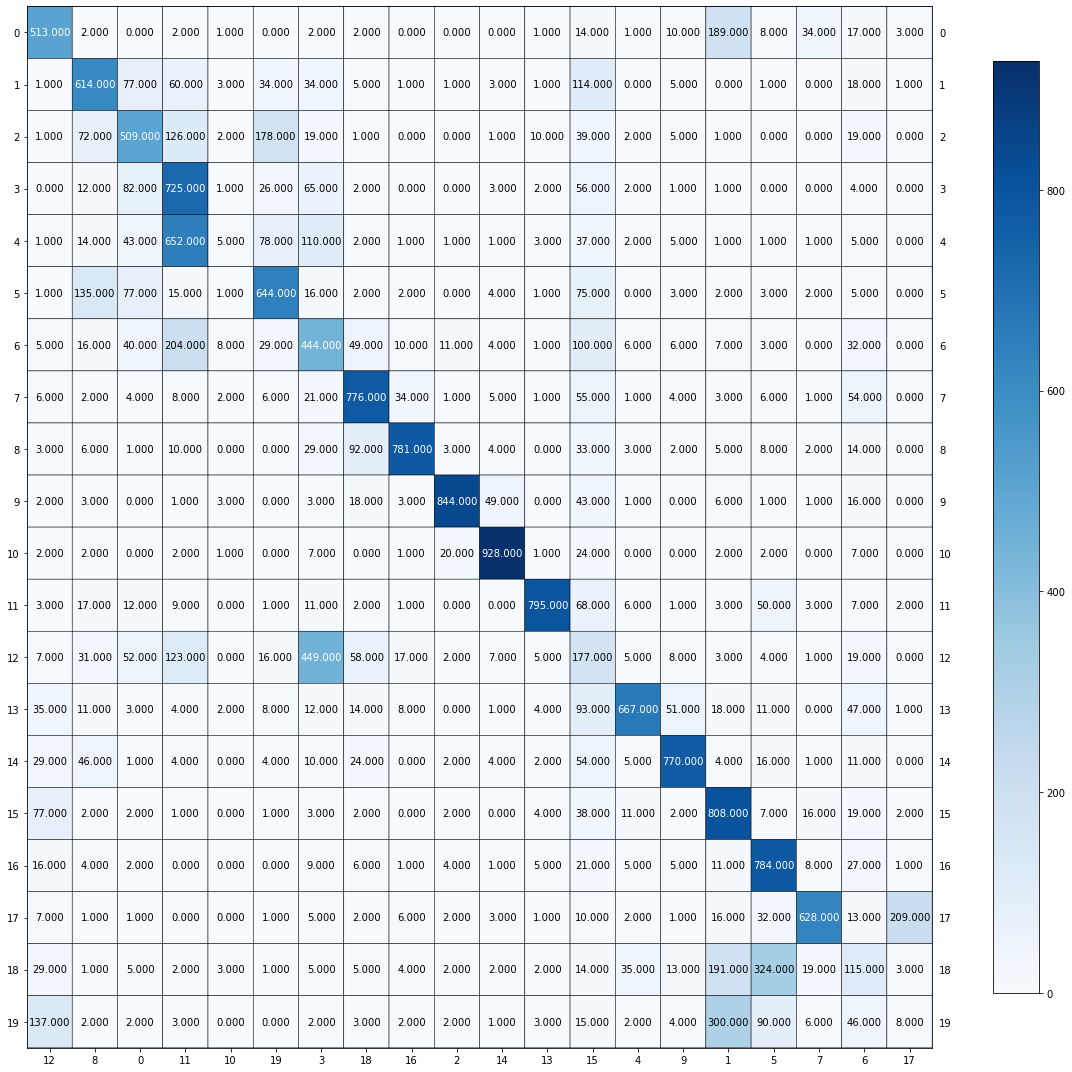
**CM for # Comp = 5:**

****

**CM for # Comp = 20:**

****

**CM for # Comp = 200:**

****

**QUESTION 12:** Analyze the contingency matrices. Which setting works best and why? What about for each metric choice?

**ANS:**

The Cosine Metric with Number of component (k) = 200 performs best with formed cluster having proper segregation and allocation having following Clustering Measures:

\* Homogeneity: 0.579

\* Completeness: 0.60

\* V-Measure: 0.589

\* Adjusted Rand Index: 0.454

\* Adjusted Mutual Information: 0.588

On analyzing the contingency matrices, we get to observe that Euclidean matrices doesn’t do well with clustering, cluster formed are quite random and are indistinguishable.

The clusters for the cosine matrix appear to be well formed, with distinct segregations and allocations. A few points cannot be assigned to a cluster well, and other clusters have few assignments, demonstrating the importance of managing Kmeans outliers carefully.

**QUESTION 13:** So far, we have attempted K-Means clustering with 4 different representation learning techniques (sparse TF-IDF representation, PCA-reduced, NMF-reduced, UMAP-reduced). Compare and contrast the clustering results across the 4 choices, and suggest an approach that is best for the K-Means clustering task on the 20-class text data. Choose any choice of clustering metrics for your comparison?

**ANS:**

Based on Adjusted Random Index and contingency metrices

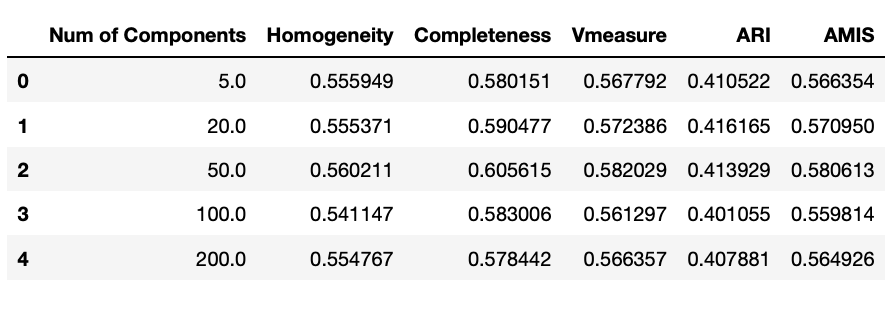
Umap with cosine metrics and with number of components as 200 performed the best having Adjusted Random Index around 0.45 much higher than SVD or NMF which ranged between (0.1, 0.15) and (0.08, 0.1) respectively. It even has highest value for other 4 metrics too.

**Agglomerative Clustering**

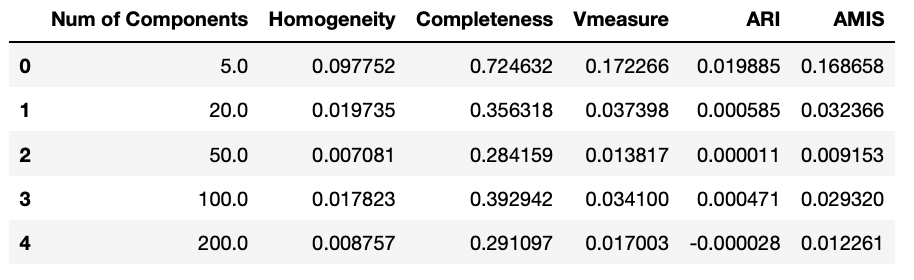
**QUESTION 14:** Use UMAP to reduce the dimensionality properly, and perform Agglom- erative clustering with n\_clusters=20 . Compare the performance of “ward” and “single” linkage criteria. Report the five clustering evaluation metrics for each case.

**ANS:**

**For Ward Linkage:** 5 Scores for Agglomerative Clustering with Ward Linkage for different number of components and number of cluster = 20



**For Single Linkage:** 5 Scores for Agglomerative Clustering with Single Linkage for different number of components and number of cluster = 20



Ward linkage is performing much better as compared to single linkage criteria with having best Adjusted Random Index around 0.413 with 50 number of components.

**HDBSCAN**

**QUESTION 15:** Apply HDBSCAN on UMAP-transformed 20-category data.

Use min\_cluster\_size=100. Vary the min cluster size among 20, 100, 200 and report your findings in terms of the five clustering evaluation metrics - you will plot the best contingency matrix in the next question. Feel free to try modifying other parameters in HDBSCAN to get better performance.

**ANS:**

The Hyper parameter experimented for HDBSCAN with Umap are :

\* Min cluster sizes = [20,100, 200]

\* Min samples = [20,100, 200]

\* Epsilon values = [0.5, 0.8, 5]

\* Number of components = [5,20,200]

The best Hyper parameter and corresponding 5 scores are following:

\* Min cluster sizes = 20

\* Min samples = 100

\* Epsilon values = 0.5

\* Number of components = 5

\* Homogeneity: 0.428

\* Completeness: 0.631

\* V-Measure: 0.510

\* Adjusted Rand Index: 0.222

\* Adjusted Mutual Information: 0.5096

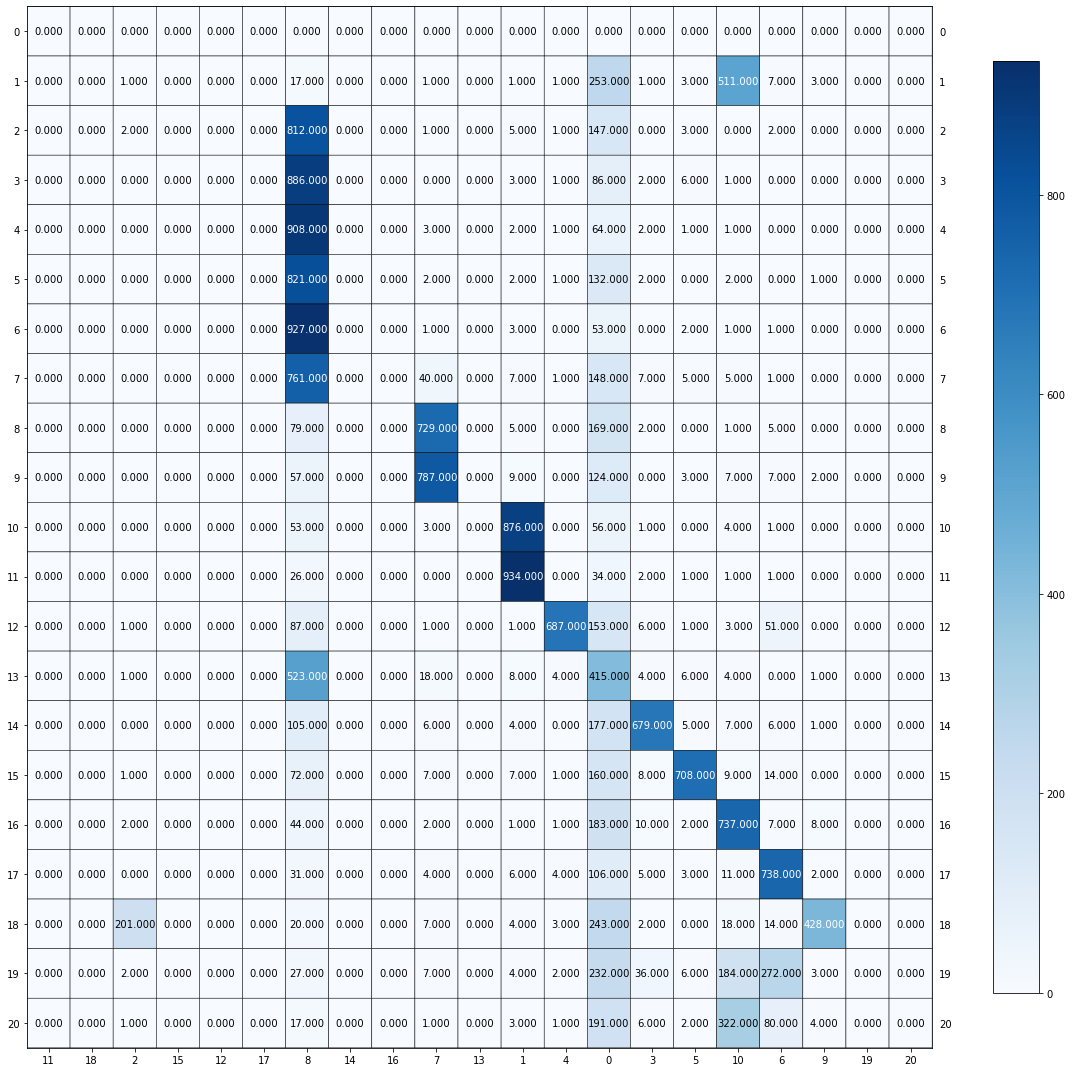
**QUESTION 16:** Plot the contingency matrix for the best clustering model from Question 15.How many clusters are given by the model? What does “-1” mean for the clustering labels? Interpret the contingency matrix considering the answer to these questions.

**ANS:**

CM for best model for Q15 is plotted below

The data points that are regarded as outliers or noisy samples by the clustering algorithm are indicated by the clustering labels "-1". They are not a part of any cluster.

Total 21 clusters are formed. Although certain data points are ignored by the algorithms because they are deemed outliers.



**QUESTION 17:** Based on your experiments, which dimensionality reduction technique and clustering methods worked best together for 20-class text data and why? Follow the table below. If UMAP takes too long to converge, consider running it once and saving the intermediate results in a pickle file.

**ANS:**

**For KMeans Clustering:** Best hyperparameter found and corresponding 5 scores are

\* Best Number of Clusters: 20

\* Best Reduction Method: UMAP with 5 number of components

\* Homogeneity: 0.5757

\* Completeness: 0.6034

\* V-Measure: 0.589

\* Adjusted Rand Index: 0.4460

\* Adjusted Mutual Information: 0.587

**For Agglomerative Clustering:** Best hyperparameter found and corresponding 5 scores are

\* Best Number of Clusters: 20

\* Best Reduction Method: UMAP with 5 number of components

\* Homogeneity: 0.559

\* Completeness: 0.594

\* V-Measure: 0.5761

\* Adjusted Rand Index: 0.4270

\* Adjusted Mutual Information: 0.5746

**For HDBSCAN:** Best hyperparameter found and corresponding 5 scores are

\* Best min clusters size: 200

\* Best Reduction Method: UMAP with 5 number of components

\* Homogeneity: 0.4327

\* Completeness: 0.6385

\* V-Measure: 0.5155

\* Adjusted Rand Index: 0.2247

\* Adjusted Mutual Information: 0.5145

From above results, the best combinations in decreasing order of average clustering measures are as follows:

1). KMeans + UMAP

2). Agglomerative clustering + UMAP

3). HDBSCAN + UMAP

**QUESTION 18:** Extra credit: If you can find creative ways to further enhance the clustering performance, report your method and the results you obtain.

**ANS:**

Two approaches were taken:

In first, we used a combination of SVD and UMAP. The 5 dimensionally reduced features were taken from both and multiplied together element wise. We also tried creating new feature by concatenating both. The idea was that feature space created might map data point far apart creating better clusters.

This approach didn’t give better results.

Score on multiplying the SVD and UMAP feature map

\* Homogeneity: 0.2834

\* Completeness: 0.3147

\* V-Measure: 0.2982

\* Adjusted Rand Index: 0.101

\* Adjusted Mutual Information: 0.295

Score on concatenating the SVD and UMAP feature map

\* Homogeneity: 0.352

\* Completeness: 0.442

\* V-Measure: 0.392

\* Adjusted Rand Index: 0.1066

\* Adjusted Mutual Information: 0.3902

In the second approach we made changes in the data itself.

A) Data without removing header and footer

This was tested on best two combinations from previous results i.e., UMAP with K-means and SVD with K-Means were tried.

\* Number of Components for SVD and UMAO: 50

\* Number of clusters: 20

\* min\_df: 3 with TF-IDF representation.

1. UMAP + KMeans:

\* Homogeneity: 0.5187

\* Completeness: 0.54338

\* V-Measure: 0.53077

\* Adjusted Rand Index: 0.53077

\* Adjusted Mutual Information: 0.52921

2. SVD + Kmeans:

\* Homogeneity: 0.5205

\* Completeness: 0.5431

\* V-Measure: 0.531

\* Adjusted Rand Index: 0.417

\* Adjusted Mutual Information: 0.5300

B. Using min\_df = 5 with TF-IDF representation.

\* Homogeneity: 0.53253

\* Completeness: 0.5512

\* V-Measure: 0.54173

\* Adjusted Rand Index: 0.43478

\* Adjusted Mutual Information: 0.5402

b). SVD + Kmeans:

\* Homogeneity: 0.3531

\* Completeness: 0.44197\* V-Measure: 0.54173

\* Adjusted Rand Index: 0.3926

\* Adjusted Mutual Information: 00.390

The Adjusted Random index for this approach is similar to the results without including the header and footer, but the results of all metrics were improved slightly on average.