The University of California, Los Angeles
Department of Electrical and Computer Engineering
ECE219 Large-scale Data Mining, Winter 2018

Project 2 Clustering

Xin Jiang (904589261), Zhiyuan Cao (304397496)

Deadline: Feburary 11th, 2018

Question 1 With `min_df=3`, the total number of terms (features) is 27768, with 7882 documents (observations) in total.

Question 2 With `random_state=43`, we have the 5 measures as follows:

- Homogeneity score: 0.752
- Completeness score: 0.755
- V-measure score: 0.754
- Adjusted Rand-Index: 0.832
- Adjusted mutual info score: 0.752

Question 3

(a) Refer to Figure 1.

(b) For 5 measure scores v.s.r and contingency matrices using SVD as the dimension reduction method, refer to Figure 2 and Figure 3; for results using NMF, refer to Figure 4 and Figure 5. The best r for SVD and NMF are 50 and 2 respectively.

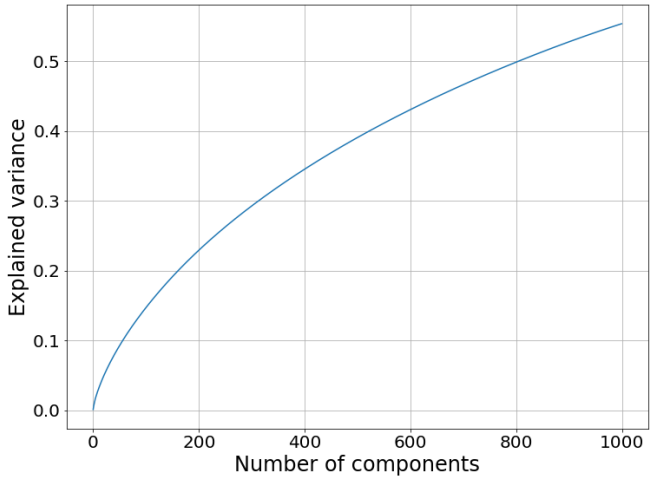

Figure 1: Cumulative explained variance over number of components involved.

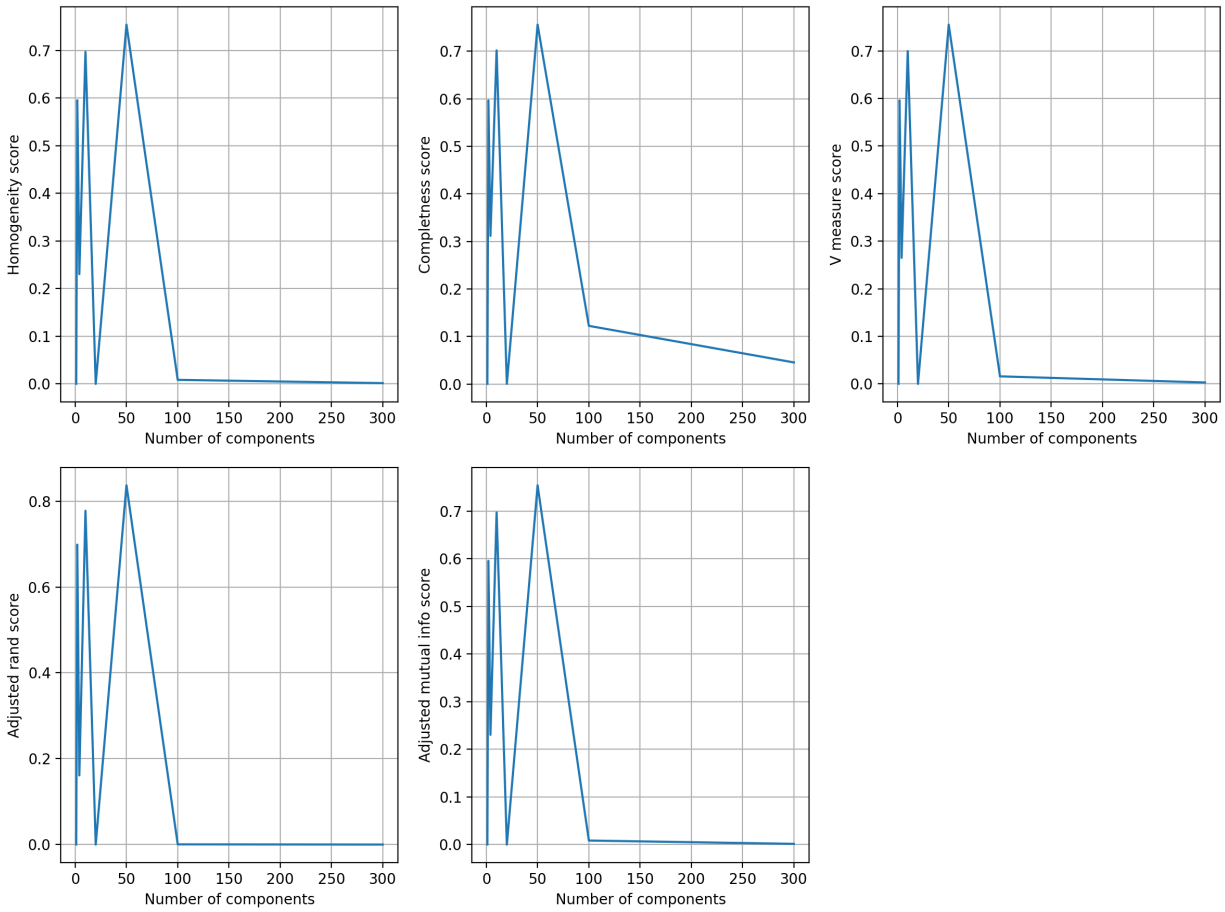

Figure 2: 5 measures using SVD against number of components r.

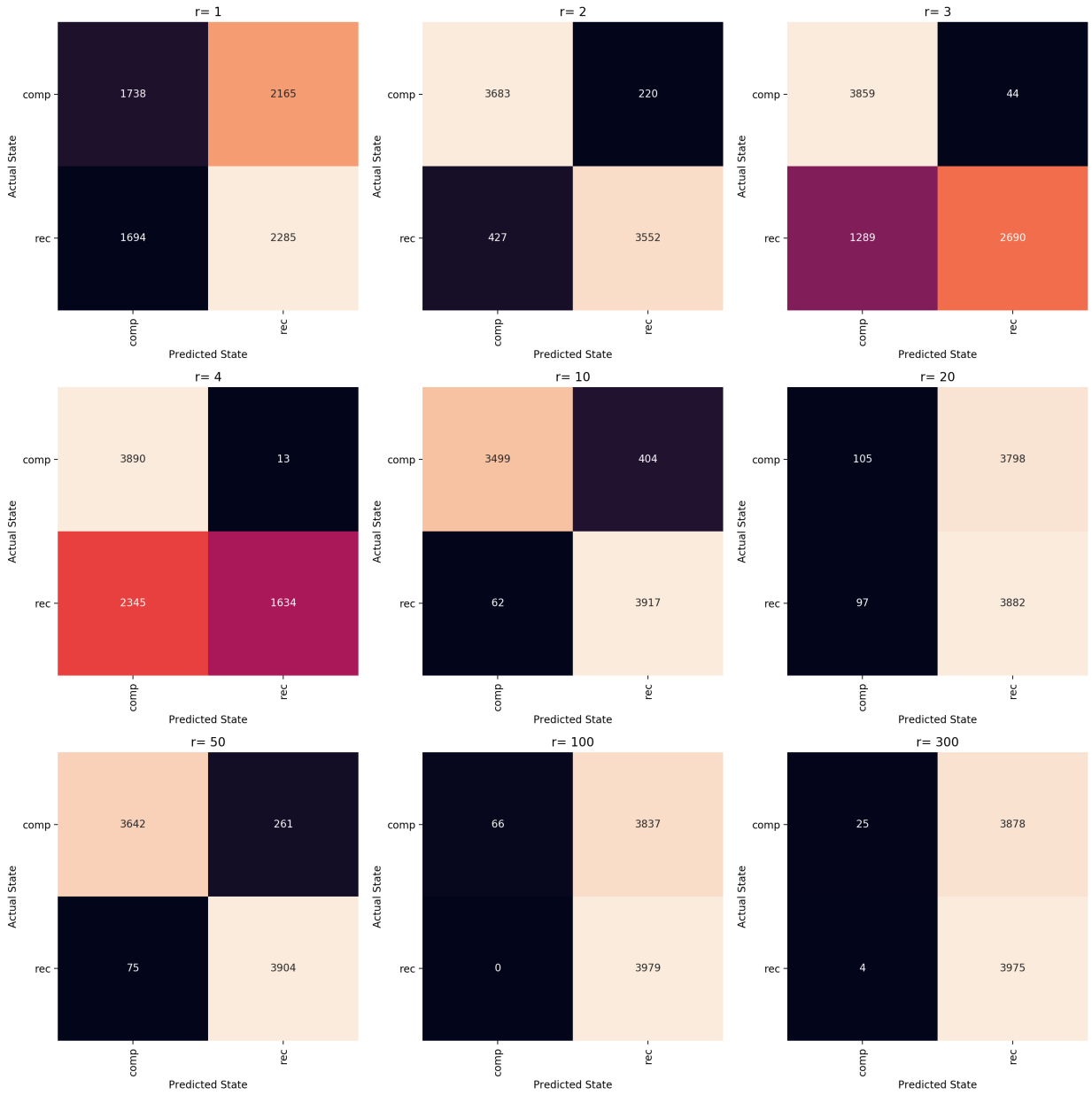

Figure 3: Contingency matrix using SVD with variation on number of components r.

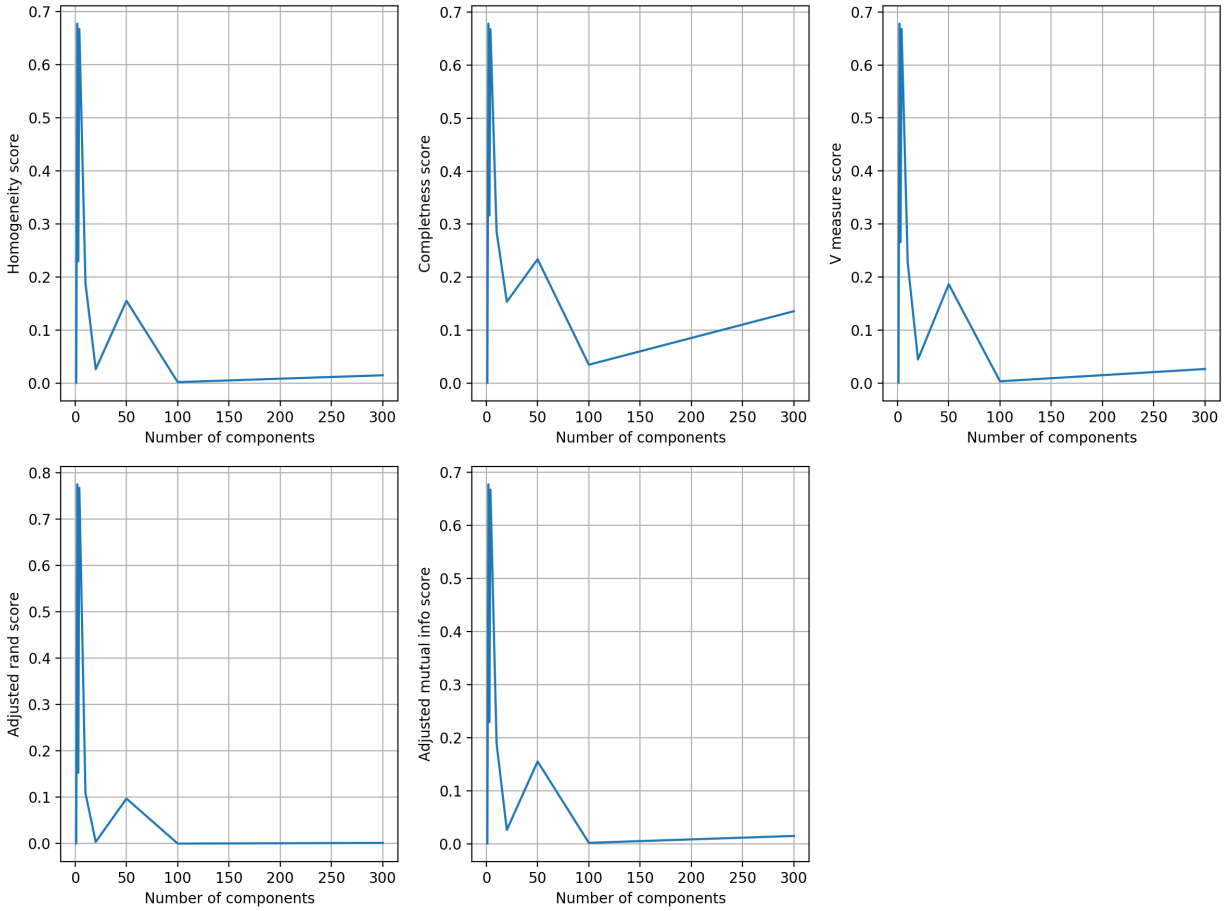

Figure 4: 5 measures using NMF against number of components r.

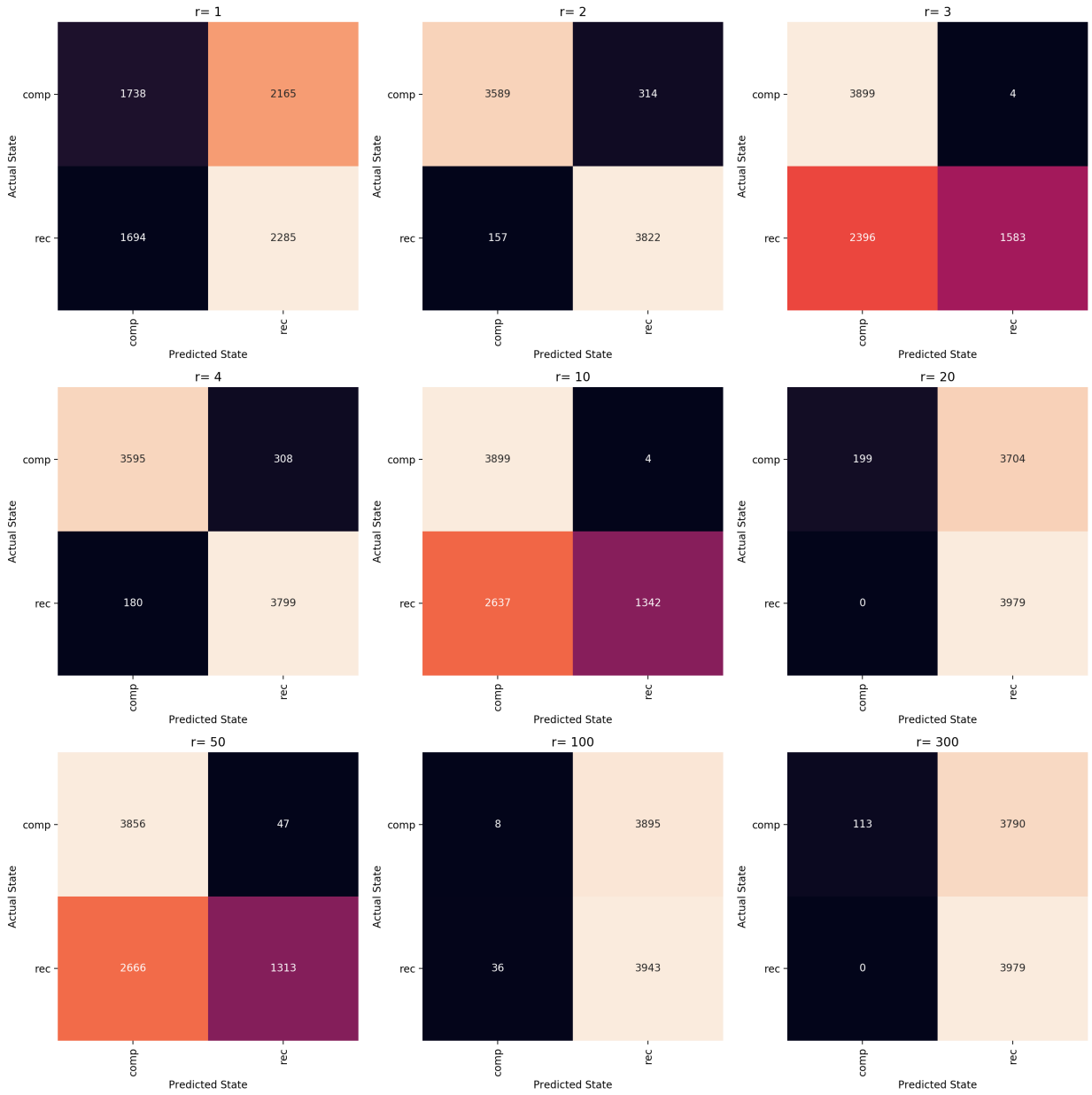

Figure 5: Contingency matrix using SVD with variation on number of components r.

Question 4

(a) Our best result comes from SVD with $r = 50$. Figure 6 shows the distribution of observations by projecting the reduced feature matrix into a 2-dimensional space. As we could observe from the figure, the two classes are mostly separated with overlaps.

(b)

- Figure 7 shows the result if we first perform unit variance scaling. Since the two classes are heavily overlapped, K-means is unable to produce meaningful results.

- Figure 8 shows the distribution of observations with log transformation. As for the general case, log-transformation makes a distribution more "normal" and therefore may have a positive effect on the results. For our case, however, we do not observe an increase, with 5 measures scores close to the unprocessed case.
- Figure 9 and 10 are the results by applying log transformation and unit variance scaling in different orders. Both of them showed minor improvements to log-transformation-only result.

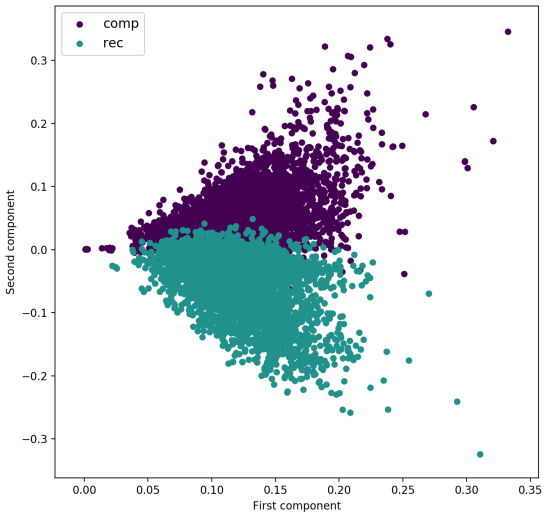

Figure 6: Distribution of observations using SVD with $r = 50$.

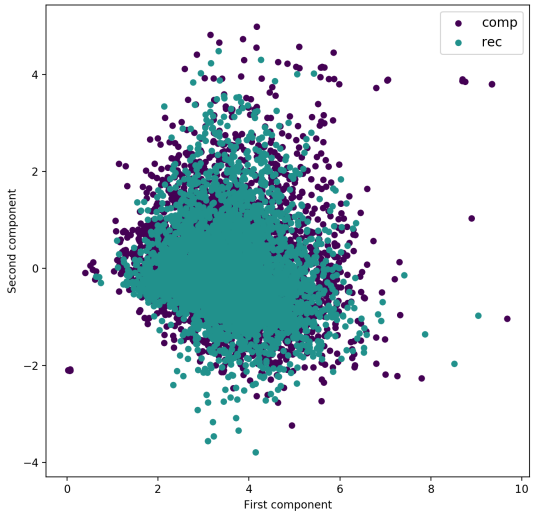

Figure 7: Distribution of observations using SVD with $r = 50$ with unit variance scaling.

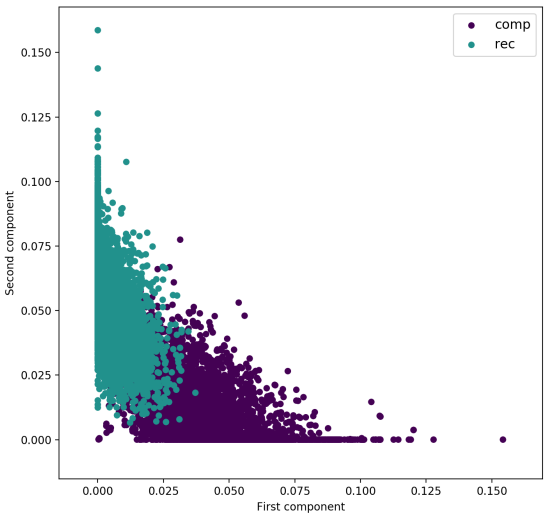

Figure 8: Distribution of observations using SVD with $r = 50$ with log transformation.

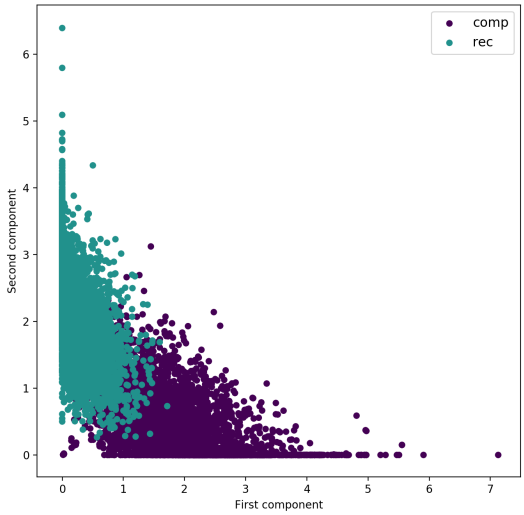

Figure 9: Distribution of observations using SVD with $r = 50$ with first log transformation followed by unit variance scaling.

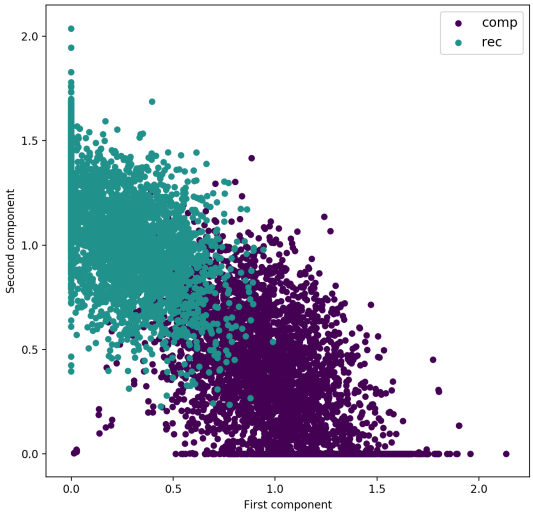

Figure 10: Distribution of observations using SVD with $r = 50$ with first unit variance scaling followed by log transformation.

Question 5 Our best result comes from NMF with $r = 20$. Figure 11 shows the by projecting the reduced feature matrix into a 2-dimensional space. The results are listed as follows.

- Figure 12 shows the result if we first perform unit variance scaling.
- Figure 13 shows the distribution of observations with log transformation.
- Figure 14 and 15 are the results by applying log transformation and unit variance scaling in different orders.

As we can see from all 5 figures, all 20 classes are sitting on top of each other, resulting that there is little K-means can do.

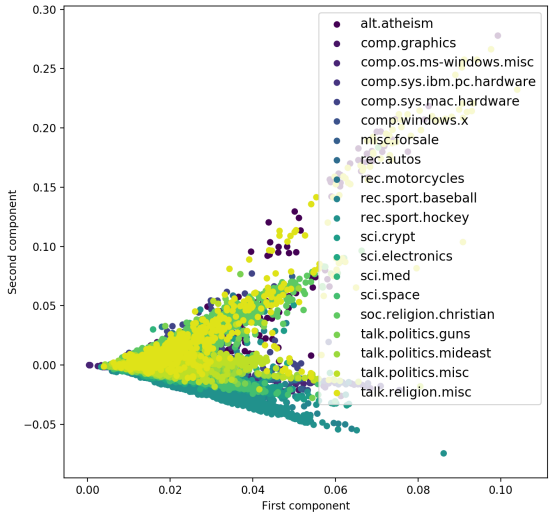

Figure 11: Distribution of observations using NMF with $r = 20$.

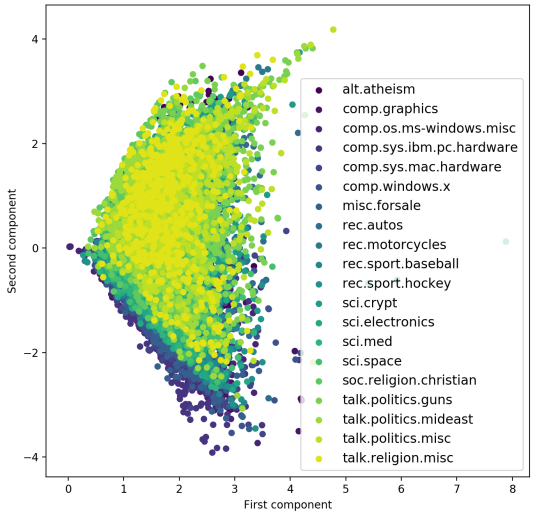

Figure 12: Distribution of observations using NMF with $r = 20$ with unit variance scaling.

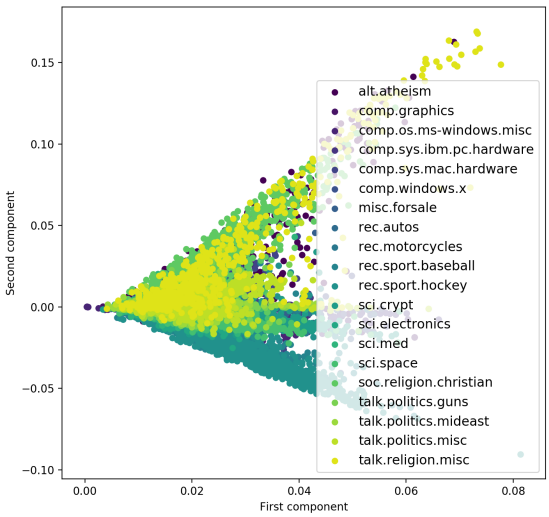

Figure 13: Distribution of observations using NMF with $r = 20$ with log transformation.

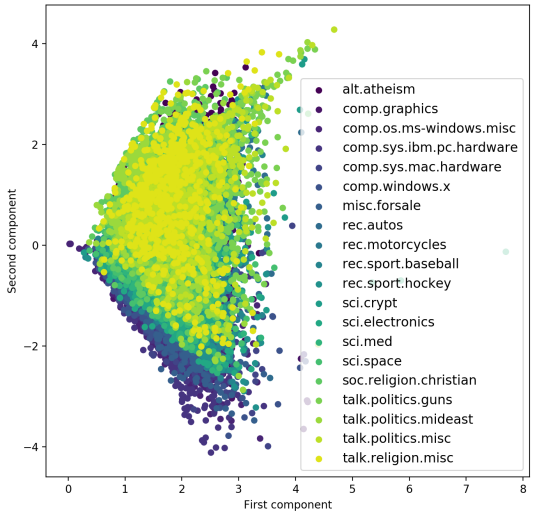

Figure 14: Distribution of observations using NMF with $r = 20$ with first log transformation followed by unit variance scaling.

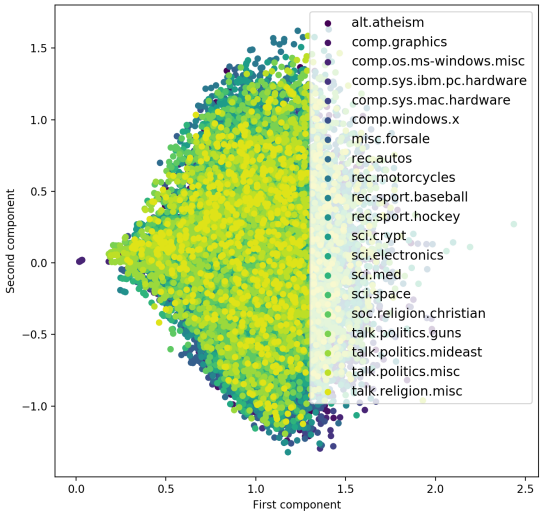

Figure 15: Distribution of observations using NMF with $r = 20$ with first unit variance scaling followed by log transformation.