EE 219
Large-Scale Data Mining: Models and Algorithms

Project 4
Clustering
Winter 2017

Guanqun Mao, Jianing Liu
204777289 804759999

March 8th 2017

Introduction

In this project, we study clustering algorithms, a group of unsupervised methods for finding groups of data point that have similar representations in a proper space. When using clustering, we assume that no priori labeling of the data points is available. K-means clustering iteratively groups data points into regions characterized by a set of cluster centroids. Each data point is assigned to the cluster with the nearest centroid.

Part 1

Similar to the problems in Project 2, we use TfidfVectorizer to transform the data into two classes: computer technology and recreational activity. For preprocessing, we exclude common stop words and punctuations in English by using Lancaster stemmer. The data contains 7882 documents and 80711 terms. We will reduce terms to 27626 by setting min_df to 2 in TfidfVectorizer().

Part 2

In this part apply K-means clustering with k = 2. Then we calculate the confusion matrix to see how well the clusters match the ground truth table. Table 2.1 shows that the cluster with actual label rec can be accurately identified by K-means clustering algorithm. However, K-means does not perform well on the other cluster.

	Predicted comp	Predicted rec
Actual comp	971	589
Actual rec	56	1534

Table 2.1 Confusion Matrix

Then we perform four measurement of purity on the processed data with respect to ground truth, which are:
Homogeneity score: a measure of how purely clusters contain only data points that belong to a single class
Completeness score: a clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster
Adjusted rand score: the Rand Index computes the similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different

clusters in the predicted and true clusterings
Adjusted mutual info score: it measures mutual information between the cluster label distribution and the ground truth label distributions. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared

Init	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
K means	0.326231701873	0.358149975875	0.348473141557	0.341818484221

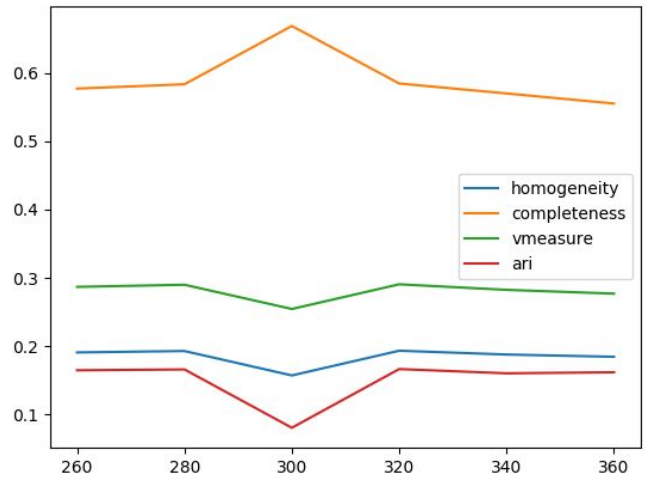
Table 2.2 Purity Measurement

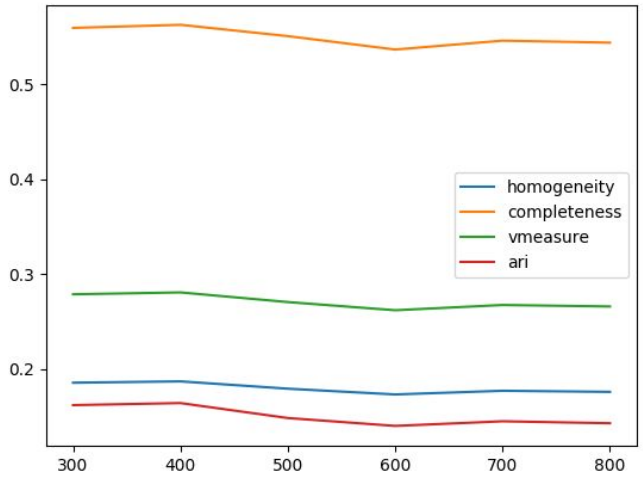
Table 2.2 shows the purity measurement. Combined with the confusion table, we can see that K-means clustering does not give very satisfying results. This is due to the fact that K-means does not perform well in high dimensions. This poor performance is due to the sparse nature of data, because K-means is based on the minimized distances in Euclidean sense.

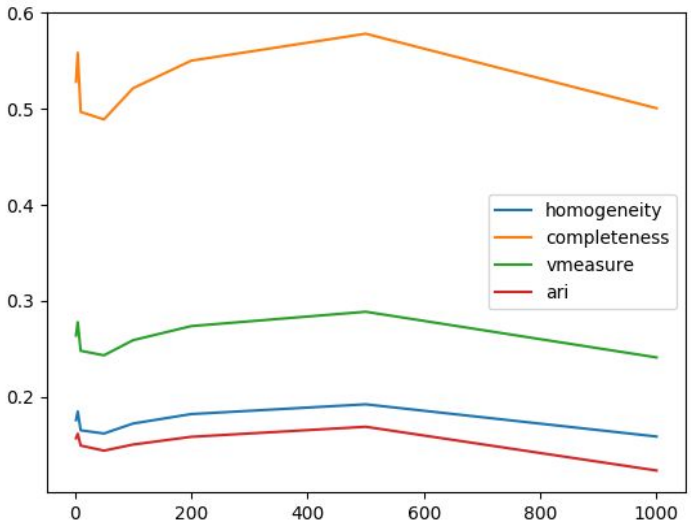
Part 3

Truncated SVD/PCA

After normalizing the data, we apply LSI to reduce dimensions with different dimension values. Figure 3.1 shows the purity measurements.


0.6
0.5
0.4
0.3
0.2
0.1
260
280
300
320
340
360
homogeneity
completeness
vmeasure
ari


0.5
0.4
0.3
0.2
300
400
500
600
700
800
homogeneity
completeness
vmeasure
ari



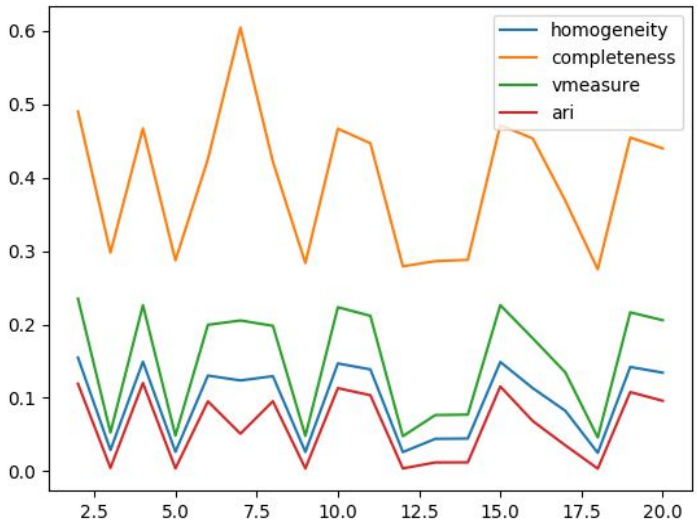
Dimension	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
5	0.185	0.558	0.162	0.185
50	0.162	0.489	0.144	0.162
100	0.172	0.521	0.151	0.172
300	0.159	0.669	0.081	0.159
500	0.192	0.578	0.169	0.192

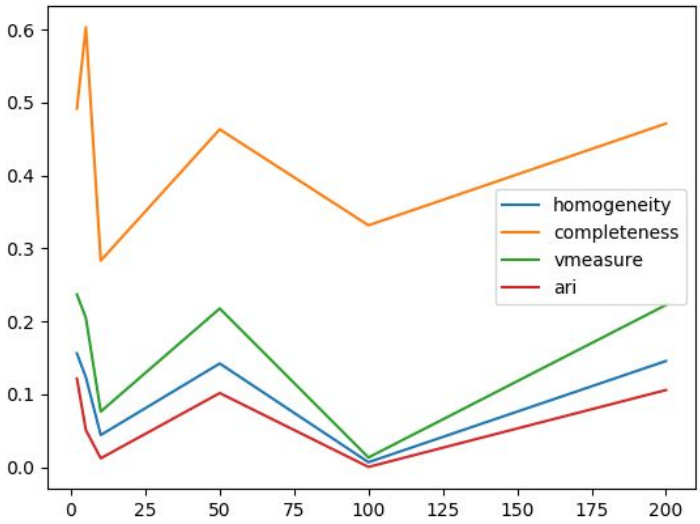
Table 3.1 Purity Measurements

Based on different values of dimension, we found that when the dimension is 300, all four measurements show fairly good results. Therefore we choose 300 as the optimal dimension.

NMF

NMF finds two non-negative matrices (W,H) whose product approximates the non-negative matrix X. It can be used for dimension reduction as well.
Similarly, we sweep over different values of dimensions. The purity measurements are shown in Table 3.2.





Dimension	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
2	0.155	0.490	0.119	0.155
5	0.027	0.287	0.004	0.027
7	0.124	0.604	0.051	0.124
15	0.149	0.472	0.116	0.149

20	0.134	0.440	0.096	0.134

Table 3.2 Purity Measurements

We will choose 7 as the optimal dimension for NMF.

Question: Why logarithm is a good candidate for your TFxIDF data?
For TFxIDF data, both Term frequency and Inverse document frequency are logarithmically scaled. Logarithmic scale is based on orders of magnitude, rather than a standard linear scale, so each mark on the scale is the previous mark multiplied by a value. We can benefit from using logarithmic scale in two ways. First, it reduces the skewness towards large values. Second, it shows percent change or multiplicative factors. In TFxIDF, the influence of very large or very small values (very rare words) is amortised by using logarithm. Furthermore, it's a convention that people intuitively perceive scoring functions to be somewhat additive. Using logarithms will make probability of different independent terms from $P(A,B)=P(A)P(B)$ to look more like $\log(P(A,B))=\log(P(A))+\log(P(B))$. Therefore, logarithm is a good candidate for TFxIDF data because of nature of our data points and that of K-means clustering.

By projecting the data points onto a 2D plane, we can see that the majority of the data points are located densely around the original of the coordinate system. Since K-means is based on the Euclidean distance, distribution clustered like this one will make it difficult for this algorithm to assign data points to different clusters. Applying a logarithm transform, we can scale the data points, essentially spreading them over the plane, and give K-mean clustering a better starting point.

Part 4

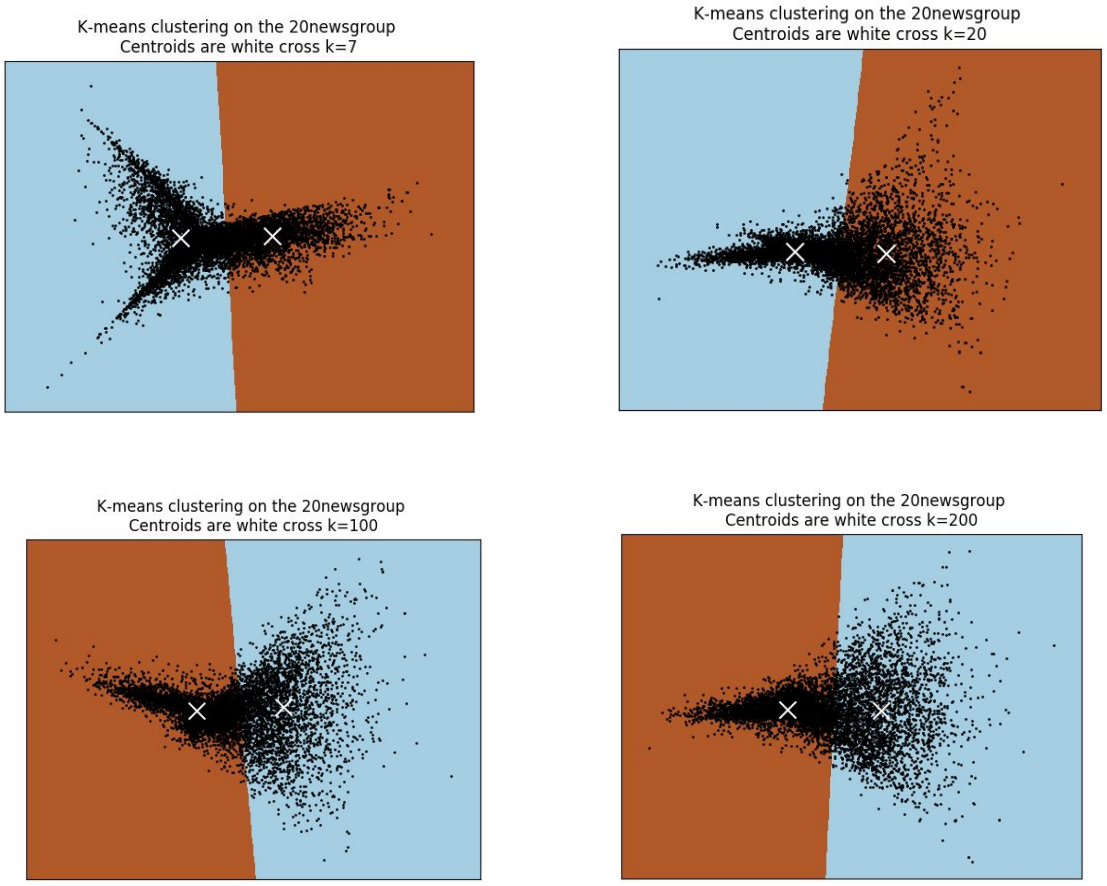


Figure 4.1 Clustering on two labels

The figures show that on a 2D plane, the data points are densely clustered around the original, making it difficult for K-means to assigning them to different clusters after dimension reduction.

Question: Why a non-linear transform is useful?
In this case, a non-linear transform will be useful, as we experimented with different non-linear functions and found that logarithm provides the best result. By applying non-linear transformation, we turn this problem to a linear problem. Smaller differences of distances between clusters will be scaled bigger when logarithm transform is applied.

Part 5

In this part we will vectorize the entire data set. We set k = 20, because there are 20 subclasses in the original data set. We will also reduce the data dimension accordingly.

Truncated SVD/PCA

We use SVD to reduce dimension and sweep through different space dimensions across [200, 300, 500, 1000] as in Part 3. The result is shown in Figure 5.1.

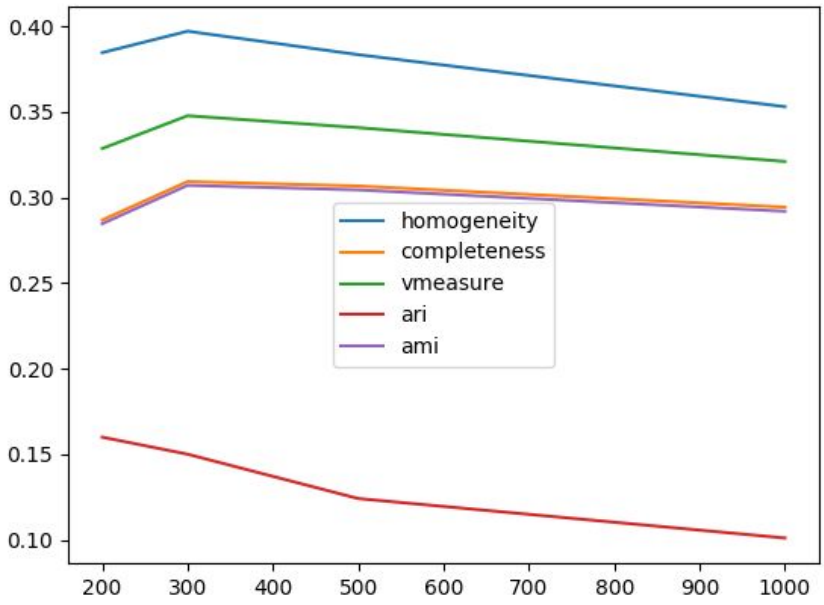


Figure 5.1 Purity Measurements of SVD

We can see that homogeneity, v-measure, completeness and adjusted mutual information score reaches their peaks when the dimension is 300. However, as the dimension increases, adjusted random index decreases. Therefore to make all purity measures at a fairly high level, we still choose 300 to be the optimal dimension, same as in Part 3. Table 5.1 shows all the purity measurements:

Number of Clusters	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
200	0.385	0.287	0.160	0.285
300	0.397	0.309	0.150	0.307
500	0.383	0.307	0.124	0.304
1000	0.353	0.294	0.101	0.292

Table 5.1 Purity Measurements

NMF

Similarly, we use NMF to reduce dimension and sweep through different space dimensions across [2, 5, 10, 15, 20]. The result is shown in Figure 5.2.

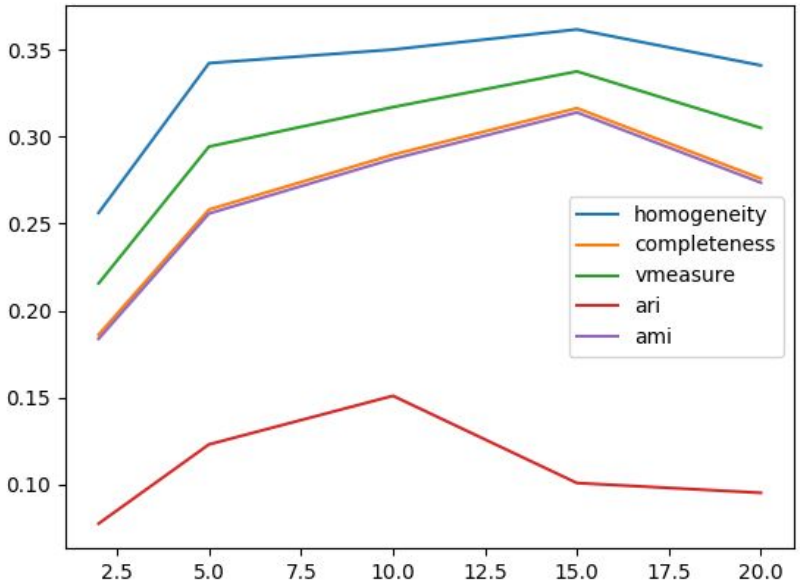


Figure 5.2 Purity Measurements of NMF

We can see that homogeneity, v-measure, completeness and adjusted mutual information score increases as the dimension increases until the dimension reaches 15, after which point all four measurements begin to drop. Adjusted random index, on the other hand, reaches its peak at dimension 10 and begins to drop after that point. Its decreasing rate becomes lower as the dimension reaches 15. Overall, the optimal dimension should be 15 as 3 out of 4 purity measurements we want (homogeneity, completeness and adjusted mutual information score) are the best at this dimension. The purity measurements are recorded in the Table 5.2.

Number of Clusters	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
2	0.256	0.186	0.077	0.184
5	0.342	0.258	0.123	0.256

10	0.350	0.290	0.151	0.287
15	0.362	0.316	0.101	0.314
20	0.341	0.276	0.095	0.274

Table 5.2 Purity Measurements

In this problem, adjusted random index increases until the dimension reaches 10 and begins to drop. Adjusted random index computes the similarity between the clustering labels and ground truth labels. As the dimension increases, it is implied the number of words increases, and thus the gap between clustering labels and ground truth labels begin to enlarge and thus the change in adjusted mutual information score.

Part 6

In this part we set k = 6, as it is the actual number of classes in our dataset. We first map the labels of different classes. Then we perform both SVD and NMF dimension reduction separately on the dataset. Finally, we use the same purity measurements to evaluate the performance of our clustering.

Figure 6.1 and Table 6.1 shows the purity measurements under different dimensions.

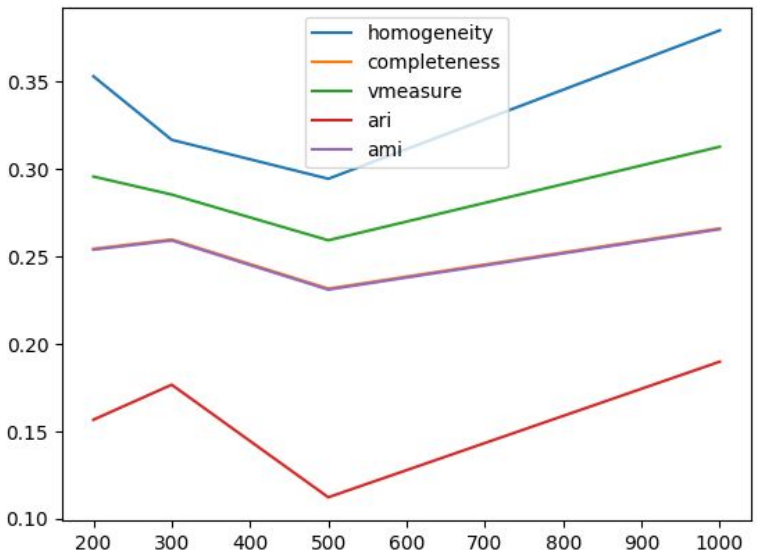


Figure 6.1 Purity Measurements of SVD

Number of Clusters	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
200	0.353	0.254	0.157	0.254
300	0.317	0.260	0.177	0.259
500	0.294	0.232	0.112	0.231
1000	0.379	0.266	0.190	0.266

Table 6.1 Purity Measurements

We found that for SVD, the optimal dimension is actually 1000. This is intuitively easy to comprehend: 1000 words are sufficient enough to summarize an article and identify which category it belongs to.

Then we use NMF for dimension reduction and the purity measurements are shown in Figure 6.2 and Table 6.2.

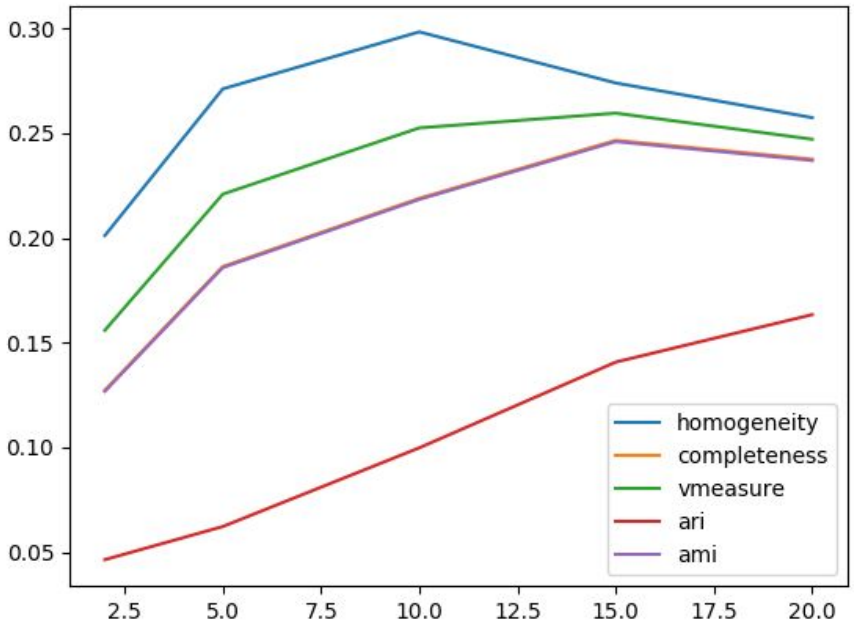


Figure 6.2 Purity Measurements of NMF

Number of Clusters	Homogeneity	Completeness	Adjusted Rand Score	Adjusted Mutual Info Score
2	0.201	0.127	0.046	0.127
5	0.271	0.186	0.062	0.186
10	0.298	0.219	0.100	0.218
15	0.274	0.247	0.141	0.246
20	0.257	0.238	0.163	0.237

Table 6.2 Purity Measurements

The adjusted random index in the only parameter that shows a consistent increasing trend after applying NMF dimension reduction. It is probably because as the dimension increases, more clusters match the actual classes and thus the similarity between clustering labels and ground truth labels. To maintain a balance between relatively high adjusted random index and high values for other purity measures, we still choose 15 to be the optimal dimension.

As we have explained in Part 3, both Term frequency and Inverse Document Frequency are logarithmically scaled (TFxIDF). Logarithmic scale is based on orders of magnitude and therefore can reduce the skewness towards large values. As such, similarly in Part 6, the use of logarithms in TFxIDF enables our processing of datasets to be optimized.

Appendix

Part 3
3. /Library/Frameworks/Python.framework/Versions/2.7/bin/python2.7
/Users/guanqunmao/UCLA/EE219/hw4/src/part3.py
Reduce feature dimension by setting min_df = 2
Number of samples: 7882, number of features: 27626

Desired dimensionality: 2
Explained variance of the SVD step: 0%
Homogeneity: 0.176
Completeness: 0.528
V-measure: 0.264
Adjusted Rand-Index: 0.157
Silhouette Coefficient: 0.595

Desired dimensionality: 5
Explained variance of the SVD step: 1%
Homogeneity: 0.185
Completeness: 0.558
V-measure: 0.278
Adjusted Rand-Index: 0.162
Silhouette Coefficient: 0.303

Desired dimensionality: 10
Explained variance of the SVD step: 2%
Homogeneity: 0.165
Completeness: 0.497
V-measure: 0.248
Adjusted Rand-Index: 0.150
Silhouette Coefficient: 0.192

Desired dimensionality: 50
Explained variance of the SVD step: 8%
Homogeneity: 0.162
Completeness: 0.489
V-measure: 0.243
Adjusted Rand-Index: 0.144
Silhouette Coefficient: 0.051

Desired dimensionality: 100
Explained variance of the SVD step: 13%
Homogeneity: 0.172

Completeness: 0.521
V-measure: 0.259
Adjusted Rand-Index: 0.151
Silhouette Coefficient: 0.032

Desired dimensionality: 200
Explained variance of the SVD step: 20%
Homogeneity: 0.182
Completeness: 0.550
V-measure: 0.274
Adjusted Rand-Index: 0.159
Silhouette Coefficient: 0.020

Desired dimensionality: 500
Explained variance of the SVD step: 35%
Homogeneity: 0.192
Completeness: 0.578
V-measure: 0.289
Adjusted Rand-Index: 0.169
Silhouette Coefficient: 0.013

Desired dimensionality: 1000
Explained variance of the SVD step: 52%
Homogeneity: 0.159
Completeness: 0.501
V-measure: 0.241
Adjusted Rand-Index: 0.124
Silhouette Coefficient: 0.010

Part 5

Number of samples: 7882, number of features: 27626

Desired dimensionality: 200
Homogeneity: 0.385
Completeness: 0.287
Adjusted Rand-Index: 0.160
Adjusted_Mutual_Info_Score: 0.285

Desired dimensionality: 300
Homogeneity: 0.397
Completeness: 0.309

Adjusted Rand-Index: 0.150
Adjusted_Mutual_Info_Score: 0.307

Desired dimensionality: 500
Homogeneity: 0.383
Completeness: 0.307
Adjusted Rand-Index: 0.124
Adjusted_Mutual_Info_Score: 0.304

Desired dimensionality: 1000
Homogeneity: 0.353
Completeness: 0.294
Adjusted Rand-Index: 0.101
Adjusted_Mutual_Info_Score: 0.292

Desired dimensionality: 2
Homogeneity: 0.256
Completeness: 0.186
Adjusted Rand-Index: 0.077
Adjusted_Mutual_Info_Score: 0.184

Desired dimensionality: 5
Homogeneity: 0.342
Completeness: 0.258
Adjusted Rand-Index: 0.123
Adjusted_Mutual_Info_Score: 0.256

Desired dimensionality: 10
Homogeneity: 0.350
Completeness: 0.290
Adjusted Rand-Index: 0.151
Adjusted_Mutual_Info_Score: 0.287

Desired dimensionality: 15
Homogeneity: 0.362
Completeness: 0.316
Adjusted Rand-Index: 0.101
Adjusted_Mutual_Info_Score: 0.314

Desired dimensionality: 20
Homogeneity: 0.341
Completeness: 0.276

Adjusted Rand-Index: 0.095
Adjusted_Mutual_Info_Score: 0.274

Part 6

Number of samples: 7882, number of features: 27626

Desired dimensionality: 200
Homogeneity: 0.353
Completeness: 0.254
Adjusted Rand-Index: 0.157
Adjusted_Mutual_Info_Score: 0.254

Desired dimensionality: 300
Homogeneity: 0.317
Completeness: 0.260
Adjusted Rand-Index: 0.177
Adjusted_Mutual_Info_Score: 0.259

Desired dimensionality: 500
Homogeneity: 0.294
Completeness: 0.232
Adjusted Rand-Index: 0.112
Adjusted_Mutual_Info_Score: 0.231

Desired dimensionality: 1000
Homogeneity: 0.379
Completeness: 0.266
Adjusted Rand-Index: 0.190
Adjusted_Mutual_Info_Score: 0.266

Desired dimensionality: 2
Homogeneity: 0.201
Completeness: 0.127
Adjusted Rand-Index: 0.046
Adjusted_Mutual_Info_Score: 0.127

Desired dimensionality: 5

Homogeneity: 0.271
Completeness: 0.186
Adjusted Rand-Index: 0.062
Adjusted_Mutual_Info_Score: 0.186

Desired dimensionality: 10
Homogeneity: 0.298
Completeness: 0.219
Adjusted Rand-Index: 0.100
Adjusted_Mutual_Info_Score: 0.218

Desired dimensionality: 15
Homogeneity: 0.274
Completeness: 0.247
Adjusted Rand-Index: 0.141
Adjusted_Mutual_Info_Score: 0.246

Desired dimensionality: 20
Homogeneity: 0.257
Completeness: 0.238
Adjusted Rand-Index: 0.163
Adjusted_Mutual_Info_Score: 0.237
