EE-219 Project 1
Shrey Agarwal (004943082)
Varun Saboo (505028591)
February 12, 2018

Project 2: Clustering

Objective –

To find proper representations of the data, s.t. the clustering is efficient and gives out reasonable results.

To perform K-means clustering on the dataset, and evaluate the performance of the clustering.

To try different preprocess methods which may increase the performance of the clustering.

Dataset –

We work with “20 Newsgroups” dataset. It is a collection of approximately 20,000 documents, partitioned (nearly) evenly across 20 different newsgroups, each corresponding to a different topic. Each topic can be viewed as a “class”.

TASK 1:

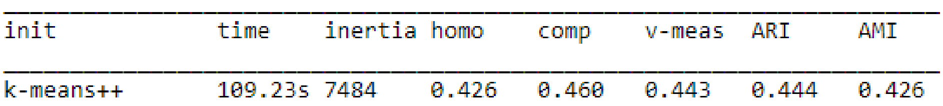
We perform TF-IDF on the documents by performing certain preprocessing steps such as stemming, word tokenizing, special characters removal and stopwords removal.

With min_df = 3, the dimension of the TF-IDF matrix is (7882, 16564), where 7882 are the total number of documents in 8 classes and 16564 are the unique tokens identified.

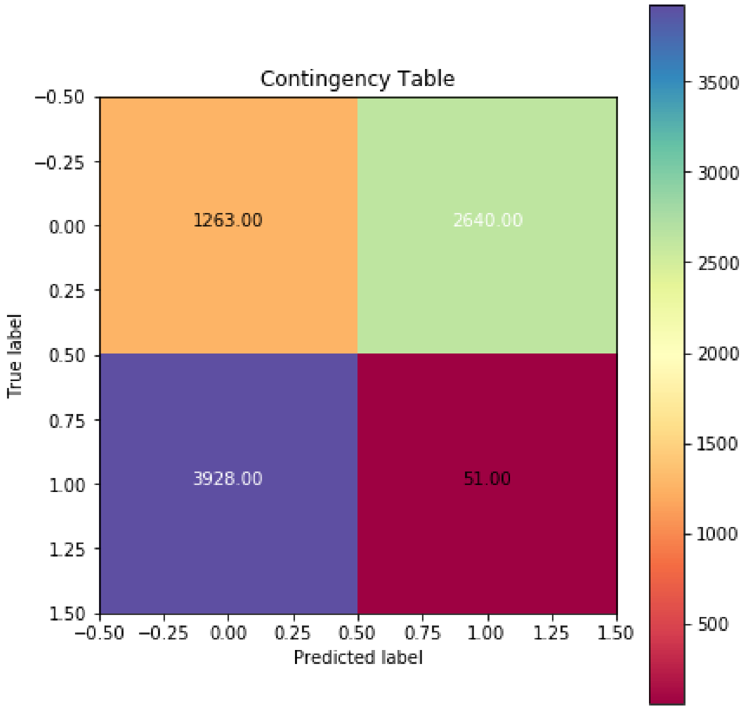
TASK 2:

After applying k-means clustering on the above dataset, we got the following results.

With number of clusters = 2:



```
____________________________________________________________________________
init            time     inertia homo     comp     v-meas   ARI      AMI
____________________________________________________________________________
k-means++       109.23s  7484     0.426    0.460    0.443    0.444    0.426
```



TASK 3:

Preprocess the data:

Dimensionality reduction:

a. LSI (Latent Semantic Indexing):
   After performing LSI on the TF-IDF matrix, following are the obtained results:

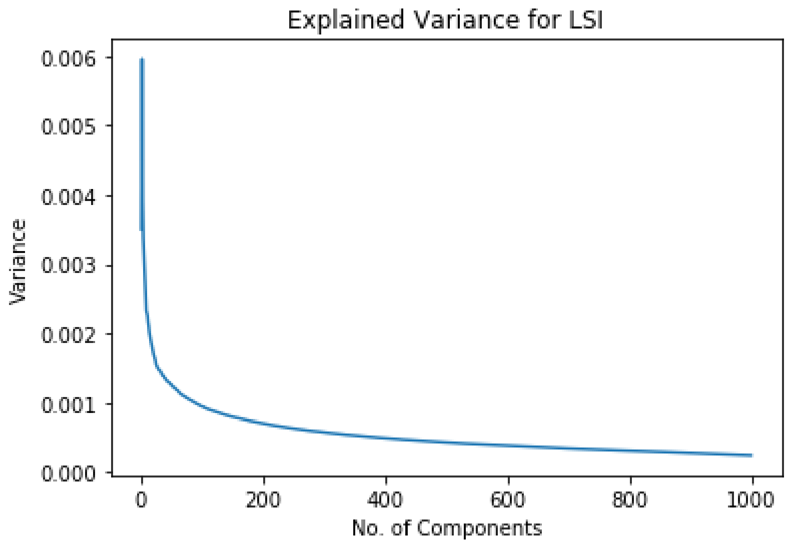
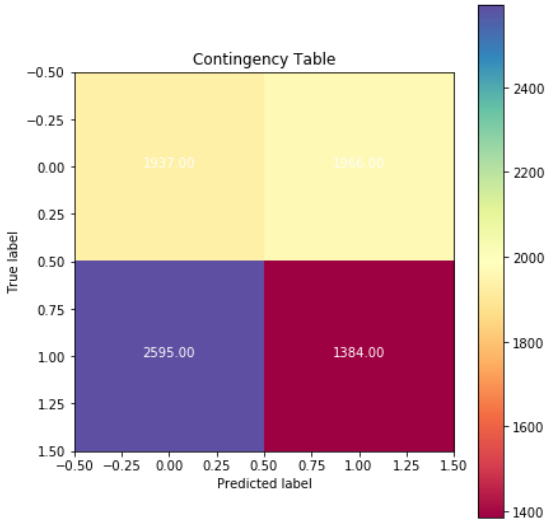
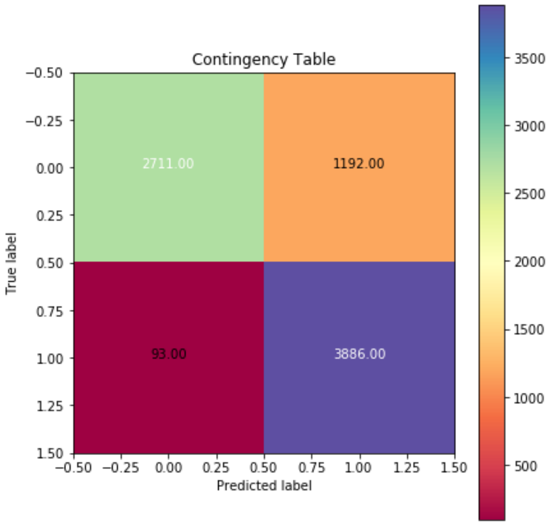


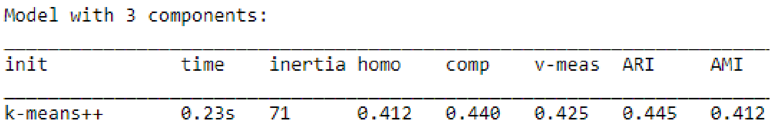
Figure 1: We can observe that the variance decreases as the no. of components increase.

```
Model with 1 components:
______________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________
k-means++       0.24s   9       0.018   0.018   0.018   0.025   0.018
```

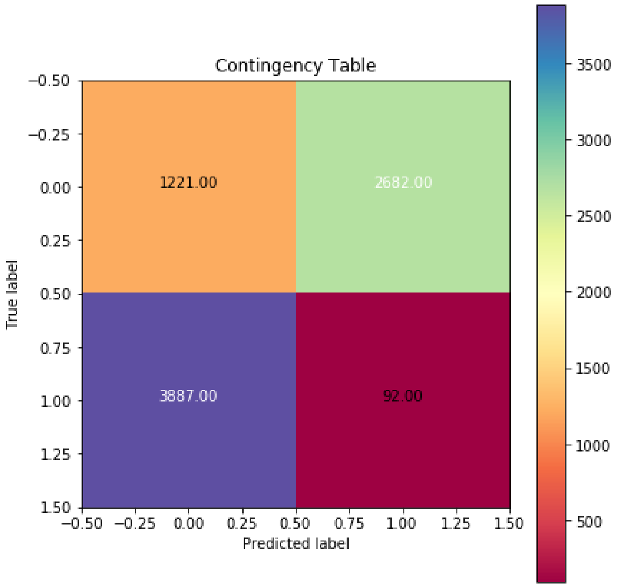


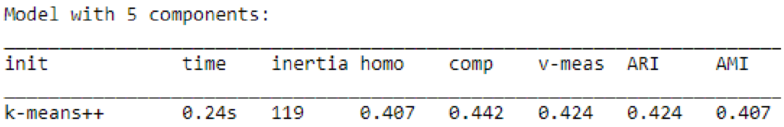
```
Model with 2 components:
______________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________
k-means++       0.40s   41      0.419   0.446   0.432   0.454   0.419
```



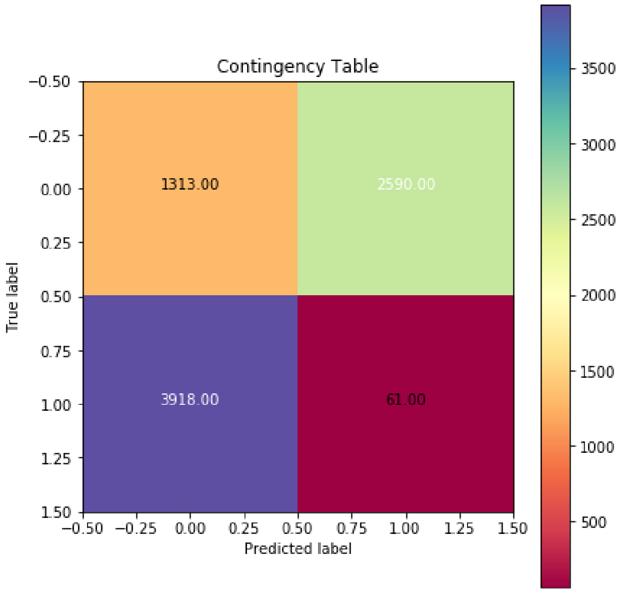


```
Model with 3 components:
______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________________
k-means++       0.23s   71      0.412   0.440   0.425   0.445   0.412
```

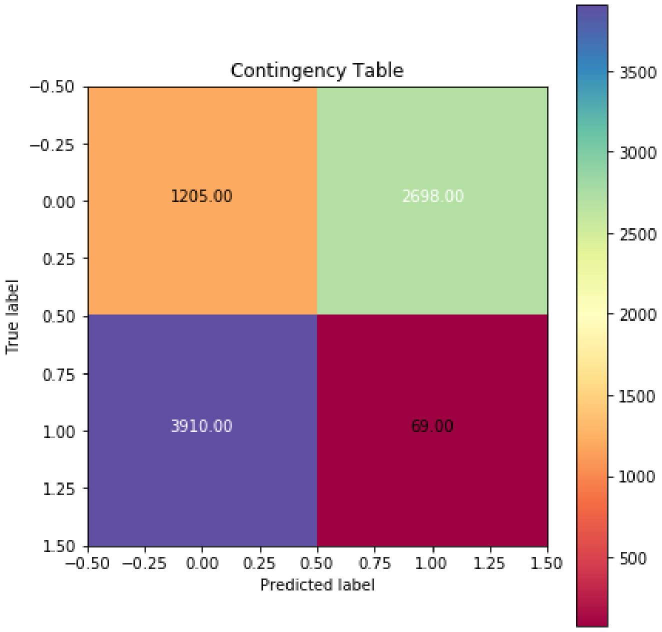




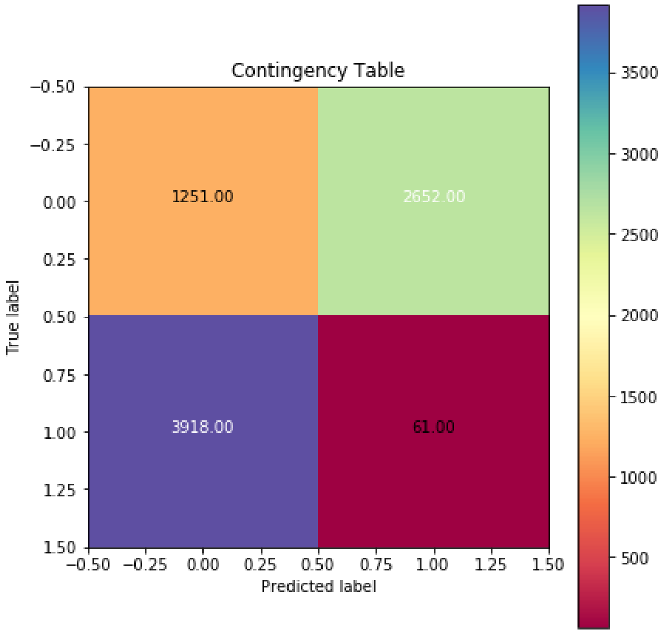
```
Model with 5 components:
______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________________
k-means++       0.24s   119     0.407   0.442   0.424   0.424   0.407
```



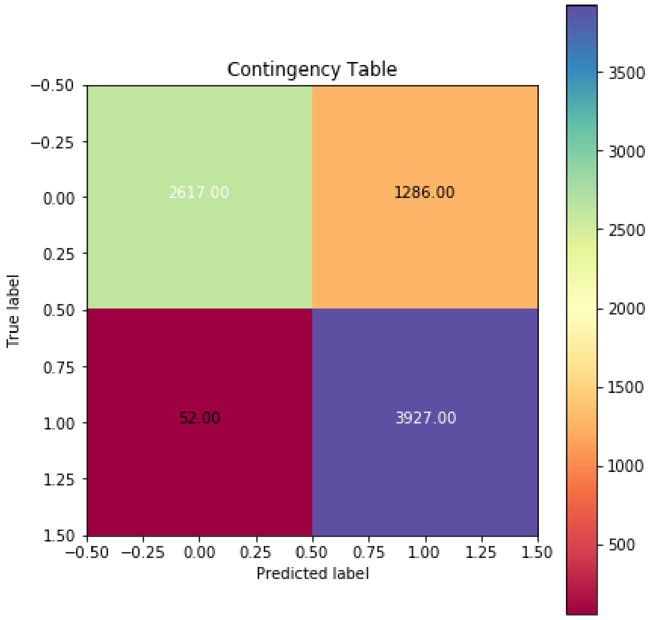
```
Model with 10 components:
______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________________
k-means++       0.56s   216     0.430   0.460   0.444   0.458   0.430
```



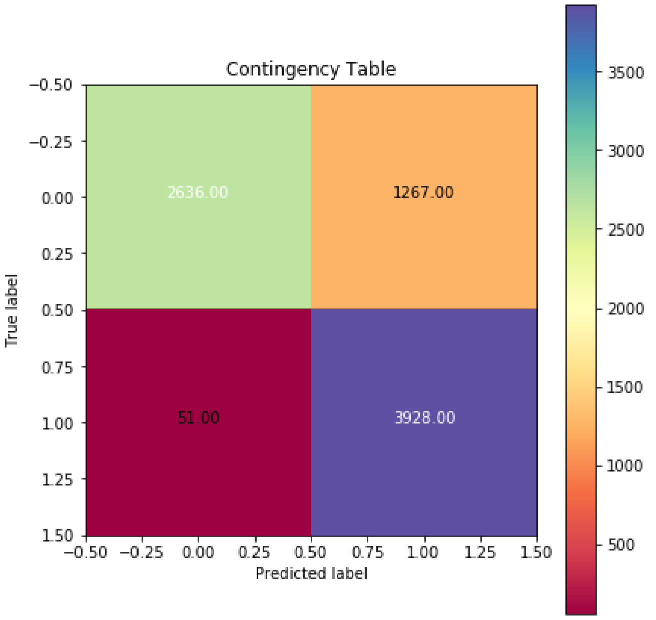
```
Model with 20 components:
______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________________
k-means++       0.74s   363     0.423   0.455   0.439   0.445   0.423
```



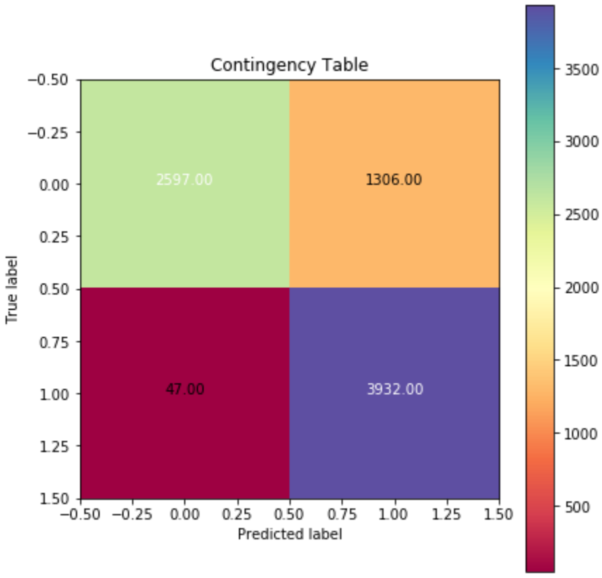
```
Model with 50 components:
___________________________________________________________________________
init             time     inertia homo     comp     v-meas   ARI      AMI
___________________________________________________________________________
k-means++        1.34s    684      0.420    0.455    0.437    0.436    0.420
```

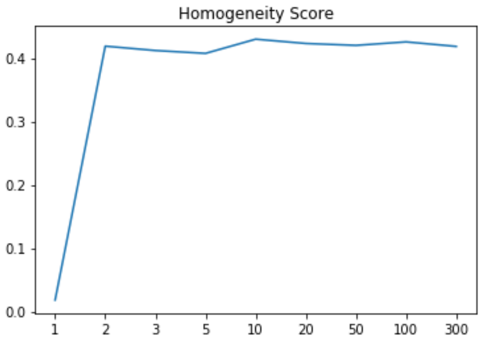


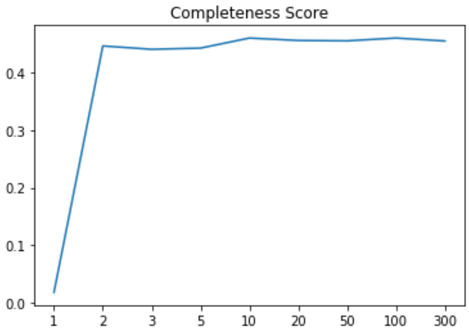
```
Model with 100 components:
___________________________________________________________________________
init             time     inertia homo     comp     v-meas   ARI      AMI
___________________________________________________________________________
k-means++        1.64s    1090     0.425    0.460    0.442    0.443    0.425
```

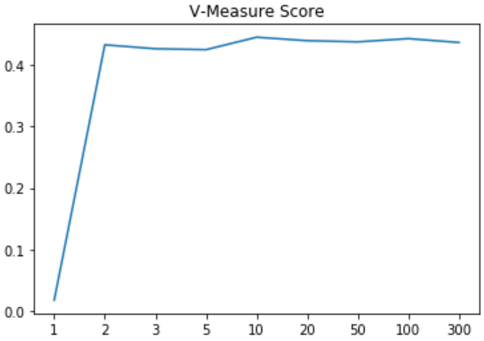


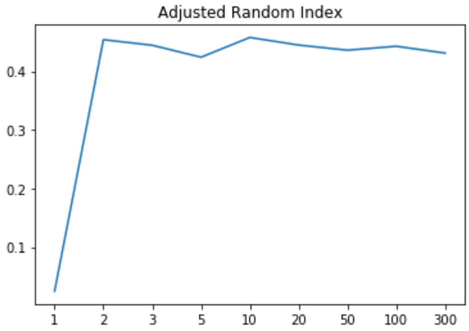
```
Model with 300 components:
_____________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_____________________________________________________________________________
k-means++       5.07s   2172    0.418   0.454   0.436   0.431   0.418
```

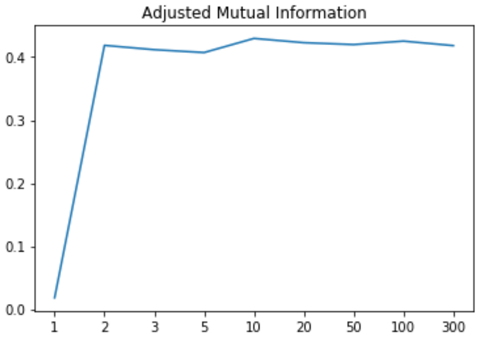






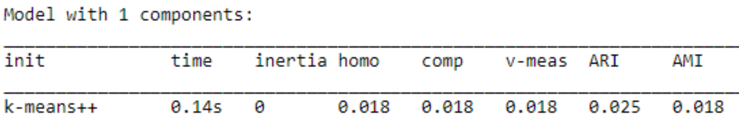




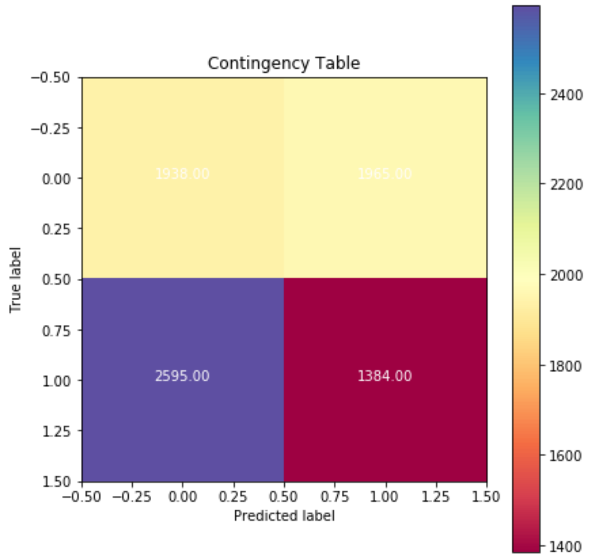


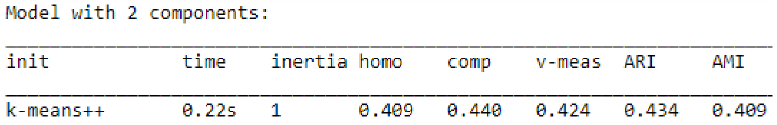
As we can observe from the contingency matrix as well as the 4 measures above, the best r value for LSI is 10. i.e, 10 components are best suitable to represent each document and cluster with maximum accuracy.

b. NMF (Non-Negative Matrix Factorization):
After performing NMF on the TF-IDF matrix, following are the obtained results:

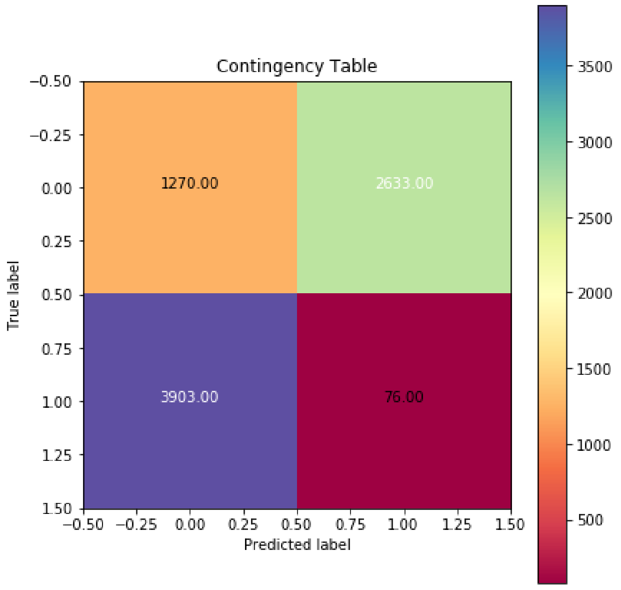


```
Model with 1 components:
_____________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_____________________________________________________________________________________
k-means++       0.14s   0       0.018   0.018   0.018   0.025   0.018
```

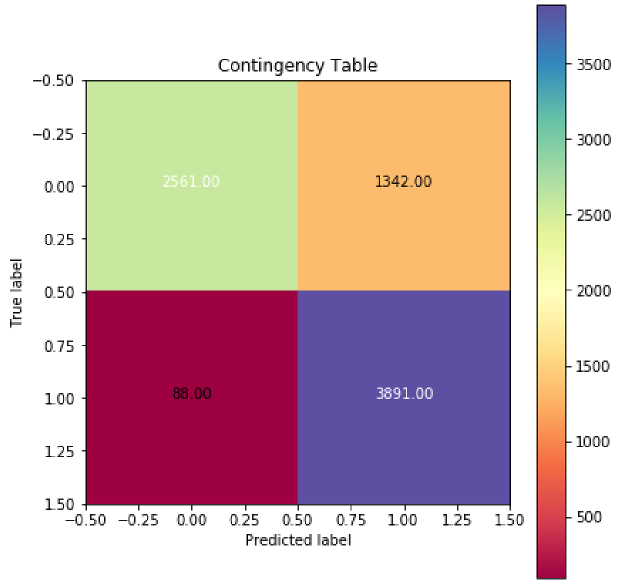




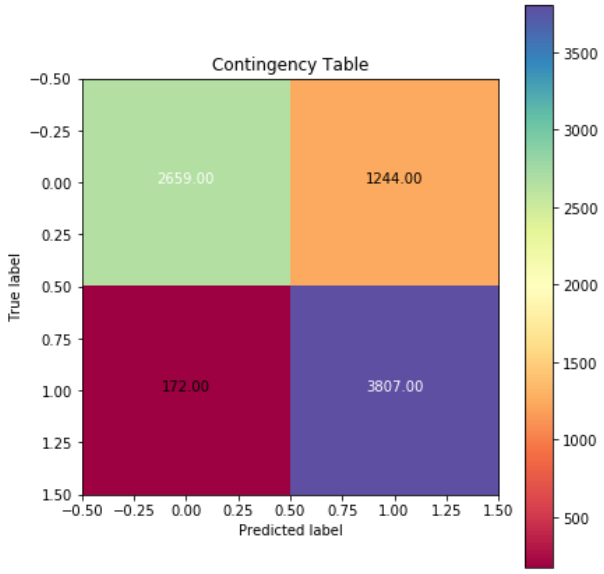
```
Model with 2 components:
_____________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_____________________________________________________________________________________
k-means++       0.22s   1       0.409   0.440   0.424   0.434   0.409
```

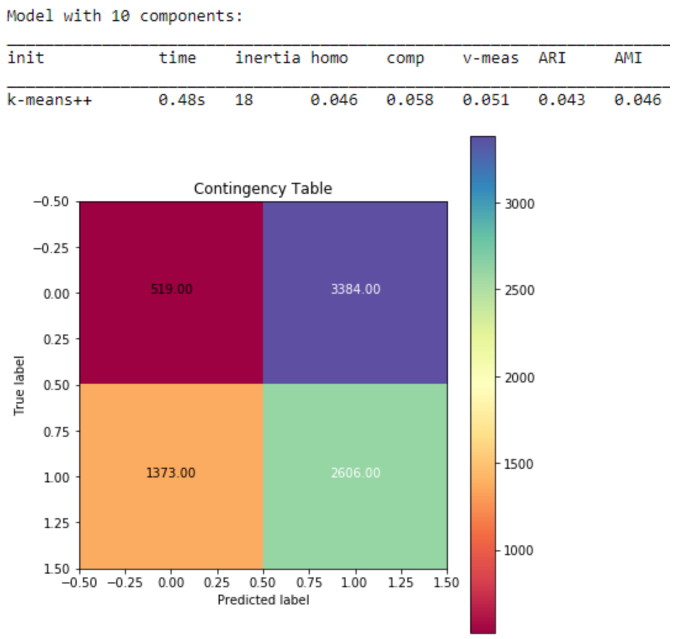


```
Model with 3 components:
__________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
__________________________________________________________________________________
k-means++       0.15s   3       0.384   0.417   0.400   0.406   0.384
```

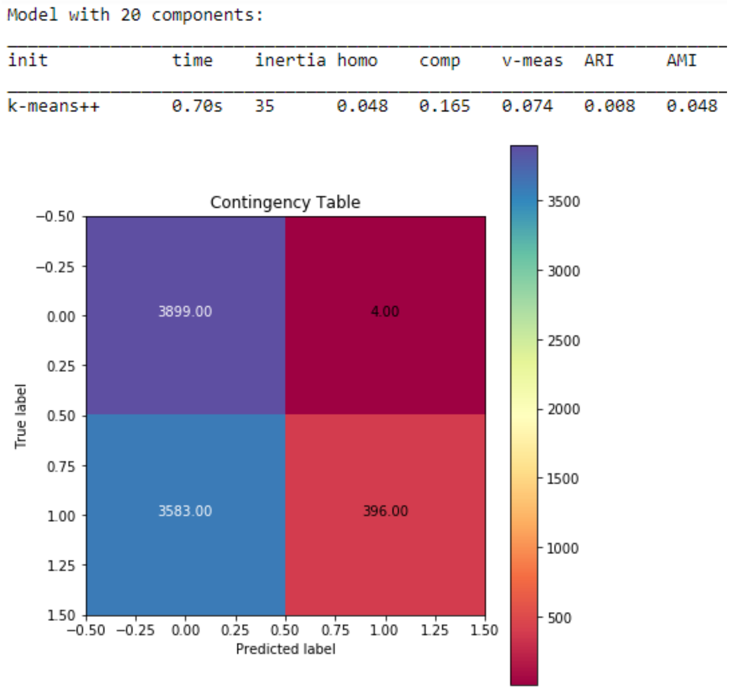


```
Model with 5 components:
__________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
__________________________________________________________________________________
k-means++       0.17s   8       0.365   0.388   0.376   0.410   0.365
```



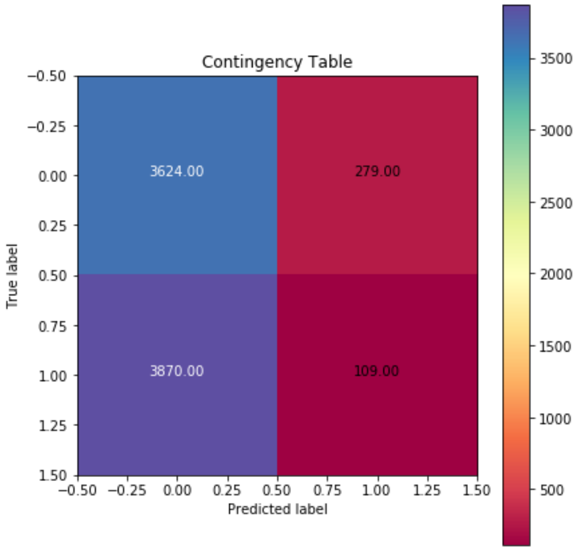


```
Model with 10 components:
________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
________________________________________________________________________________
k-means++       0.48s   18      0.046   0.058   0.051   0.043   0.046
```

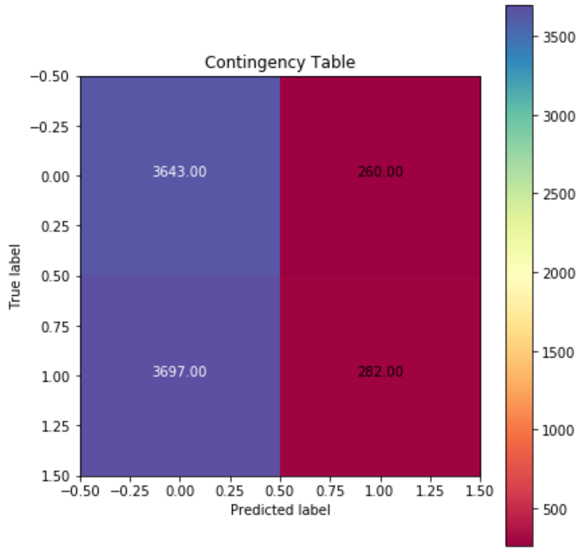


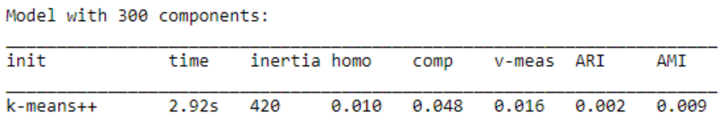
```
Model with 20 components:
________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
________________________________________________________________________________
k-means++       0.70s   35      0.048   0.165   0.074   0.008   0.048
```

```
Model with 50 components:
________________________________________________________________________________
init              time     inertia homo     comp     v-meas   ARI      AMI
________________________________________________________________________________
k-means++         0.81s    81        0.008    0.027    0.012    0.003    0.008
```

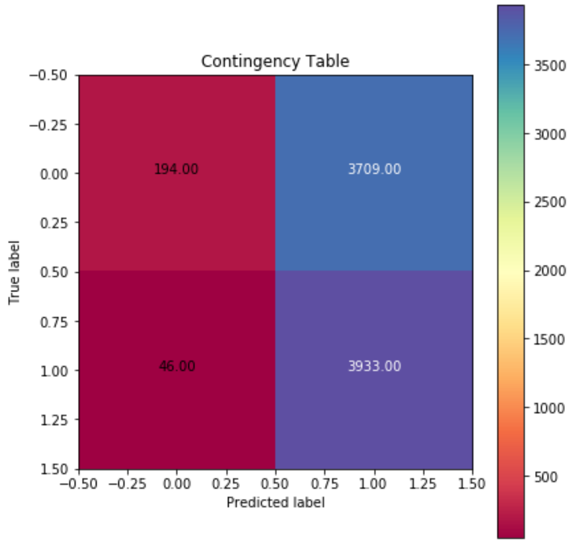


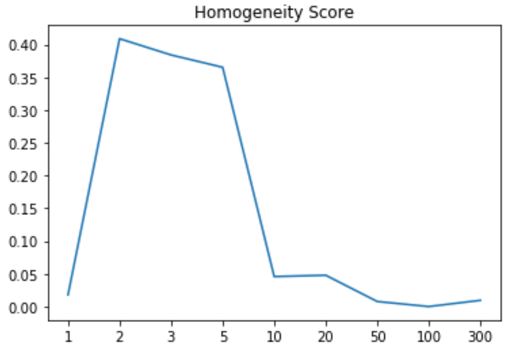
```
Model with 100 components:
________________________________________________________________________________
init              time     inertia homo     comp     v-meas   ARI      AMI
________________________________________________________________________________
k-means++         1.55s    159       0.000    0.000    0.000    -0.000   -0.000
```

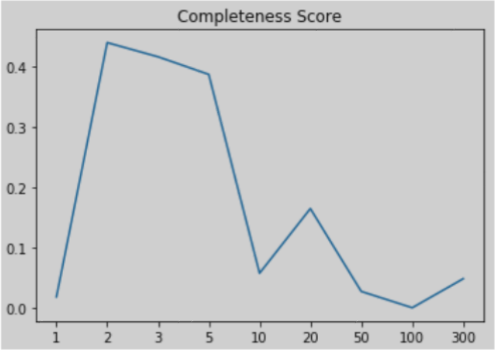


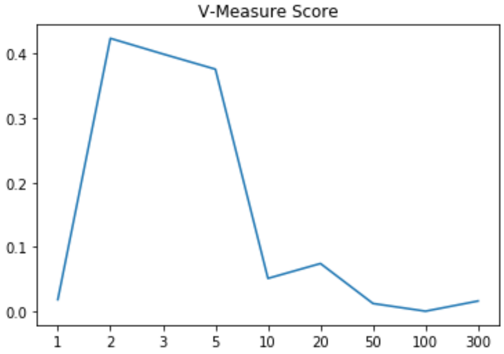


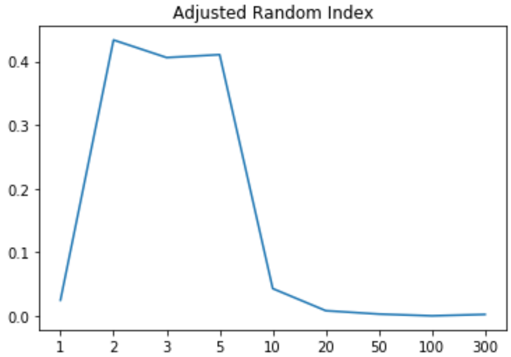
```
Model with 300 components:
____________________________________________________________________________
init            time     inertia homo     comp     v-meas   ARI      AMI
____________________________________________________________________________
k-means++       2.92s    420     0.010    0.048    0.016    0.002    0.009
```

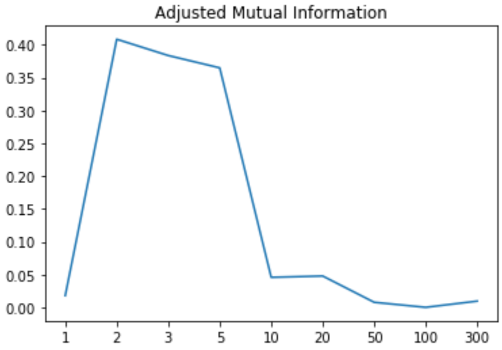












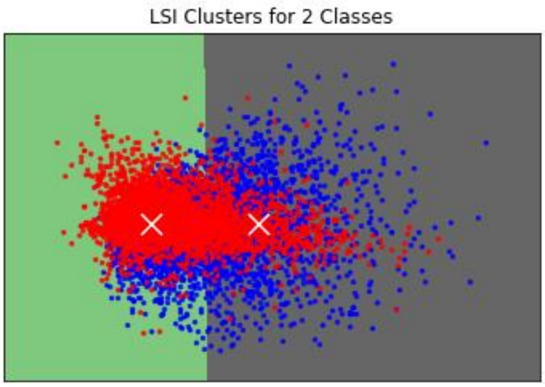
As we can observe, the best r value is at r = 2. i.e, only 2 components/features are more than enough to best cluster the documents in two classes with maximum accuracy.

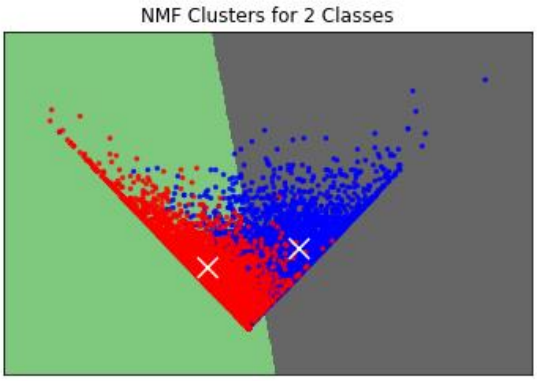
The non-monotonic behavior can be explained by the following reasoning.
The Curse of Dimensionality:
The distance to the nearest neighbor and the distance to the farthest neighbor tend to converge as the dimension increases. On the other hand, increasing the dimension of data representation provides intricate details which helps in distinguishing each data point. Hence, a proper balance needs to be found to represent data in best dimension space. Here, we observed that the clustering improved performance as we increase dimension but started performing worse for higher dimensions. Hence, it exhibits the non-monotonic behavior.

TASK 4:

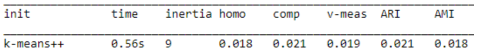
Visualize the performance:

a. We found that,
   r=5 works best in case of LSI, and
   r=2 works best in case of NMF.
   Hence, following are the clusters with their decision boundary for the two methods.

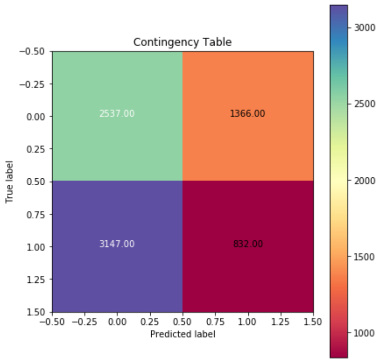


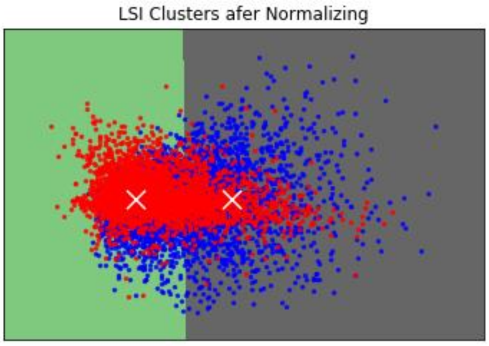


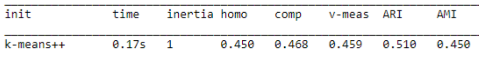
b. 3 Methods:
   i. Normalizing:
      We performed scaling using sklearn “scale” to get unit variance and then normalized the data. Following are the results for LSI and NMF:



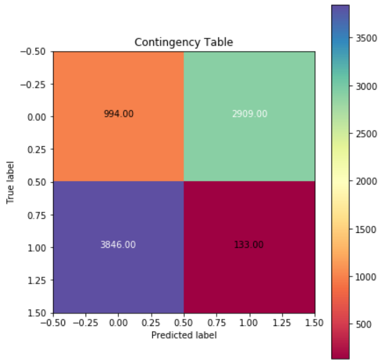
```
init            time    inertia homo    comp    v-meas  ARI     AMI

k-means++       0.56s   9       0.018   0.021   0.019   0.021   0.018
```



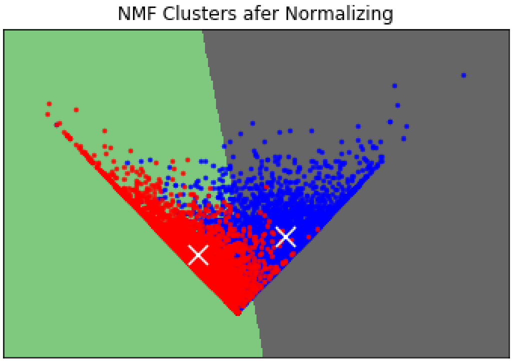




```
__________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
__________________________________________________________________________________
k-means++       0.17s   1       0.450   0.468   0.459   0.510   0.450
```



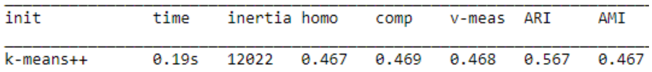
Contingency Table



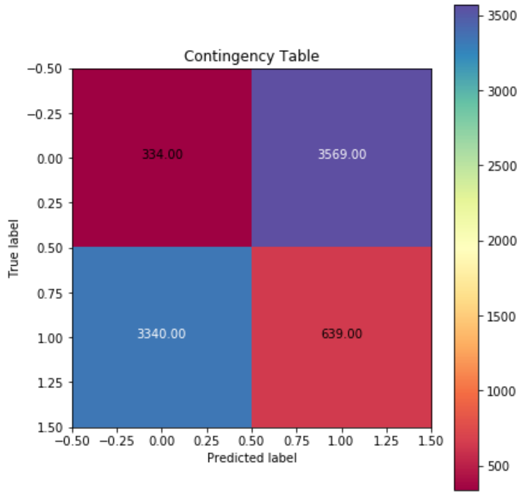
NMF Clusters afer Normalizing

As we can observe, the performance has improved from 83.69% to 85.7% which is a good improvement.

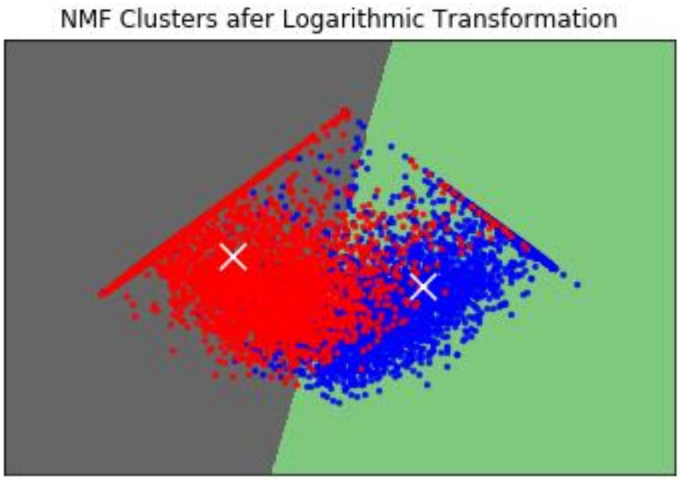
ii. Logarithmic Transformation:



```
___________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
___________________________________________________________________________________
k-means++       0.19s   12022   0.467   0.469   0.468   0.567   0.467
```



Contingency Table



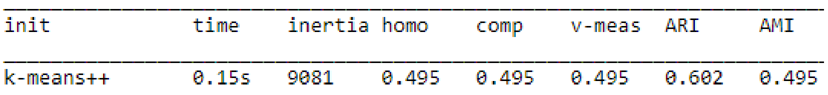
NMF Clusters afer Logarithmic Transformation

As we can observe, the performance has improved from 83.69% to 87.65% which is a very good improvement.

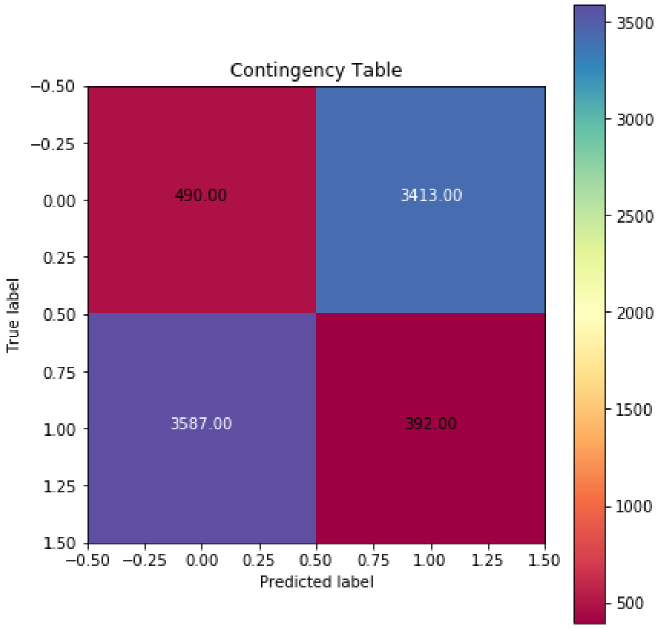
iii. Combination

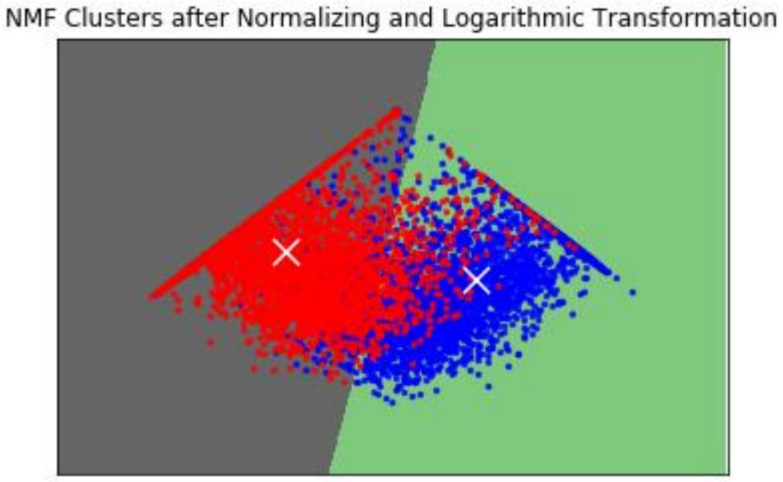
1. Normalization and then Log Transformation

We perform normalization and then apply log transformation on NMF data to get the following results:



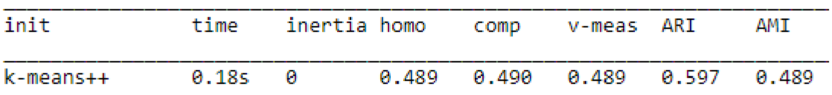
```
________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
________________________________________________________________________________
k-means++       0.15s   9081    0.495   0.495   0.495   0.602   0.495
```



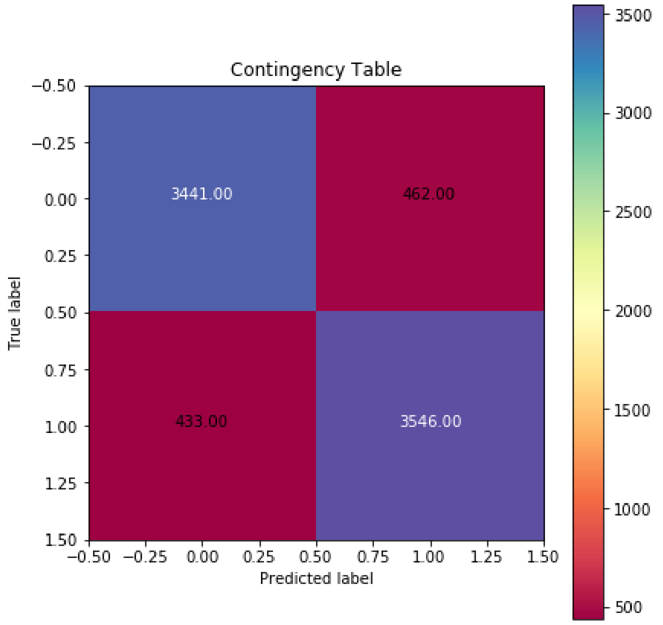


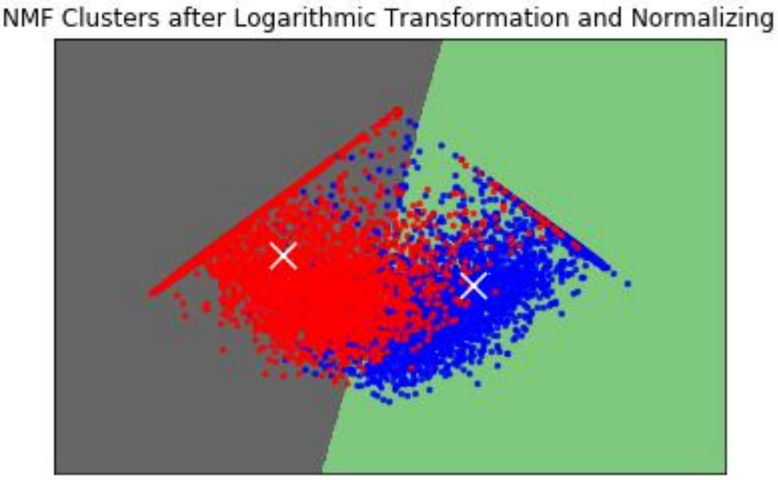
As we can observe, the performance has improved from 83.69% to 88.80% which is higher than normalization or log transformation separately. Infact, it is the highest and best combination so far.

2. Log Transformation and then Normalization



```
______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________________
k-means++       0.18s   0       0.489   0.490   0.489   0.597   0.489
```





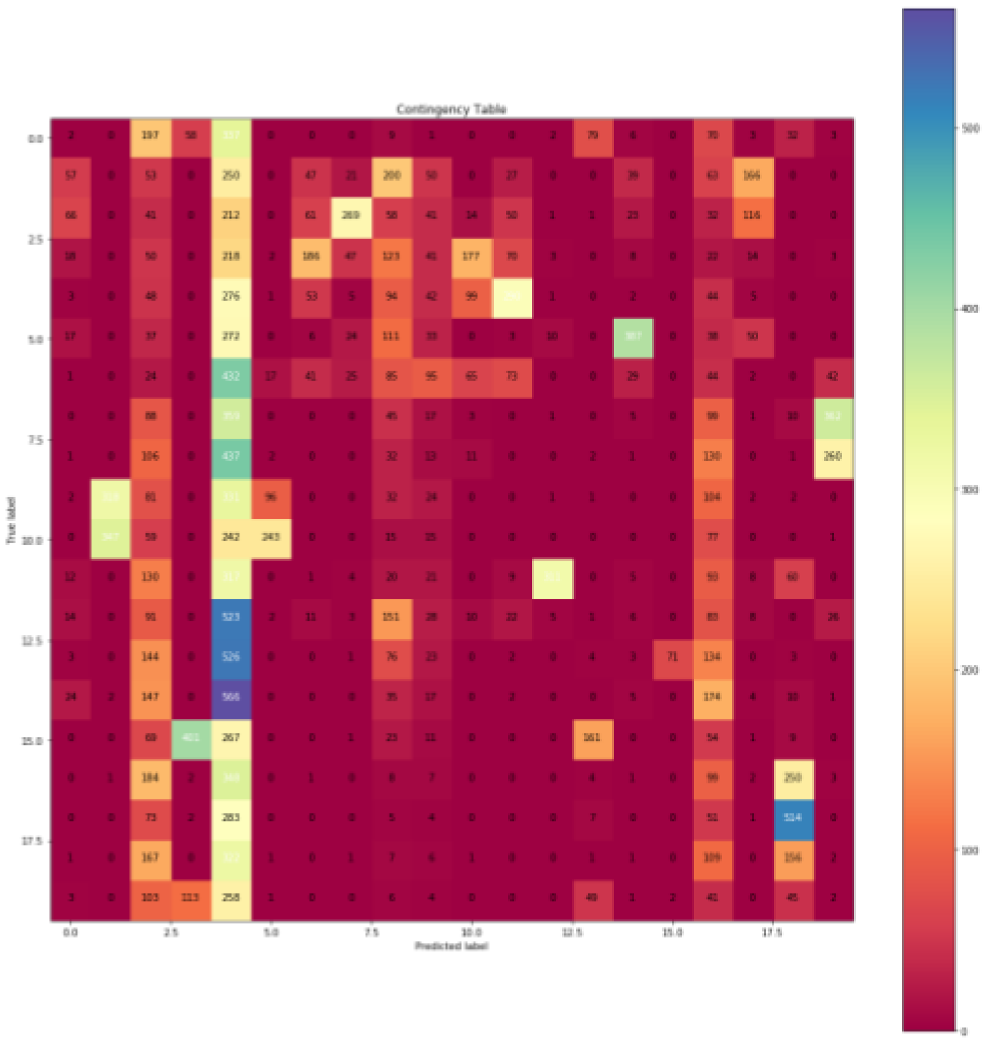
As we can observe, the performance has improved from 83.69% to 88.64% which is higher than normalization or log transformation separately.

TASK 5: Expand to 20 categories

We follow the same workflow as above with all 20 categories and produce following results and observations:

1. We perform TF-IDF on the documents by performing certain preprocessing steps such as stemming, word tokenizing, special characters removal and stopwords removal.
   With min_df = 3, the dimension of the TF-IDF matrix is (18846, 33158), where 18846 are the total number of documents in 20 categories and 33158 are the unique tokens identified.
2. After applying k-means clustering on the above dataset, we got the following results.
   With number of clusters = 2:

```
________________________________________________________________________________
init             time     inertia homo     comp     v-meas   ARI      AMI

________________________________________________________________________________
k-means++        570.00s  17686   0.246    0.306    0.273    0.050    0.243
```



3. Preprocess the data:
   Dimensionality reduction:
   a. LSI (Latent Semantic Indexing):
   After performing LSI on the TF-IDF matrix, following are the obtained results:

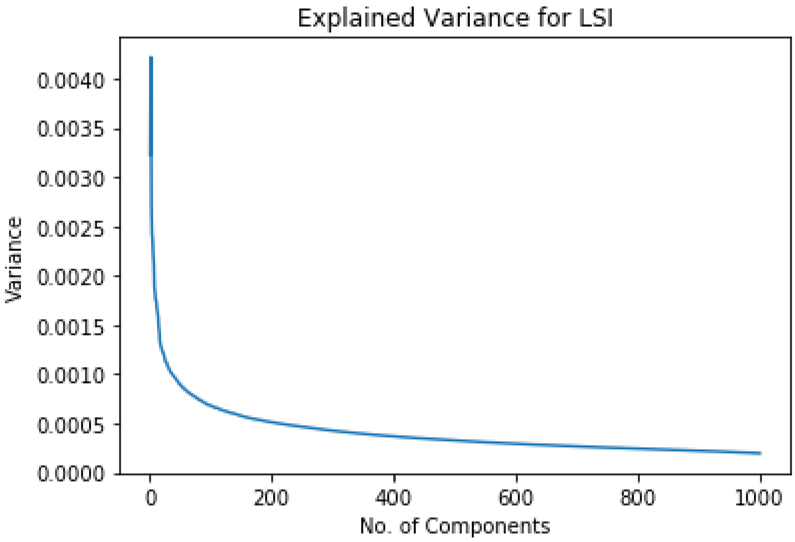
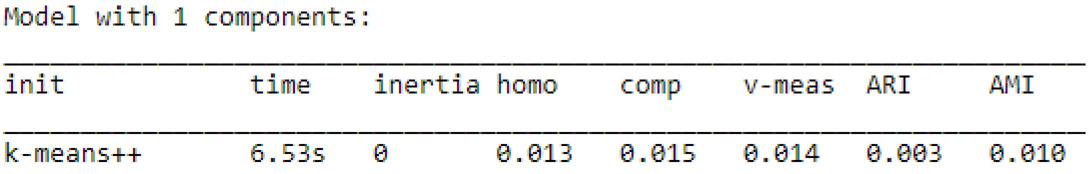
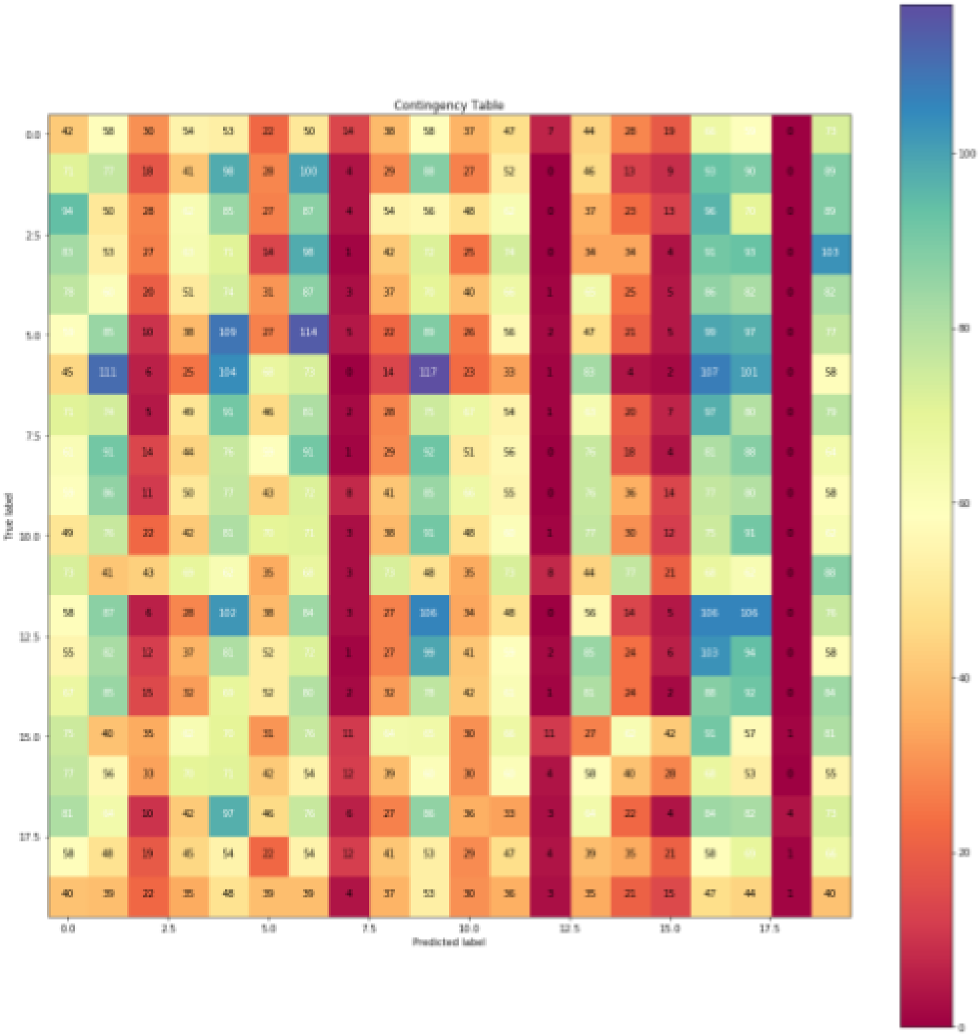


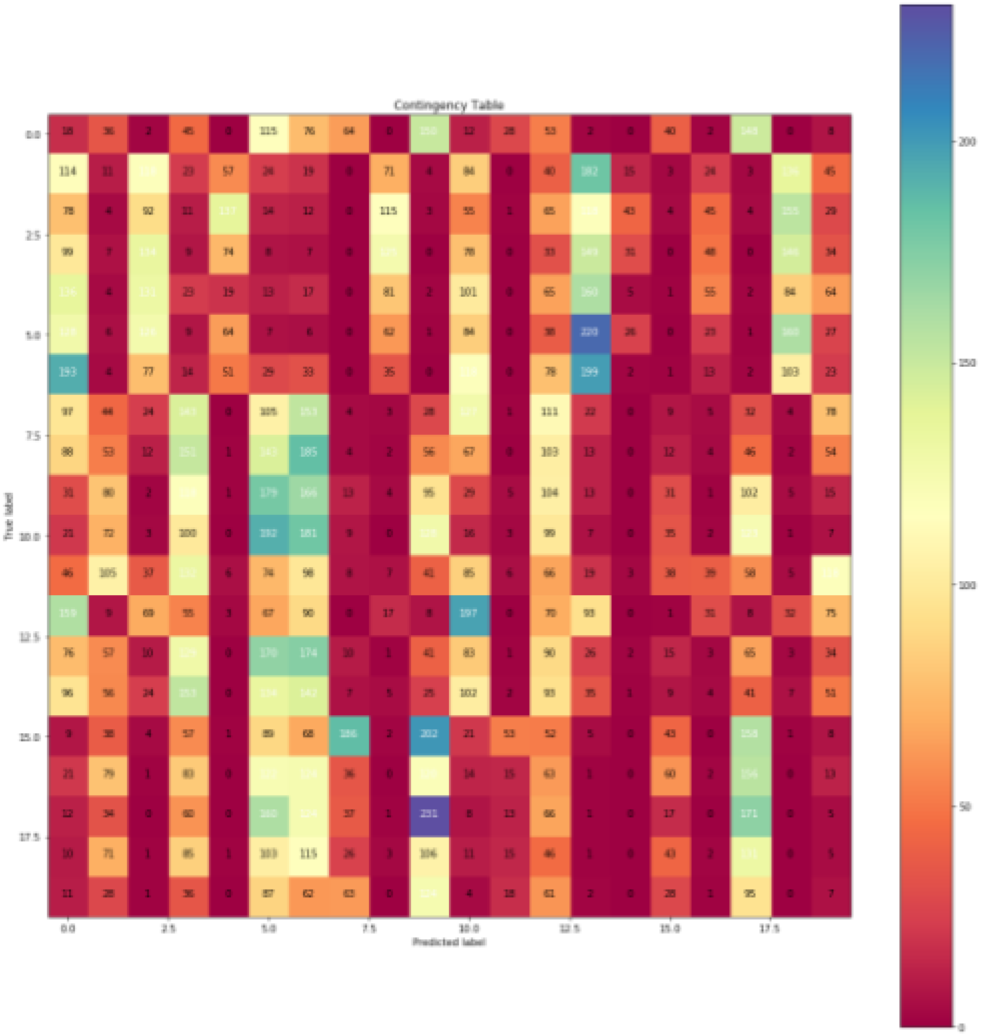
Figure 2: We can observe that the variance decreases as the no. of components increase. Interestingly, the variance is even more lesser than previous 8 categories.



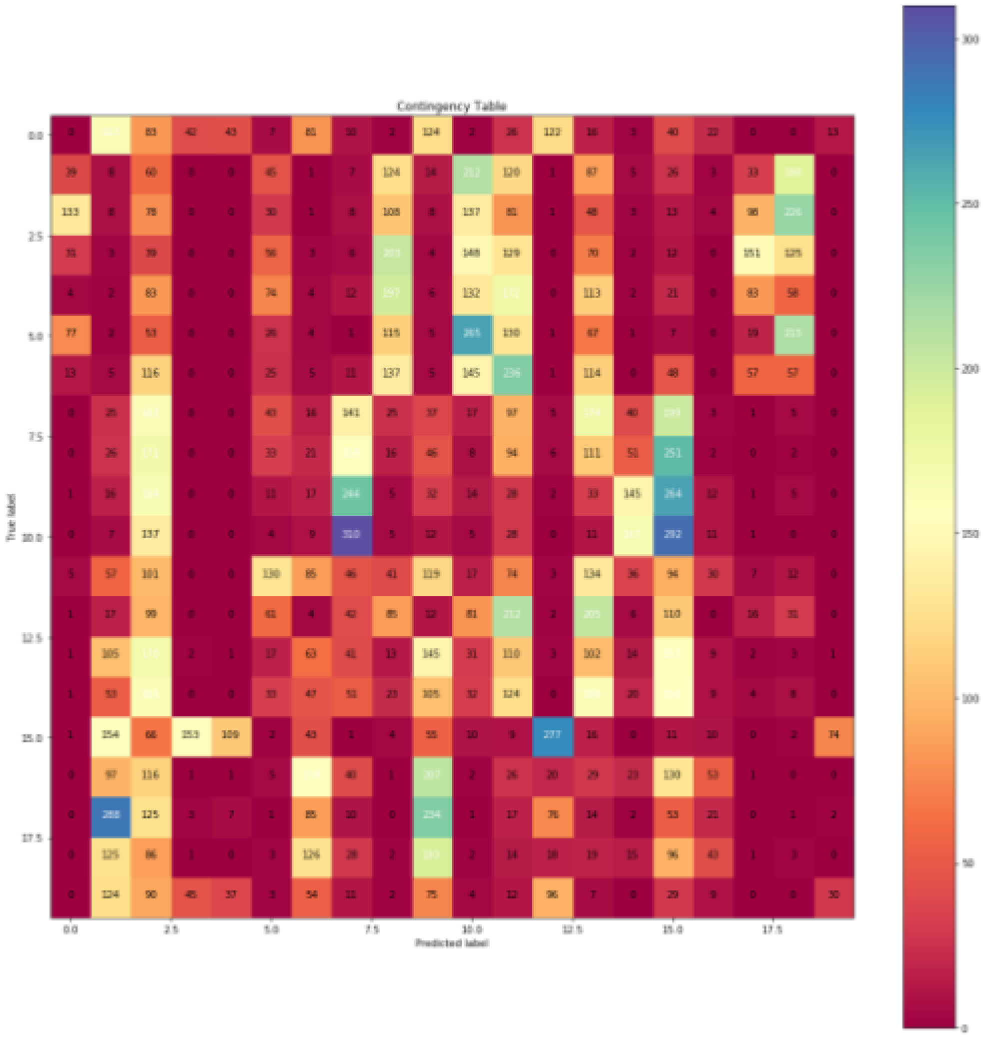
```
Model with 1 components:
________________________________________________________________________
init            time     inertia homo     comp     v-meas   ARI      AMI
________________________________________________________________________
k-means++       6.53s    0        0.013    0.015    0.014    0.003    0.010
```



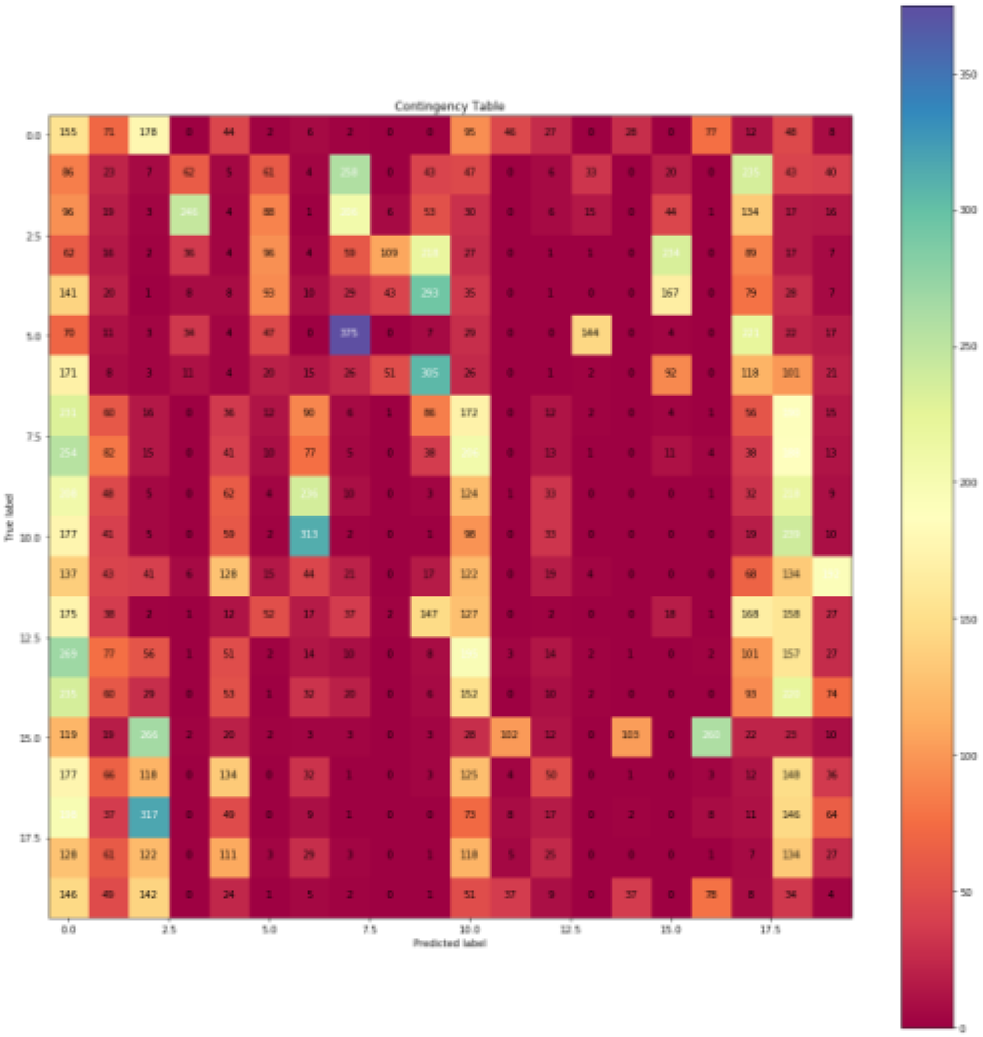
```
Model with 2 components:
_______________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_______________________________________________________________________
k-means++       10.54s  9       0.152   0.162   0.157   0.043   0.150
```



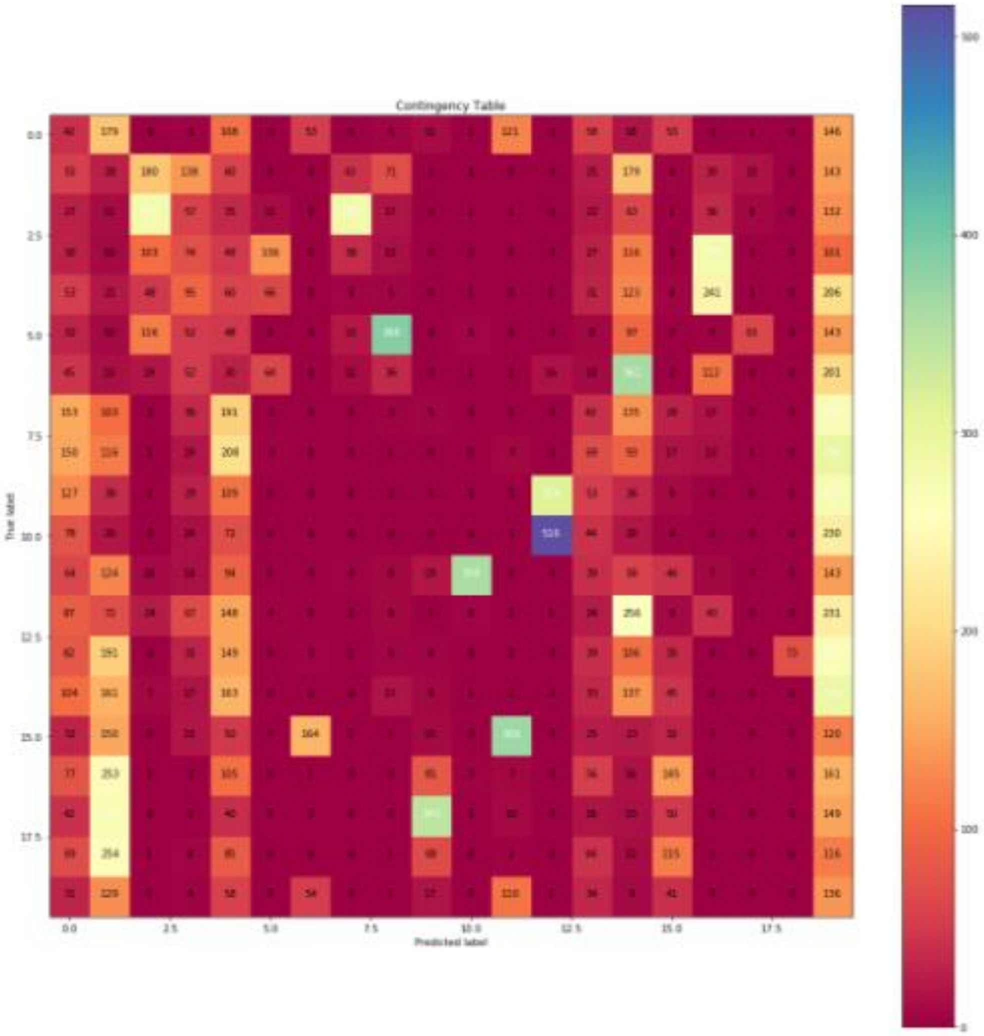
```
Model with 3 components:
______________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________
k-means++       10.64s  23      0.201   0.217   0.209   0.064   0.199
```



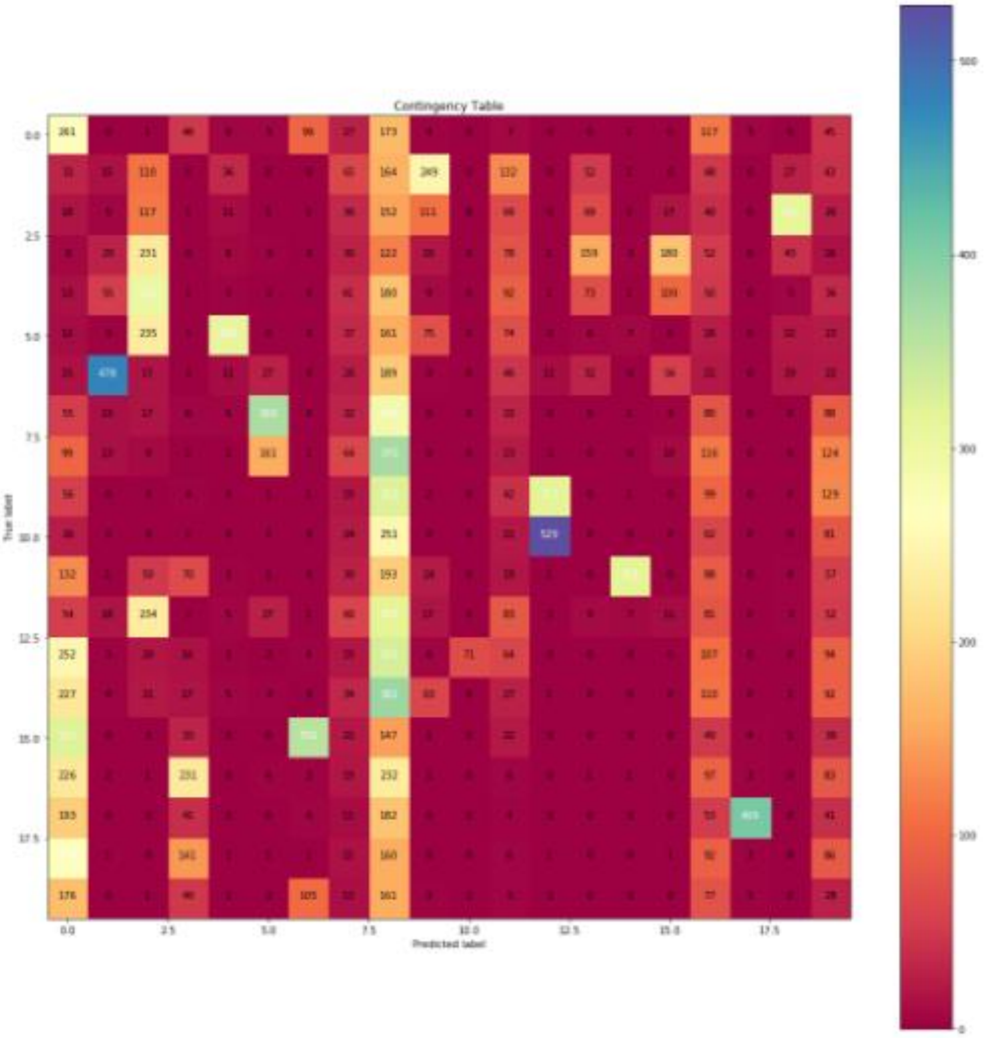
```
Model with 5 components:
_____________________________________________________________________________
init            time     inertia homo     comp     v-meas   ARI      AMI
_____________________________________________________________________________
k-means++       12.64s   61       0.199    0.221    0.209    0.061    0.197
```



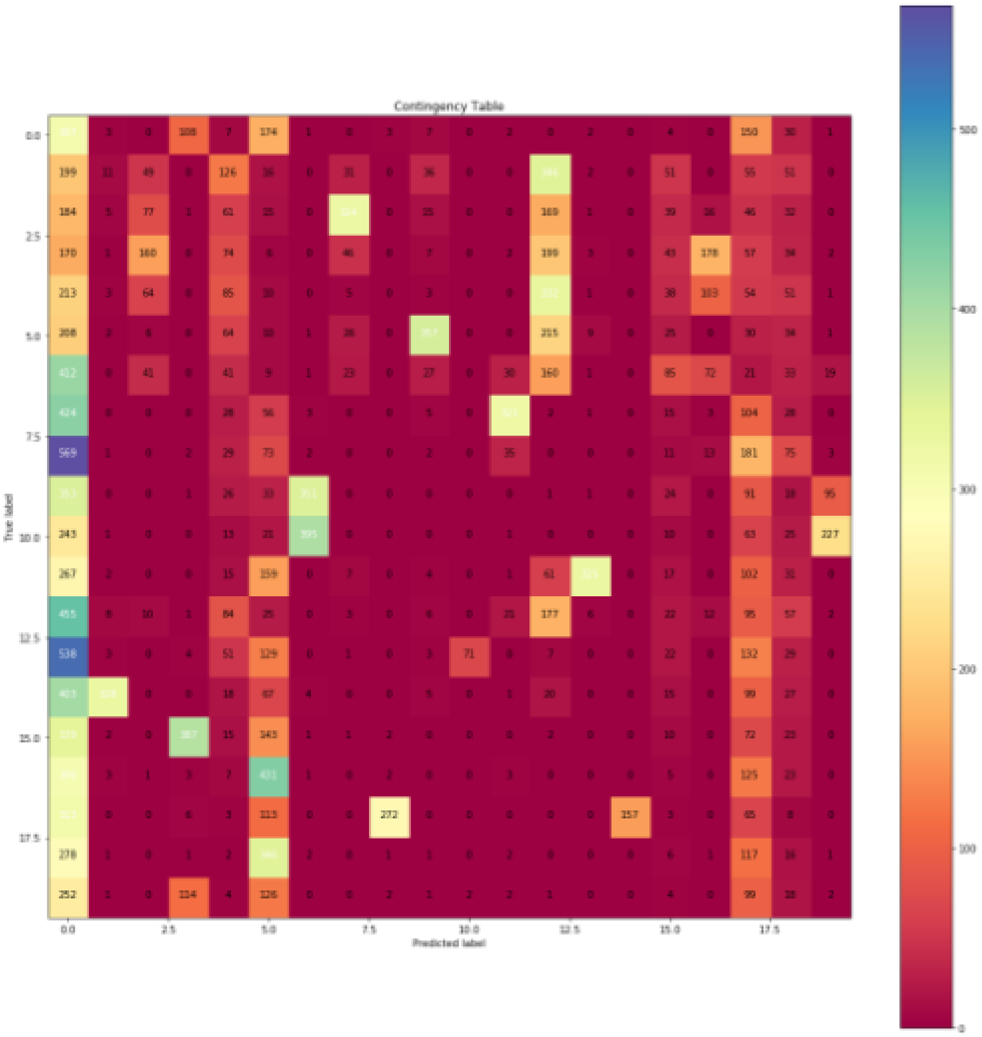
```
Model with 10 components:
________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
________________________________________________________________________________
k-means++       12.95s  148     0.230   0.259   0.244   0.072   0.227
```



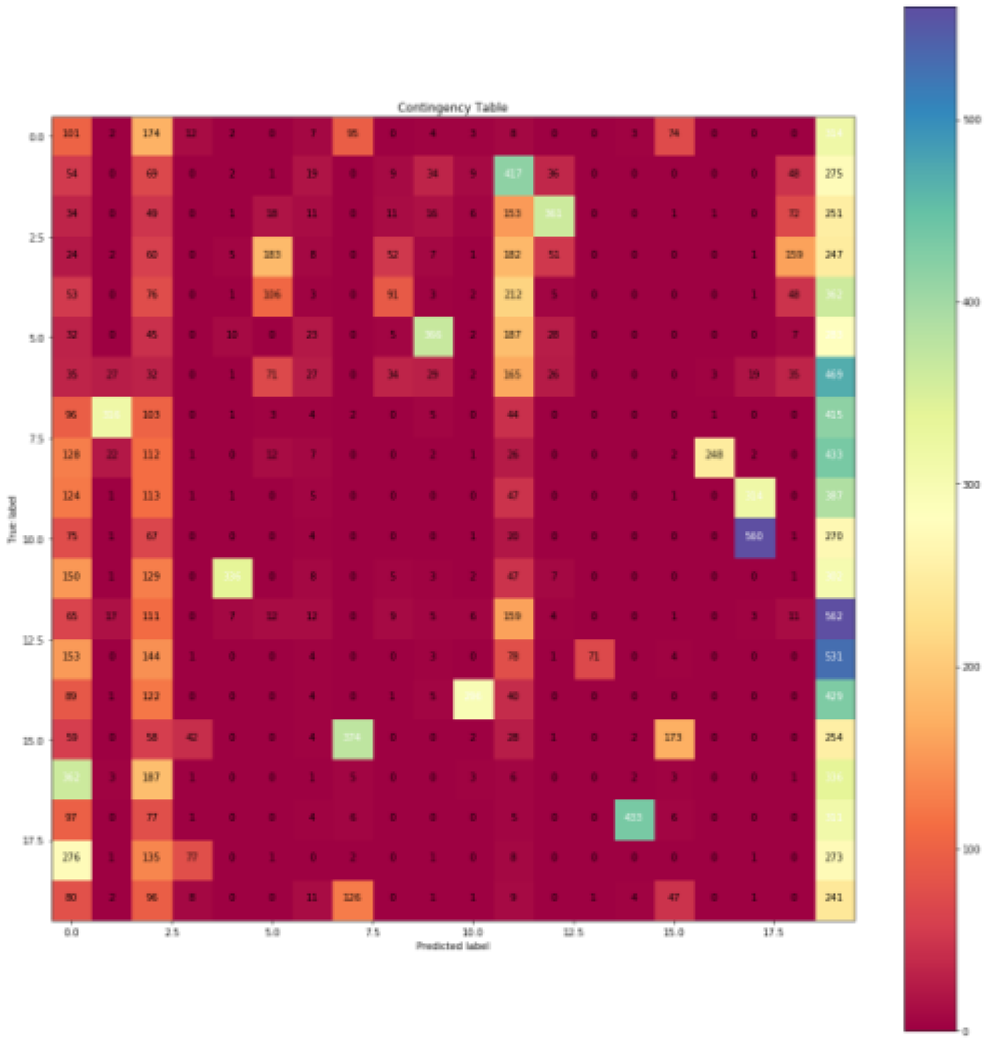
```
Model with 20 components:
___________________________________________________________________________________
init            time     inertia homo     comp     v-meas   ARI      AMI
___________________________________________________________________________________
k-means++       11.56s   334      0.264    0.299    0.280    0.081    0.261
```



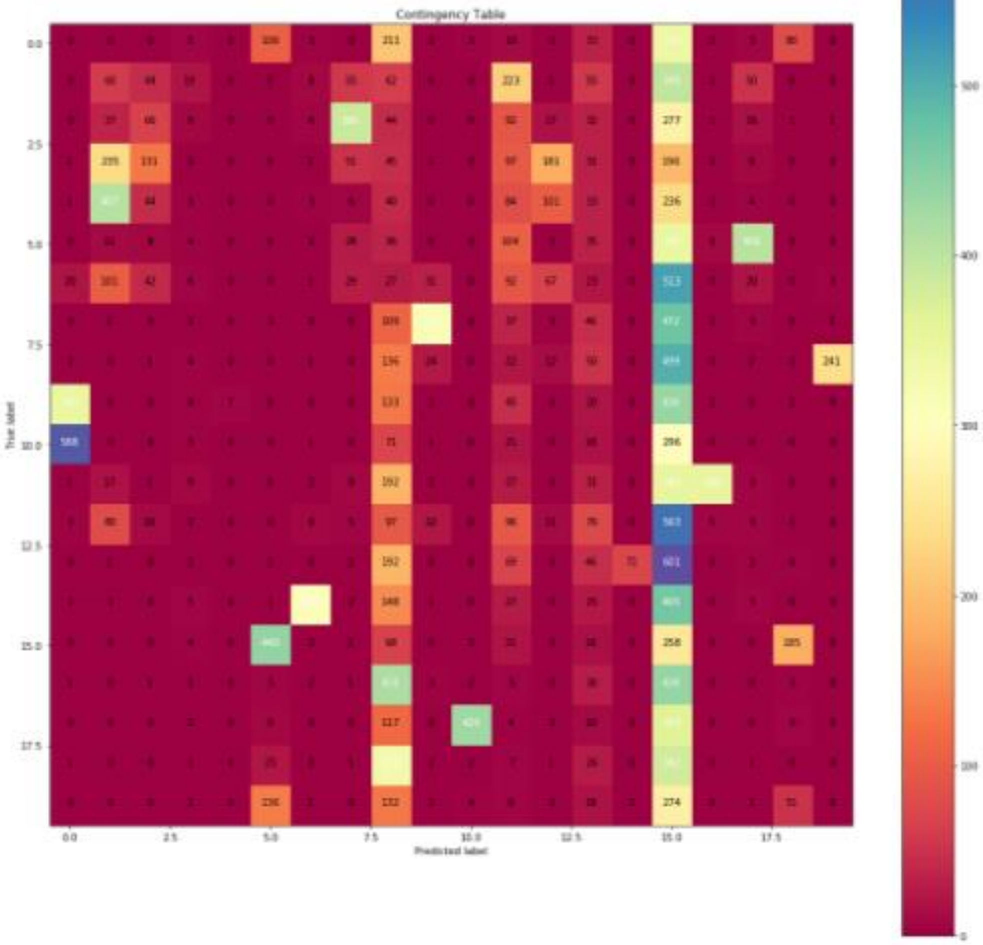
```
Model with 50 components:

________________________________________________________________
init             time     inertia homo     comp     v-meas   ARI      AMI
________________________________________________________________
k-means++        15.05s   860      0.253    0.315    0.281    0.057    0.250
```

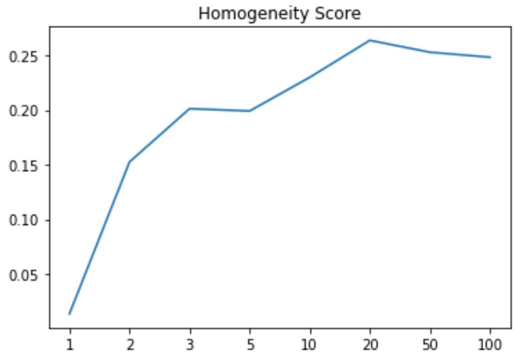


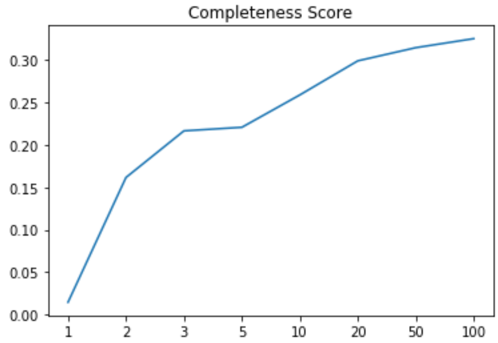
```
Model with 100 components:
_______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_______________________________________________________________________________
k-means++       21.82s  1560    0.248   0.326   0.282   0.052   0.246
```

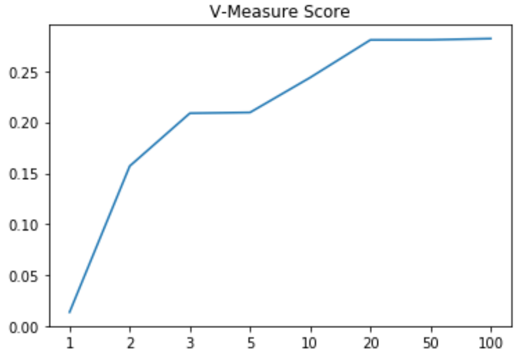


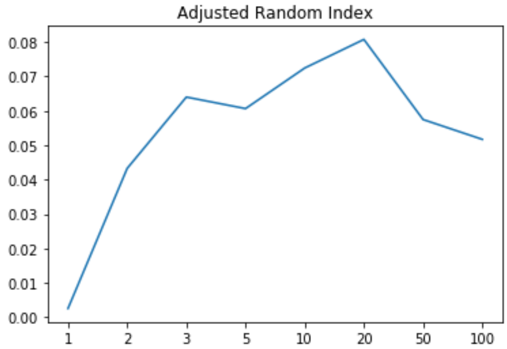
```
Model with 300 components:
______________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________
k-means++       42.33s  3461    0.261   0.354   0.300   0.053   0.259
```

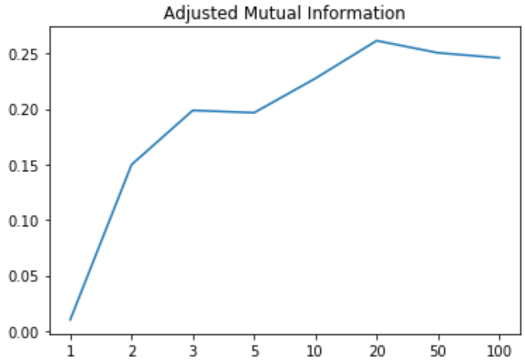

Contingency Table
True label
Predicted label
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
100
200
300
400
500
600





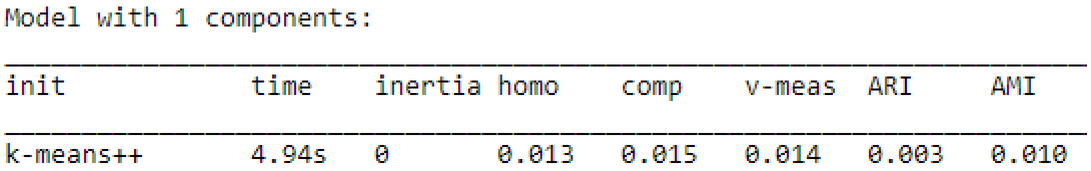




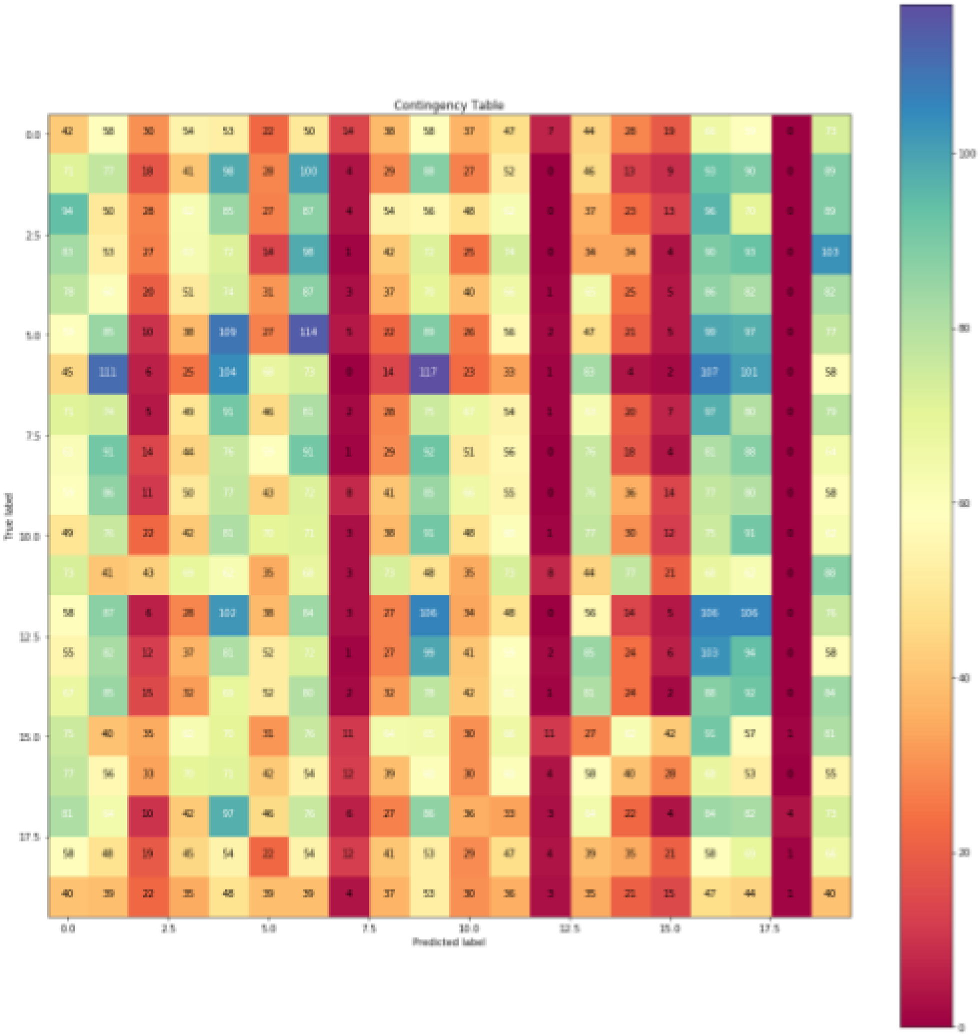


As we can observe from the contingency matrix as well as the 5 measures above, the best r value for LSI is 20. i.e, 20 components are best suitable to represent each document and cluster with maximum accuracy. Although, the graph for completeness score and V-Measure are somewhat monotonic in nature, we consider all the measures and decide 20 to be the best value.

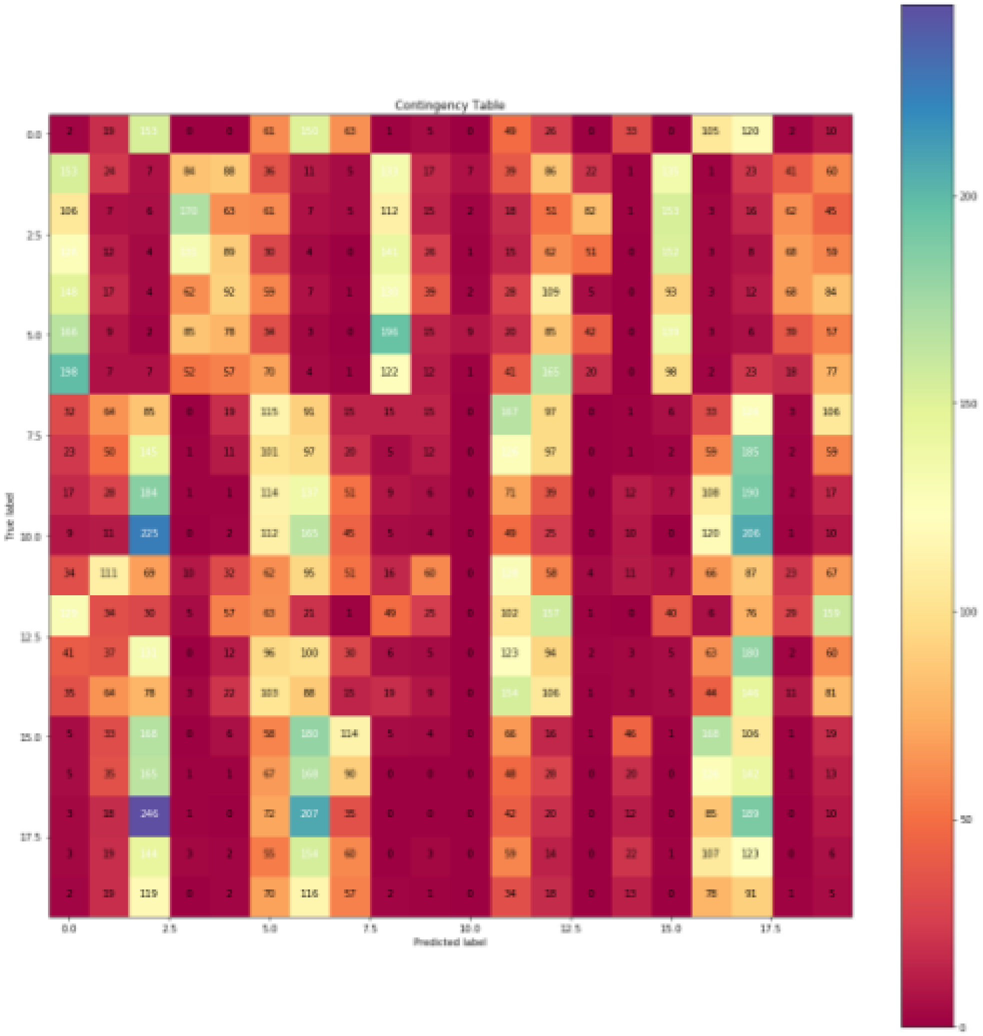
b. NMF (Non-Negative Matrix Factorization):
After performing NMF on the TF-IDF matrix, following are the obtained results:



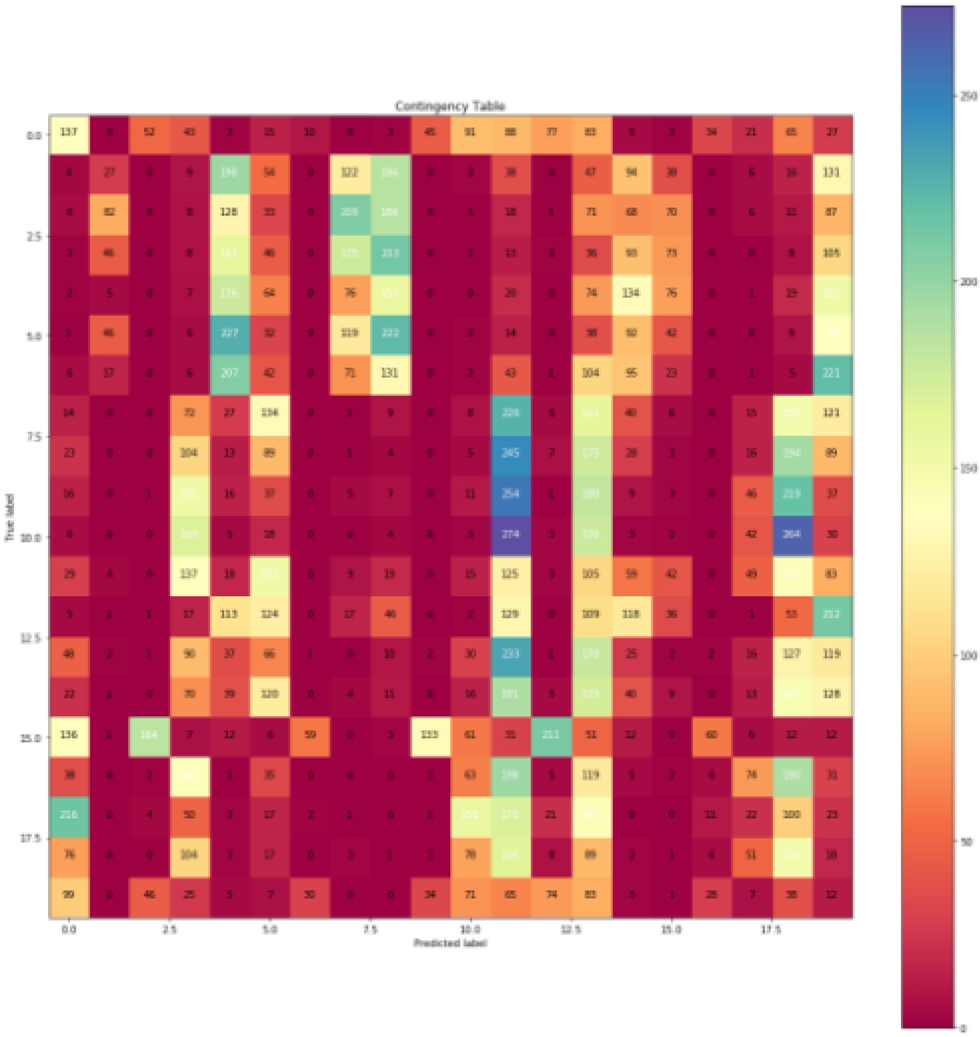
```
Model with 1 components:
________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
________________________________________________________________________________
k-means++       4.94s   0       0.013   0.015   0.014   0.003   0.010
```



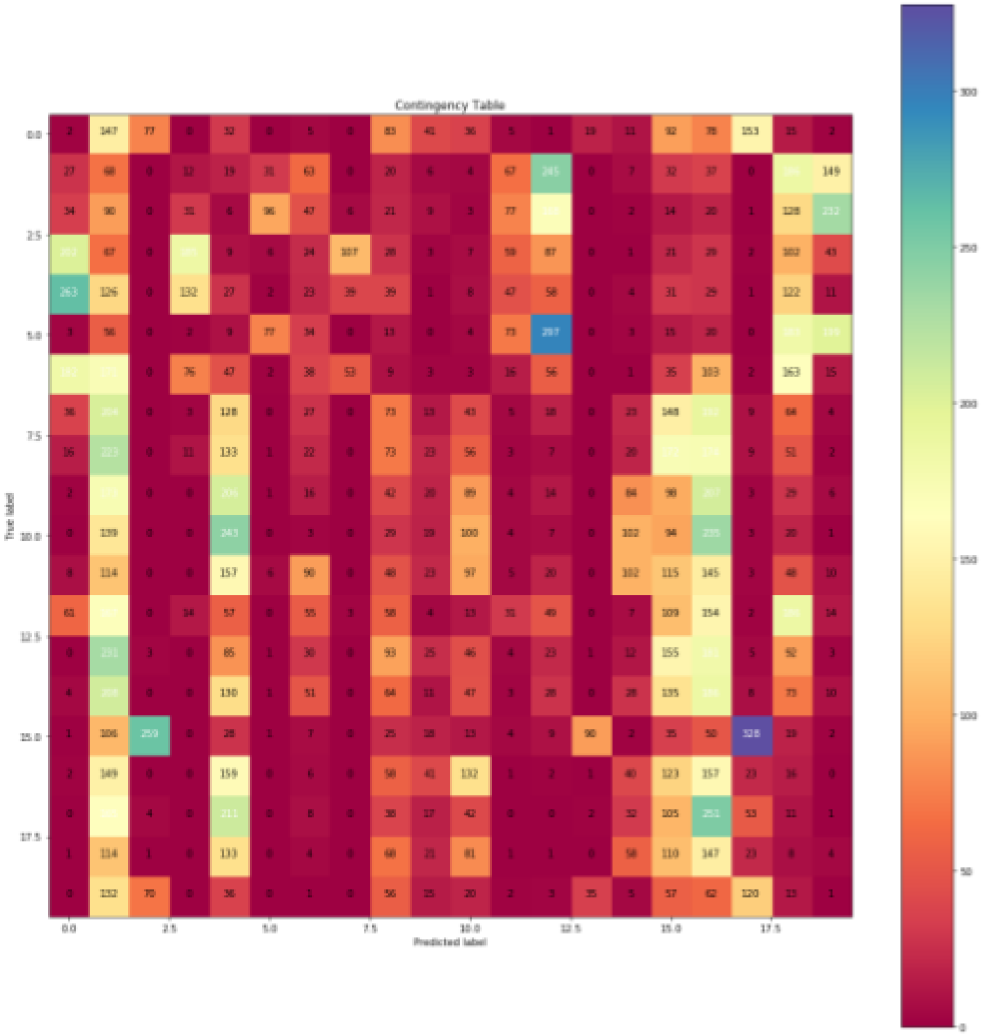
```
Model with 2 components:
_____________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_____________________________________________________________________
k-means++       10.66s  0       0.143   0.154   0.149   0.041   0.141
```



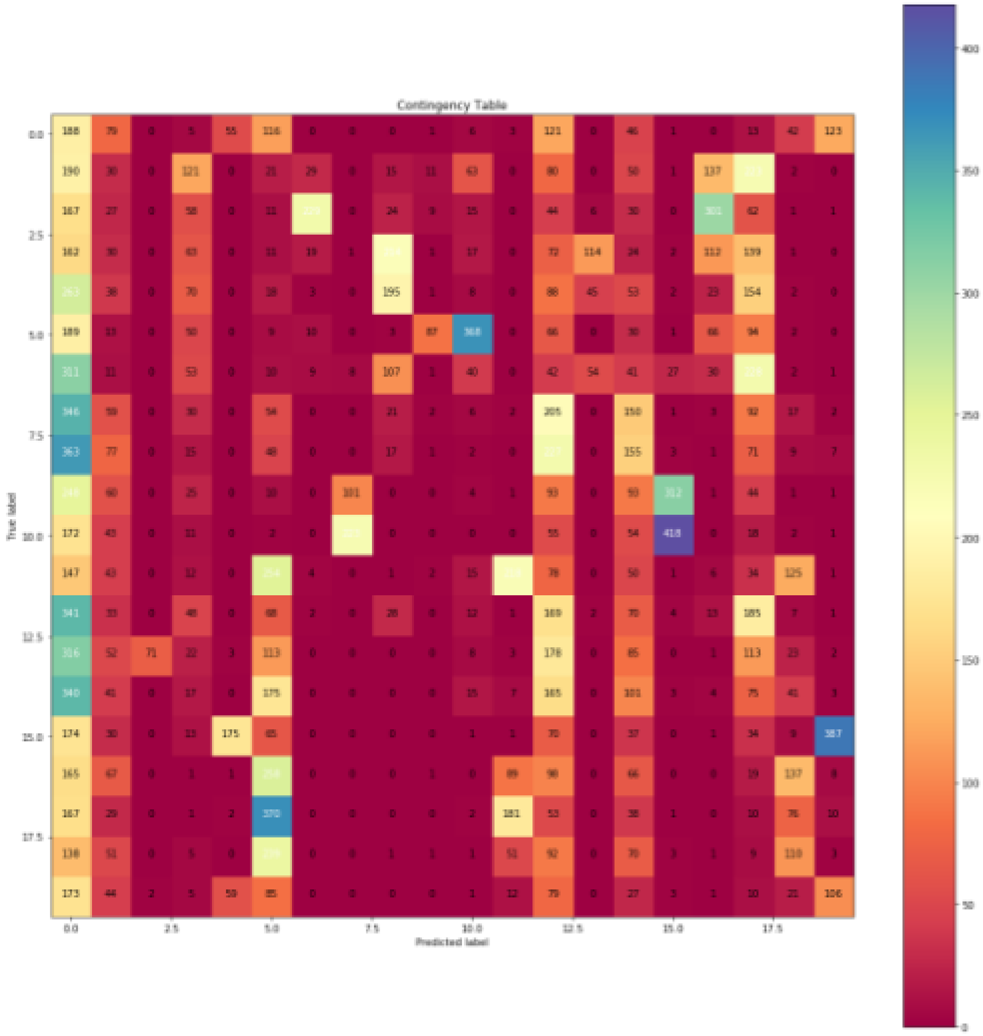
```
Model with 3 components:
_______________________________________________________________________________
init           time    inertia homo    comp    v-meas  ARI     AMI
_______________________________________________________________________________
k-means++      8.95s   0       0.181   0.199   0.190   0.052   0.178
```



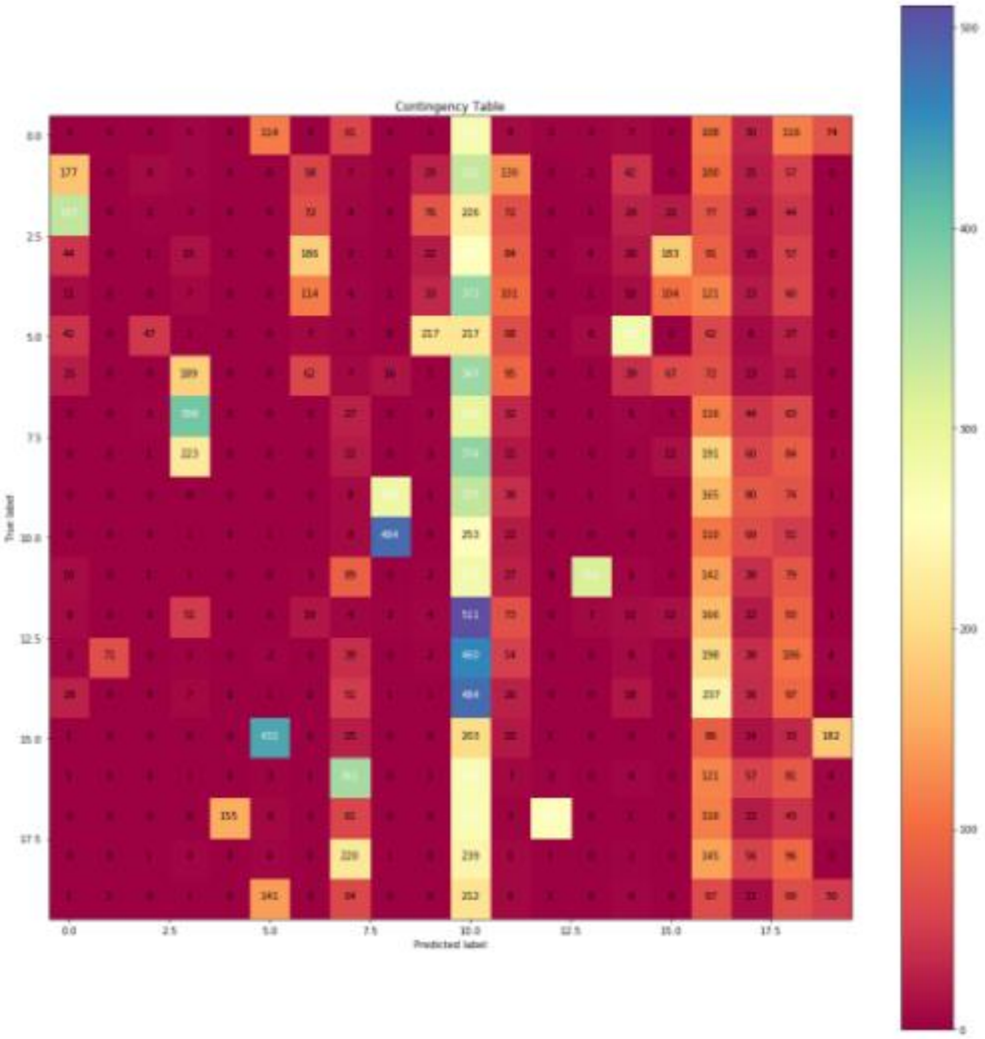
```
Model with 5 components:
________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
________________________________________________________________________________
k-means++       9.11s   1       0.174   0.191   0.182   0.049   0.171
```



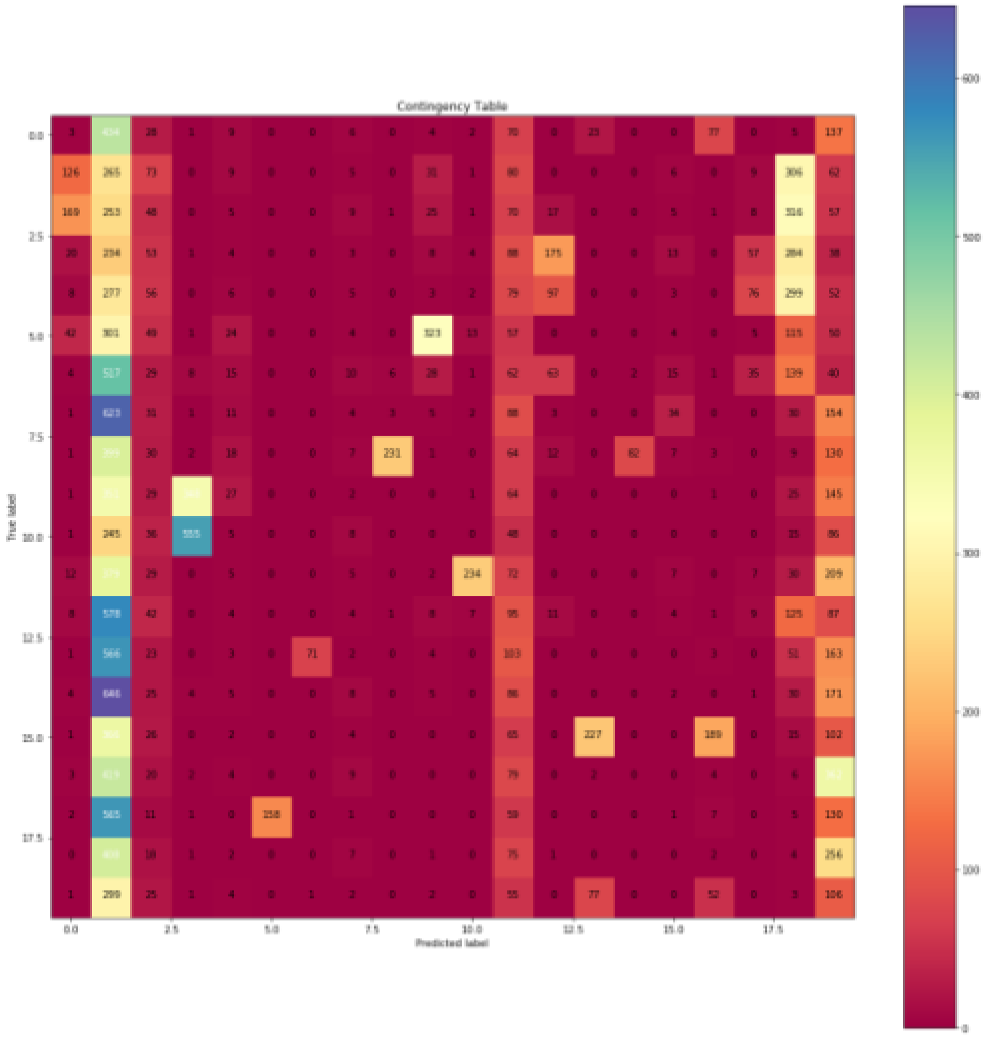
```
Model with 10 components:
_______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_______________________________________________________________________________
k-means++       6.06s   6       0.214   0.248   0.230   0.057   0.212
```



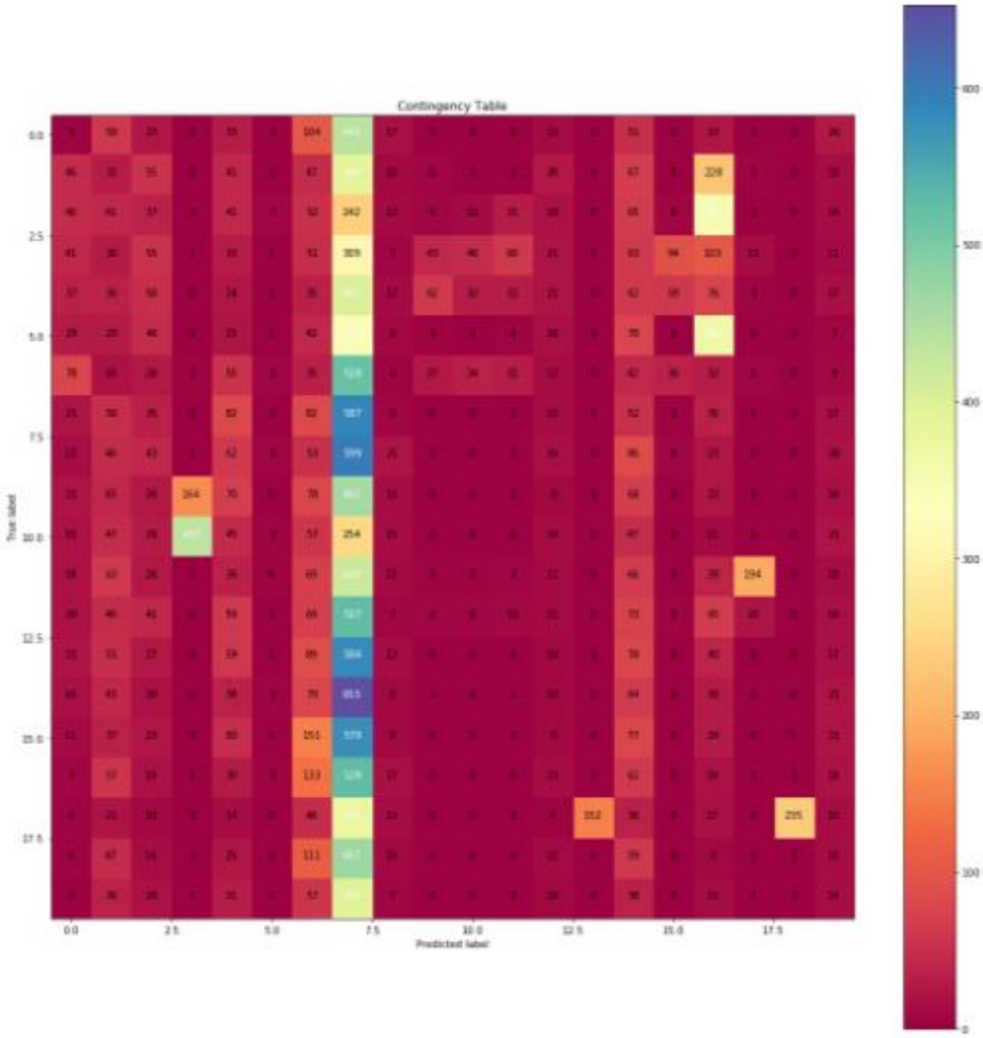
```
Model with 20 components:
__________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
__________________________________________________________________________________
k-means++       7.28s   19      0.227   0.281   0.251   0.049   0.225
```

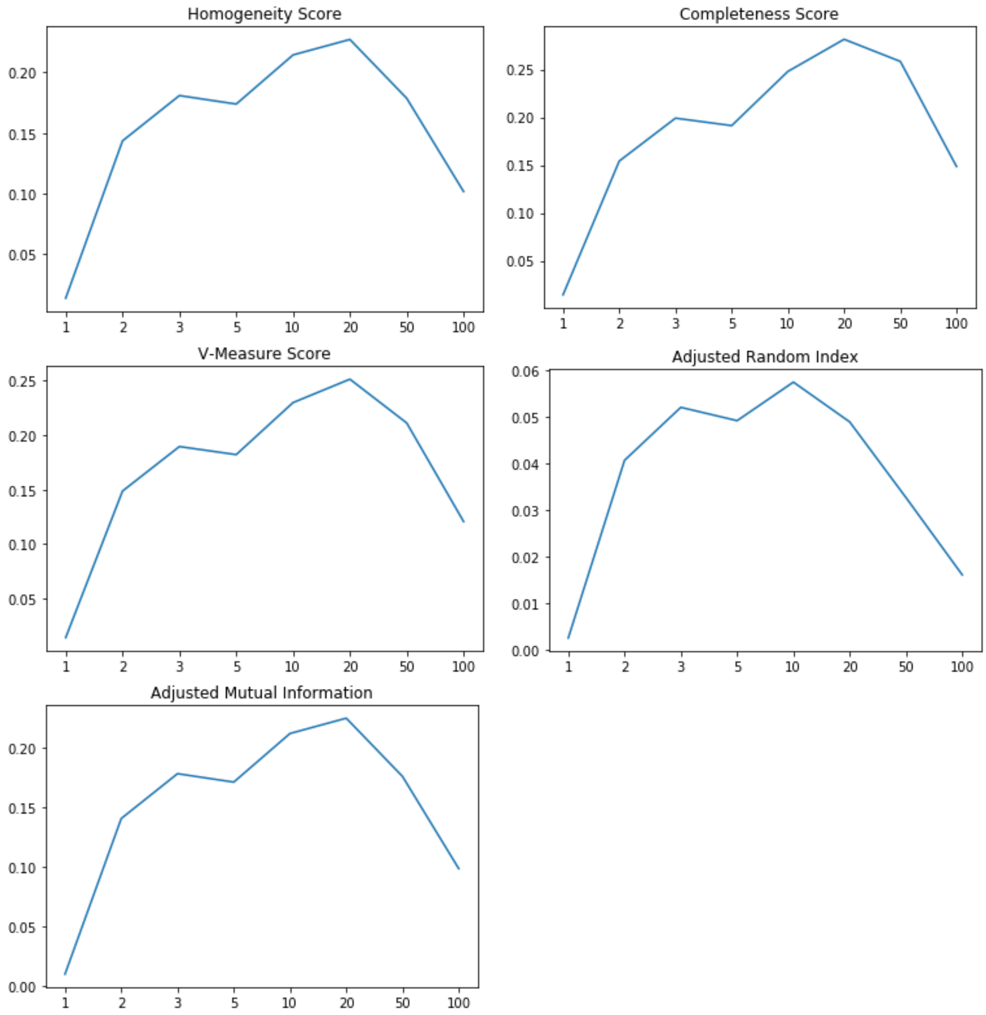


```
Model with 50 components:
__________________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
__________________________________________________________________________________
k-means++       11.42s  70      0.179   0.258   0.211   0.033   0.176
```



```
Model with 100 components:
________________________________________________________________________________
init              time    inertia homo     comp     v-meas   ARI      AMI
________________________________________________________________________________
k-means++         15.94s  158      0.102    0.149    0.121    0.016    0.099
```


Contingency Table
True label
Predicted label
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
100
200
300
400
500
600



Homogeneity Score

Completeness Score

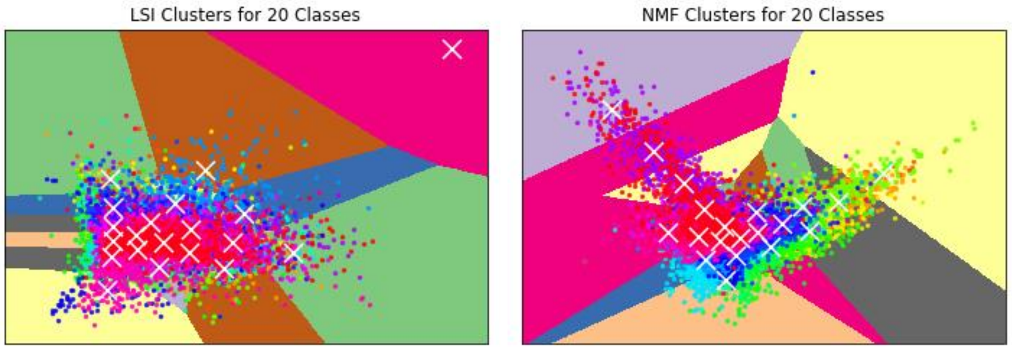
V-Measure Score

Adjusted Random Index

Adjusted Mutual Information

As we can observe, the best r value is at r = 20. i.e, only 20 components/features are more than enough to best cluster the documents in two classes with maximum accuracy.

4. Visualize the performance:
    a. We found that,
    r=5 works best in case of LSI, and
    r=2 works best in case of NMF.
    Hence, following are the clusters with their decision boundary for the two methods:

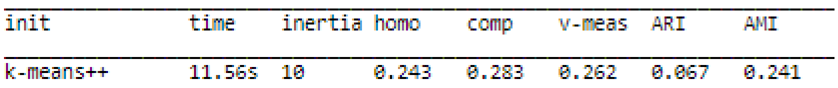


LSI Clusters for 20 Classes

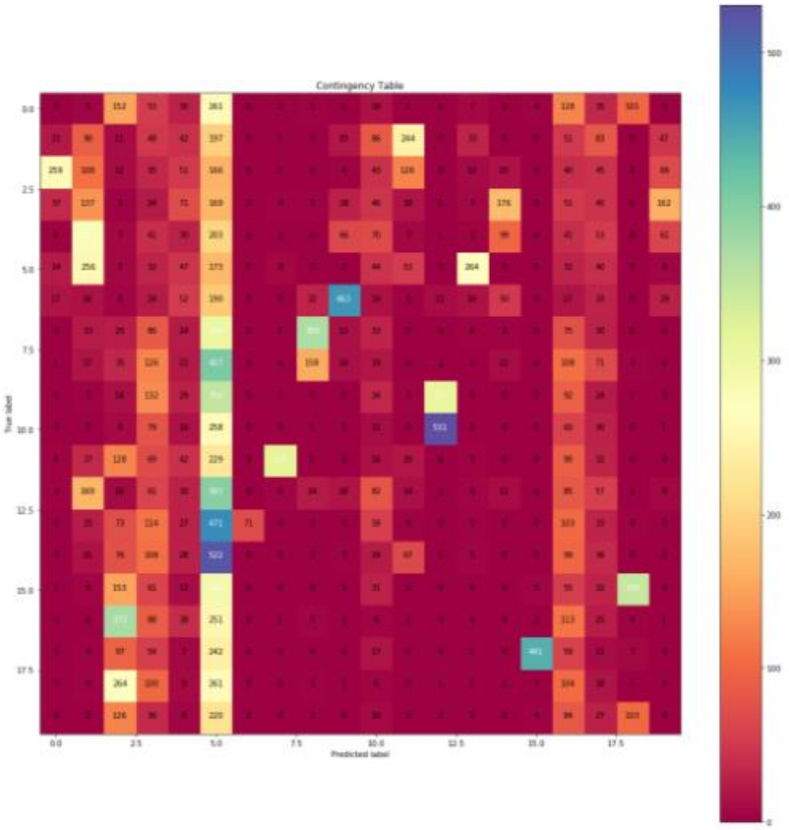
NMF Clusters for 20 Classes

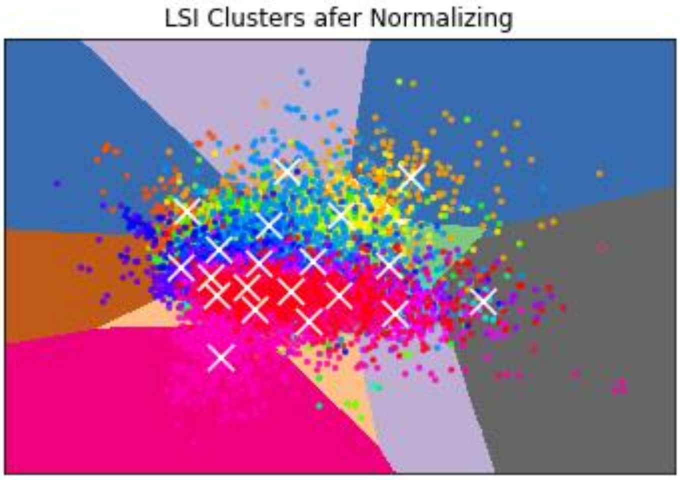
    b. 3 Methods:

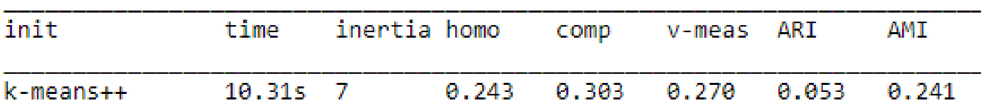
i. Normalizing:
   We performed scaling using sklearn “scale” to get unit variance and then normalized the data. Following are the results for LSI and NMF:



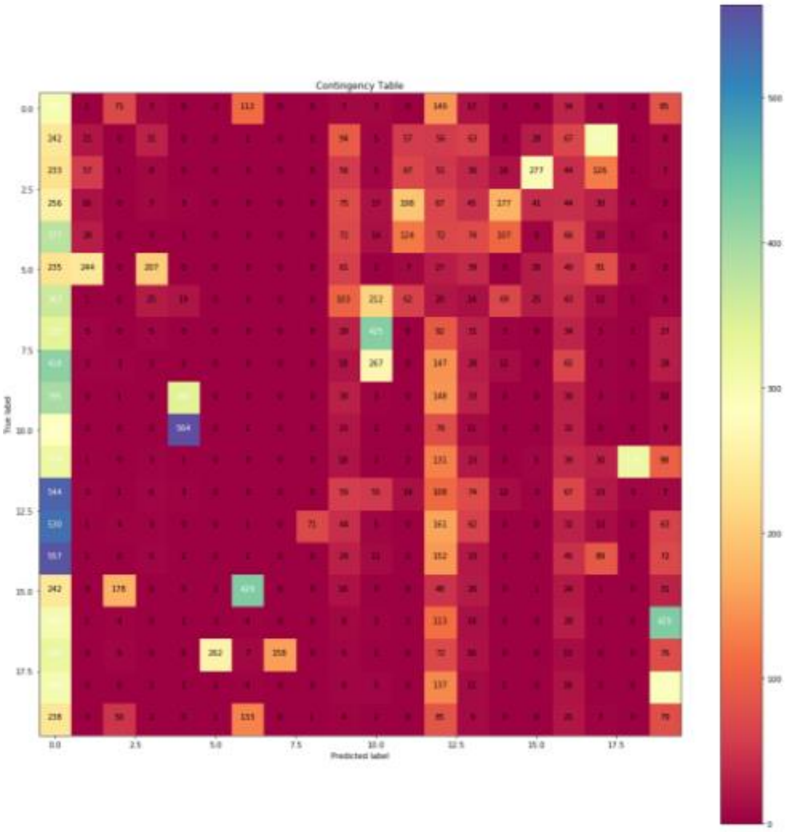
```
_____________________________________________________________________________
init            time     inertia homo     comp     v-meas   ARI      AMI
_____________________________________________________________________________
k-means++       11.56s   10       0.243    0.283    0.262    0.067    0.241
```



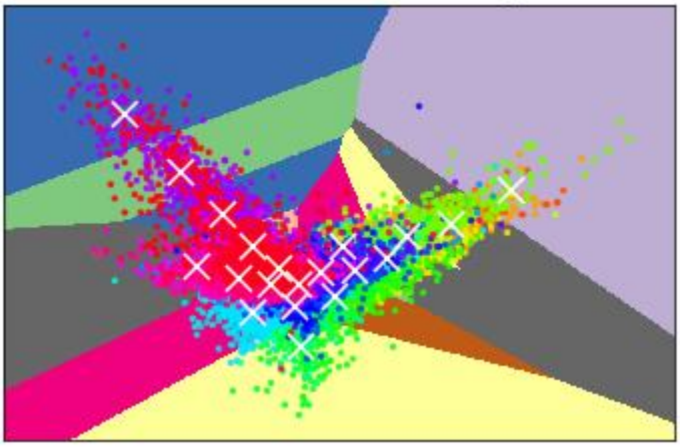




```
____________________________________________________________________________________
init              time      inertia  homo      comp      v-meas    ARI       AMI
____________________________________________________________________________________
k-means++         10.31s    7         0.243     0.303     0.270     0.053     0.241
```

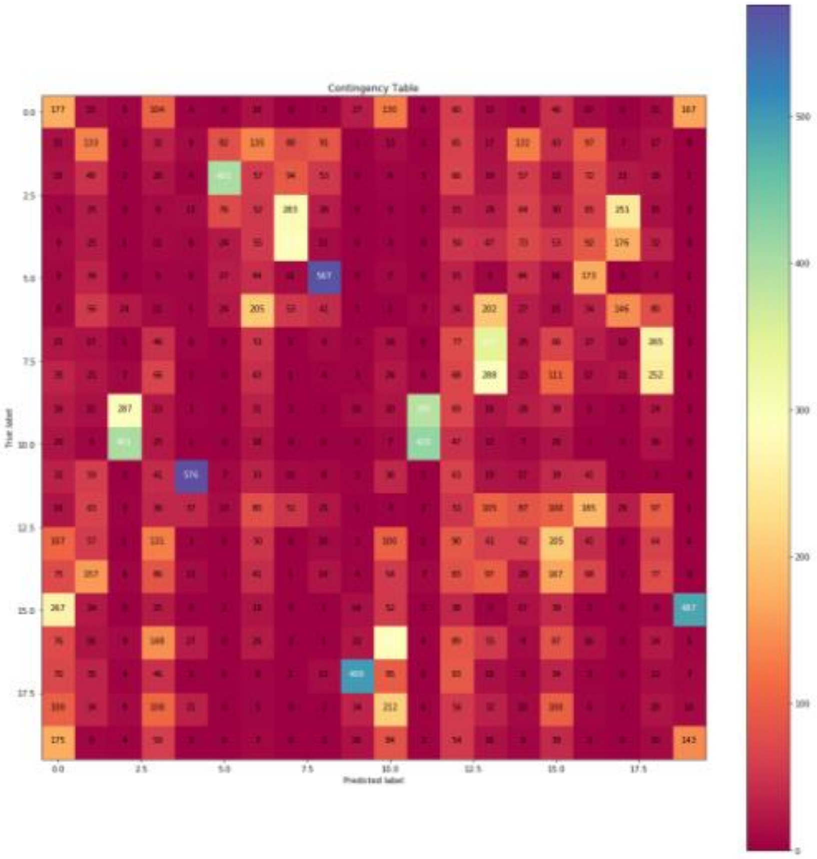


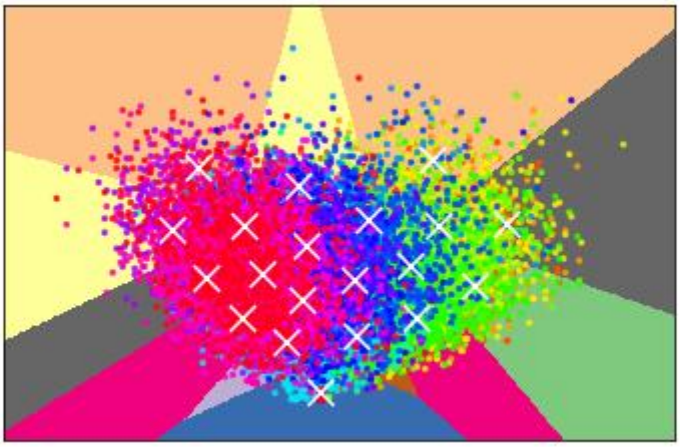
NMF Clusters afer Normalizing



ii. Logarithmic Transformation:

```
______________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
______________________________________________________________________________
k-means++       18.65s  253625  0.302   0.304   0.303   0.156   0.299
```

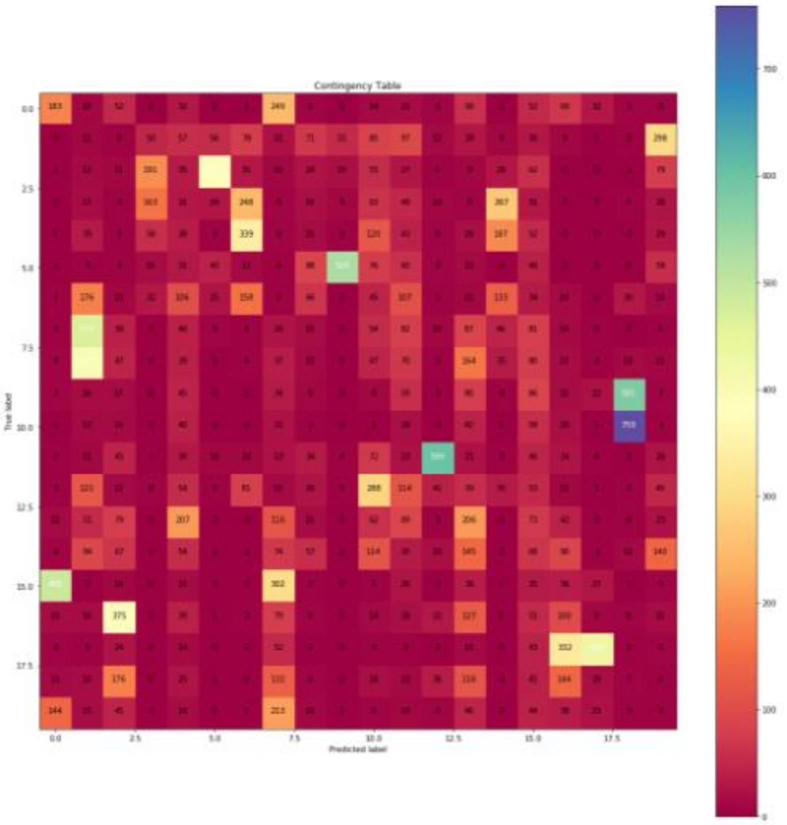


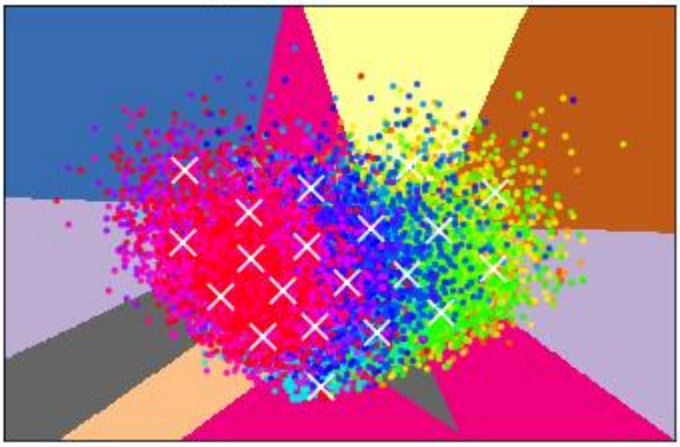

NMF Clusters afer Logarithmic Transformation

iii. Combination

1. Normalization and then Log Transformation
   We perform normalization and then apply log transformation on NMF data to get the following results:

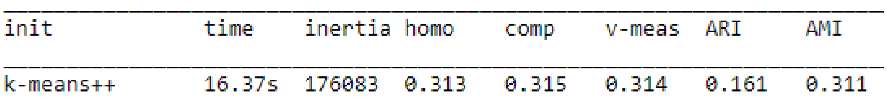
```
___________________________________________________________________________
init           time     inertia homo     comp     v-meas   ARI      AMI
___________________________________________________________________________
k-means++      12.30s   11       0.325    0.330    0.327    0.186    0.323
```



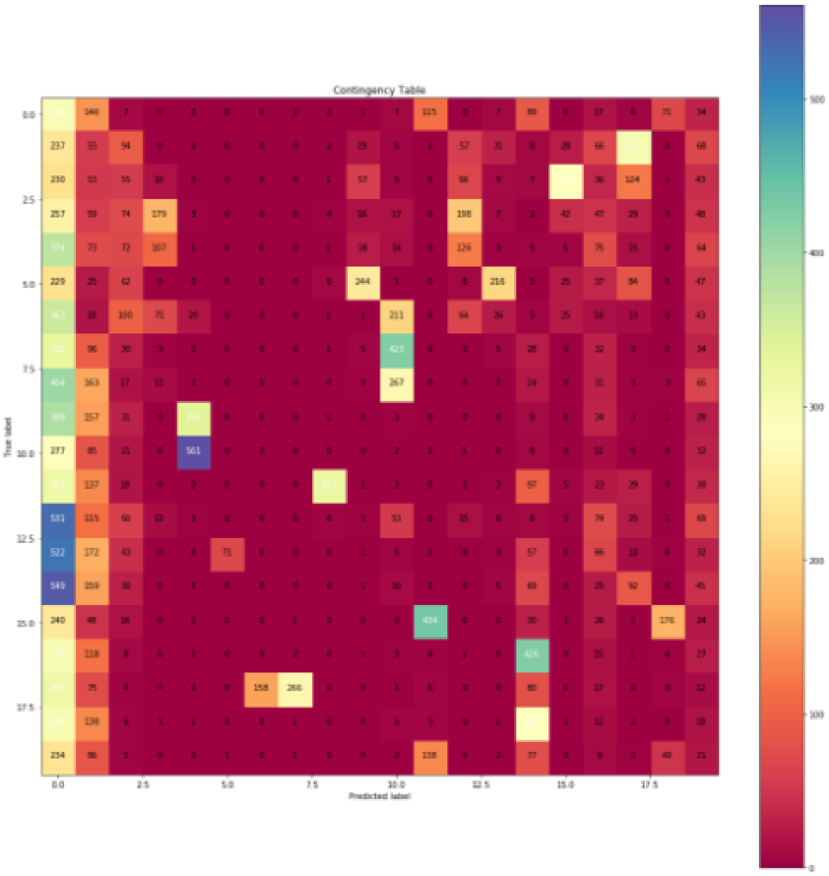


NMF Clusters afer Logarithmic Transformation and Normalizing

2. Log Transformation and then Normalization
   We apply log transformation on NMF data and then perform normalization to get the following results:



```
_____________________________________________________________________________
init            time    inertia homo    comp    v-meas  ARI     AMI
_____________________________________________________________________________
k-means++       16.37s  176083  0.313   0.315   0.314   0.161   0.311
```



285	146	7	0	0	0	0	2	2	1	7	115	0	7	89	0	17	6	71	34
237	55	94	0	0	0	0	0	2	19	5	1	57	31	8	28	66	302	0	68
230	53	55	16	0	0	0	0	1	57	5	0	66	9	7	283	36	124	1	43
257	59	74	179	3	0	0	0	4	16	17	0	198	7	2	42	47	29	0	48
374	73	72	107	1	0	0	0	1	28	14	0	126	3	5	5	75	15	0	64
229	25	62	0	0	0	0	0	8	244	1	0	8	216	2	25	37	84	0	47
363	18	100	71	20	0	0	0	1	1	211	0	64	26	5	25	14	13	0	43
330	96	30	3	0	0	0	0	1	5	423	0	0	5	28	0	32	3	0	34
404	163	17	12	2	0	0	0	0	3	267	0	0	2	24	0	31	3	3	65
389	157	31	0	339	0	0	0	1	0	3	0	0	0	9	0	34	3	1	28
277	85	21	0	561	0	0	0	0	0	2	1	1	0	8	0	11	0	0	32
315	137	18	0	0	0	0	0	321	1	2	0	2	3	57	5	23	29	0	38
531	115	60	12	3	0	0	0	6	3	53	0	15	6	8	3	74	25	1	68
522	172	43	0	0	71	0	0	0	1	5	1	0	3	57	0	66	13	4	32
549	159	30	0	1	0	0	0	0	1	10	1	0	5	69	0	25	92	0	45
240	48	16	0	0	0	0	1	0	0	0	454	0	0	30	1	26	1	176	24
299	118	8	0	1	0	0	2	0	1	3	4	1	0	426	0	15	1	4	27
315	75	4	0	0	0	158	266	0	0	1	6	0	0	80	0	17	0	6	12
297	138	6	1	1	0	0	1	0	0	5	5	0	1	287	1	12	2	0	18
234	86	5	0	0	1	0	1	0	0	2	138	0	2	77	0	9	3	49	21

NMF Clusters afer Normalizing and Logarithmic Transformation