ECE219 - Project 2 Report

Clustering

2/12/18

Jonathan Hurwitz (804258351)

Peter Kim (204271299)

1.

Here we transform our newsgroup documents into TF-IDF vector representations. We use min_df=3 and filter out stop words. Stemming was not involved. The final dimension of our matrix was 7882 documents X 27768 tokens.

2.

a) The contingency matrix after applying k-means with r=2 on the previous vectorized dataset was as follows:

4	3889
1728	2251

b) The clustering metrics were as follows:

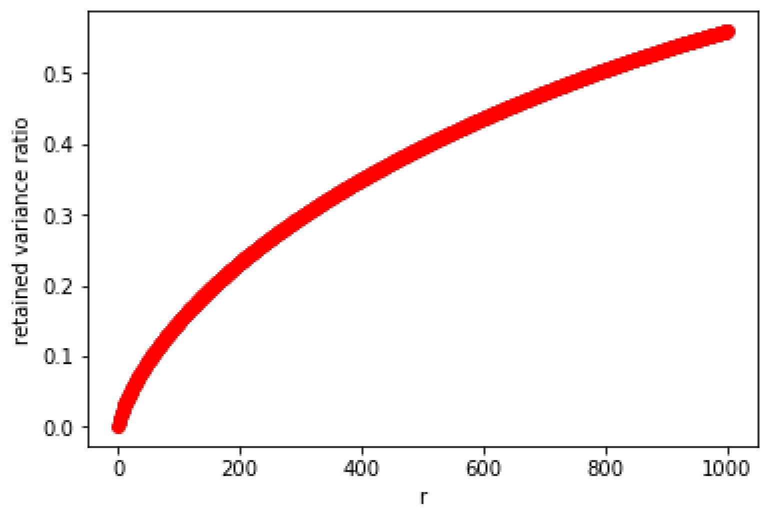
Homogeneity	0.253
Completeness	0.335
V-measure	0.288
Adjusted Rand-Index	0.181
Adjusted Mutual Info	0.253

The initial TF-IDF representations are high dimensional and sparse. The clustering performance given high dimensional and sparse data is poor. This is expected as Euclidean distances, which the k-means algorithm attempts to minimize every iteration, are essentially reduced to be the same for almost all pairs of examples in high-dimensional settings. This, therefore, renders k-means as an ineffective clustering technique for high-dimensions due to the curse of dimensionality.

3.

In this section, we preprocess the TF-IDF matrix by reducing the number of feature dimensions by applying both Latent Semantic Indexing (LSI) and Non-negative Matrix Factorization (NMF).

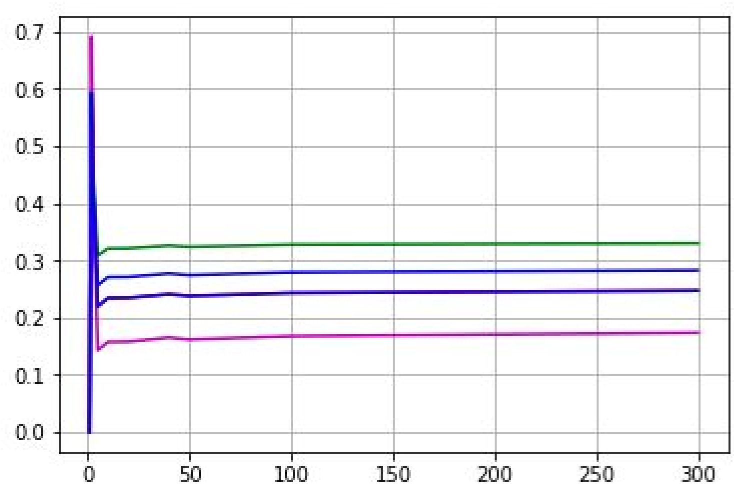
ai) We first attempted to find the most effective dimension to reduce to through inspecting change in the retained variance ratio when removing the top singular values. For the top 1000 principle components, we show the retained variance ratio in the following graph:



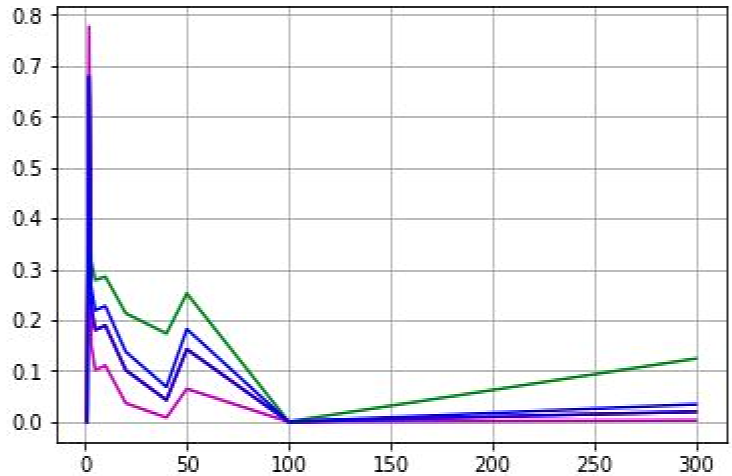
At r=1000 principal components, we retain around 60% of the total variance.

aii) Afterwards, we used the two methods of LSI and NMF and try r values in the range [1,2,3,5,10,20,50,100,300] and determined the clustering performance with respect to the five metrics mentioned above.

When reducing with LSI and NMF, we get the following performance scores across our range of r.



(LSI) r vs. metric scores



(NMF) r vs. metric scores

The non-monotonicity of the two above graphs can be explained as such: as we increase the number of dimensions r, more information is encapsulated in the feature representation just by nature of having a higher percentage of the variance due to more principal components. However, K-Means suffers from the curse of dimensionality, so as the dimensions increase, Euclidean distance becomes a worse and worse distance metric and will perform poorly. Thus, we are trading off between more information and an inherent shortcoming in the K-Means method. This is why it's possible for kmeans to perform better with lower dimension (such as r=100) than with higher dimension (such as r=300 or r=1000).

From the above graphs as well as the metrics shown in the table below, it is clear that **the best r-value for both LSI and NMF is r=2**. As mentioned before, we are trading off between incorporating more information and the high dimensionality shortcomings of K-Means. It makes sense that a low dimension value such as 2 would be effective for K-Means, due to the aforementioned curse of dimensionality.

The contingency matrices and corresponding metrics for all cases are summarized in the following table:

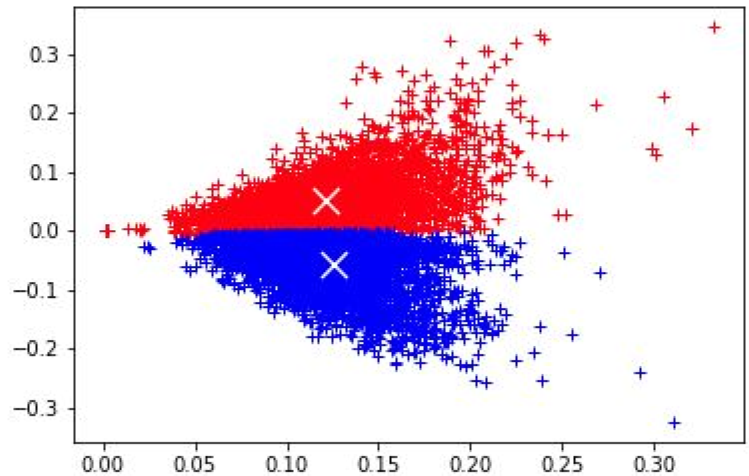
r-value	**LSI**	**NMF**
1	homogeneity=0.00030 completeness=0.00030 v-measure=0.00030 adj rand index=0.00034 adj mutual info=0.00021 [[1703 2200] [1656 2323]]	homogeneity=0.00030 completeness=0.00030 v-measure=0.00030 adj rand index=0.00034 adj mutual info=0.00021 [[2200 1703]

		[2323 1656]]
2 Best value for LSI Best value for NMF	**homogeneity=0.59263 completeness=0.59424 v-measure=0.59343 adj rand index=0.69180 adj mutual info=0.59259** **[[3713 190] [473 3506]]**	**homogeneity=0.67905 completeness=0.68013 v-measure=0.67959 adj rand index=0.77702 adj mutual info=0.67902** [[3594 309] [158 3821]]
3	homogeneity=0.40097 completeness=0.43866 v-measure=0.41897 adj rand index=0.39656 adj mutual info=0.40091 [[3866 37] [1422 2557]]	homogeneity=0.22934 completeness=0.31648 v-measure=0.26596 adj rand index=0.15280 adj mutual info=0.22927 [[3899 4] [2396 1583]]
5	homogeneity=0.21960 completeness=0.30838 v-measure=0.25653 adj rand index=0.14284 adj mutual info=0.21953 [[3898 5] [2446 1533]]	homogeneity=0.18063 completeness=0.27871 v-measure=0.21920 adj rand index=0.10196 adj mutual info=0.18056 [[3898 5] [2677 1302]]
10	homogeneity=0.23427 completeness=0.32095 v-measure=0.27084 adj rand index=0.15740 adj mutual info=0.23420 [[3900 3] [2374 1605]]	homogeneity=0.18920 completeness=0.28527 v-measure=0.22751 adj rand index=0.11056 adj mutual info=0.18913 [[3898 5] [2625 1354]]
20	homogeneity=0.23463 completeness=0.32122	homogeneity=0.10163 completeness=0.21326 v-measure=0.13766

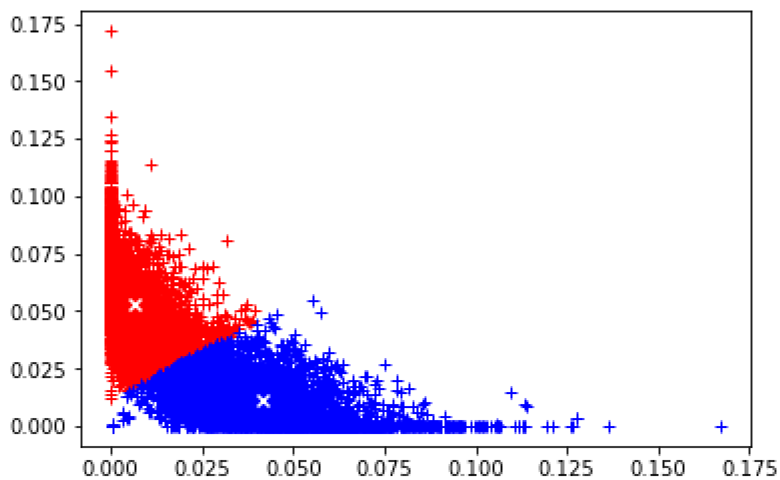
	v-measure=0.27118 adj rand index=0.15780 adj mutual info=0.23455 [[3 3900] [1607 2372]]	adj rand index=0.03660 adj mutual info=0.10155 [[7 3896] [800 3179]]
50	homogeneity=0.23801 completeness=0.32377 v-measure=0.27434 adj rand index=0.16165 adj mutual info=0.23794 [[3900 3] [2353 1626]]	homogeneity=0.14268 completeness=0.25245 v-measure=0.18232 adj rand index=0.06493 adj mutual info=0.14260 [[2 3901] [1045 2934]]
100	homogeneity=0.24321 completeness=0.32768 v-measure=0.27920 adj rand index=0.16763 adj mutual info=0.24314 [[3900 3] [2324 1655]]	homogeneity=0.00002 completeness=0.00045 v-measure=0.00003 adj rand index=0.00001 adj mutual info=-0.00008 [[3887 16] [3965 14]]
300	homogeneity=0.24740 completeness=0.33013 v-measure=0.28284 adj rand index=0.17350 adj mutual info=0.24733 [[4 3899] [1684 2295]]	homogeneity=0.01995 completeness=0.12427 v-measure=0.03438 adj rand index=0.00283 adj mutual info=0.01986 [[3723 180] [3974 5]]

4.

a) The best clustering results for LSI and NMF were obtained for r=2. This value was used in the dimensionality reduction. First, we visualized the performance of the best clustering results by projecting the final data vectors onto a 2 dimensional place and color coding the two classes.



LSI-reduced clustering result with r=2



NMF-reduced clustering result with r=2

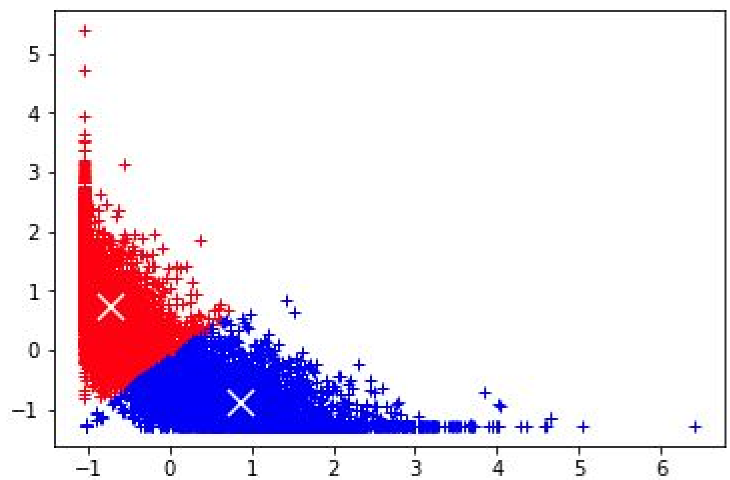
We are content with these clustering results as the two different types of classes are generally centered around their respective centroids (marked with a white X) in a cohesive manner.

b) The next step was to try several different methods to see if they increase clustering performance for r=2 NMF reduced data. The methods and their results are described below:

1) Normalizing the features such that each feature has unit variance.

   The table below shows a comparison between the baseline K-Means + NMF (r=2) results and the normalized NMF (r=2) results. NMF with normalization results in marginally better homogeneity, completeness, v-measure, and adjusted mutual information. However, adjusted random index is marginally lower.

NMF (r=2) baseline performance	NMF (r=2) with normalization performance
homogeneity=0.67905 **completeness=0.68013** **v-measure=0.67959** **adj rand index=0.77702** **adj mutual info=0.67902** [[3594 309] [158 3821]]	**homogeneity=0.68280** **completeness=0.68564** **v-measure=0.68422** **adj rand index=0.77344** **adj mutual info=0.68277** [[369, 3534], [3873, 106]]

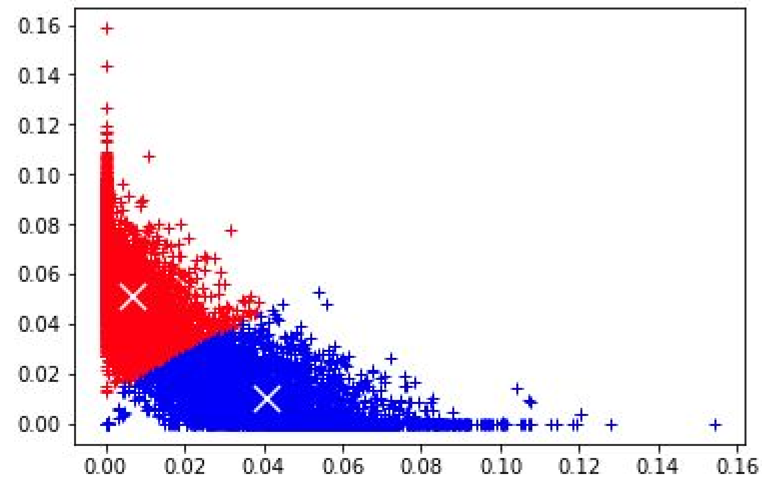


NMF (r=2) with Normalization

2) Applying non-linear transformation (log transformation).

   The table below shows a comparison between the baseline performance and K-Means after a log transformation is applied. All of the metrics decrease.

NMF (r=2) baseline performance	NMF (r=2) with log transformation performance
homogeneity=0.67905 **completeness=0.68013** **v-measure=0.67959** **adj rand index=0.77702** **adj mutual info=0.67902** [[3594 309] [158 3821]]	**homogeneity=0.67505** **completeness=0.67641** **v-measure=0.67573** **adj rand index=0.77255** **adj mutual info=0.67503** [[325, 3578], [3827, 152]]

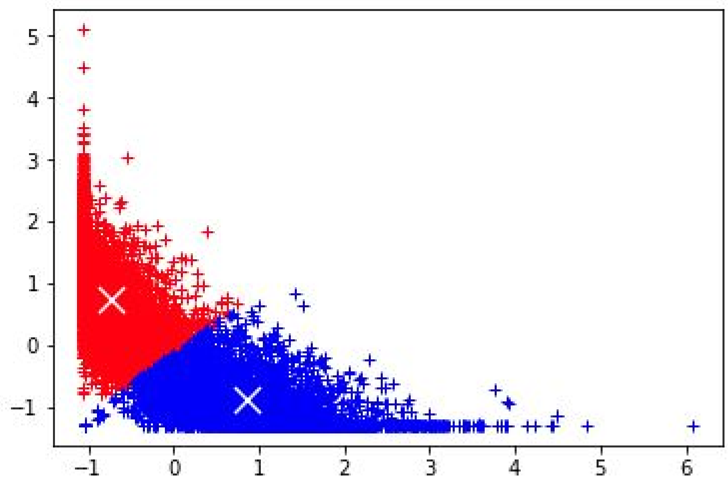


NMF (r=2) with log transformation

3) Applying log transformation then normalization.

   NMF with log transformation then normalization showed an improvement across all metrics except for adjusted rand index.

NMF (r=2) baseline performance	NMF (r=2) with log transformation then normalization performance
homogeneity=0.67905 **completeness=0.68013** **v-measure=0.67959** **adj rand index=0.77702** **adj mutual info=0.67902** [[3594 309] [158 3821]]	**homogeneity=0.68369** **completeness=0.68649** **v-measure=0.68509** **adj rand index=0.77434** **adj mutual info=0.68366** [[367, 3536], [3873, 106]]

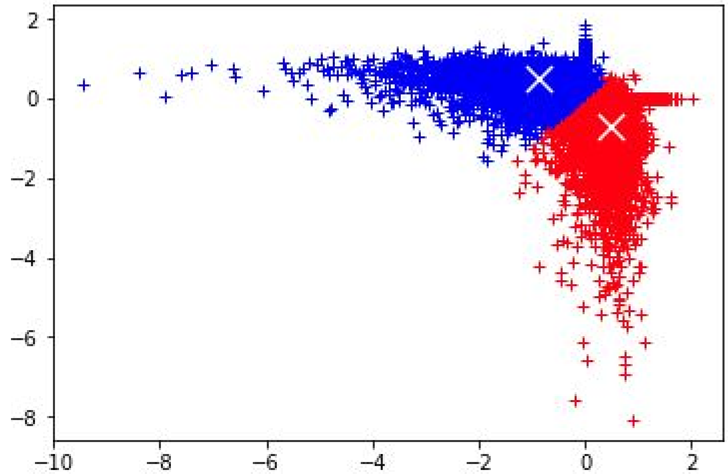


NMF (r=2) with log transformation then normalization

4) Applying normalization then log transformation.

   NMF with normalization and then log transformation showed a decrease in performance across all categories.

NMF (r=2) baseline performance	NMF (r=2) with normalization then log transformation performance
homogeneity=0.67905 **completeness=0.68013** **v-measure=0.67959** **adj rand index=0.77702** **adj mutual info=0.67902** [[3594 309] [158 3821]]	**homogeneity=0.65169** **completeness=0.65196** **v-measure=0.65183** **adj rand index=0.75525** **adj mutual info=0.65166** [[3613, 290], [226, 3753]]



NMF (r=2) with normalization then log transformation

A non-linear transformation such as a logarithmic transformation can increase the relationship between independent features, which are the case for our TFIDF matrix. This can add linearity to our data and thus, could increase the clustering results. However, in our case of applying logarithmic transformation to the NMF-reduced dataset, we empirically show that the performance was actually hurt by such a nonlinear transformation (~1% decrease).

5.

This section involved clustering of 20 categories rather than just 2. All documents were included in the data matrix. The same parameters as in part 1 were used with regards to stemming and min_df. We ran K-Means with varying values of r for both LSI and NMF dimensionality reduction to find an optimal value for each. The results from this sweep are summarized in the table below.

r-value	**LSI**	**NMF**
1	homogeneity=0.01516 completeness=0.01648 v-measure=0.01579 adj rand index=0.00313 adj mutual info=0.01196	homogeneity=0.01546 completeness=0.01666 v-measure=0.01604 adj rand index=0.00316 adj mutual info=0.01226
2	homogeneity=0.17440 completeness=0.18632	homogeneity=0.16135 completeness=0.17094 v-measure=0.16601

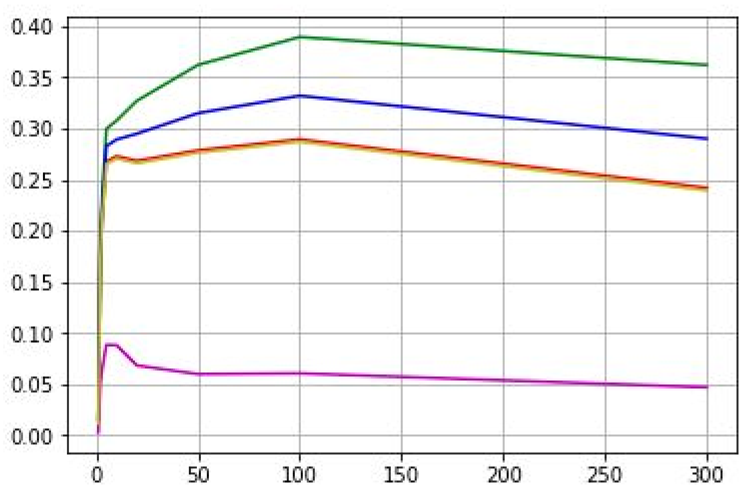
	v-measure=0.18016 adj rand index=0.05036 adj mutual info=0.17173	adj rand index=0.04657 adj mutual info=0.15863
3	homogeneity=0.21548 completeness=0.23676 v-measure=0.22562 adj rand index=0.06647 adj mutual info=0.21293	homogeneity=0.18779 completeness=0.21056 v-measure=0.19852 adj rand index=0.05472 adj mutual info=0.18515
5	homogeneity=0.26561 completeness=0.29715 v-measure=0.28050 adj rand index=0.08719 adj mutual info=0.26323	homogeneity=0.21922 completeness=0.26075 v-measure=0.23819 adj rand index=0.06541 adj mutual info=0.21666
10 Best value for NMF	homogeneity=0.27704 completeness=0.31739 v-measure=0.29584 adj rand index=0.08653 adj mutual info=0.27469	**homogeneity=0.26157** **completeness=0.31635** **v-measure=0.28636** **adj rand index=0.07140** **adj mutual info=0.25917**
20	homogeneity=0.28841 completeness=0.35187 v-measure=0.31699 adj rand index=0.07260 adj mutual info=0.28609	homogeneity=0.22945 completeness=0.29031 v-measure=0.25632 adj rand index=0.04733 adj mutual info=0.22694
50	homogeneity=0.29119 completeness=0.38053 v-measure=0.32992 adj rand index=0.06287 adj mutual info=0.28888	homogeneity=0.16791 completeness=0.24519 v-measure=0.19932 adj rand index=0.02413 adj mutual info=0.16520
100 Best value for LSI	**homogeneity=0.29706** **completeness=0.38877** **v-measure=0.33679**	homogeneity=0.06814 completeness=0.11786 v-measure=0.08636

	adj rand index=0.06609 **adj mutual info=0.29477**	adj rand index=0.00549 adj mutual info=0.06504
300	homogeneity=0.28491 completeness=0.39868 v-measure=0.33233 adj rand index=0.06989 adj mutual info=0.28256	homogeneity=0.07278 completeness=0.12840 v-measure=0.09290 adj rand index=0.00654 adj mutual info=0.06973

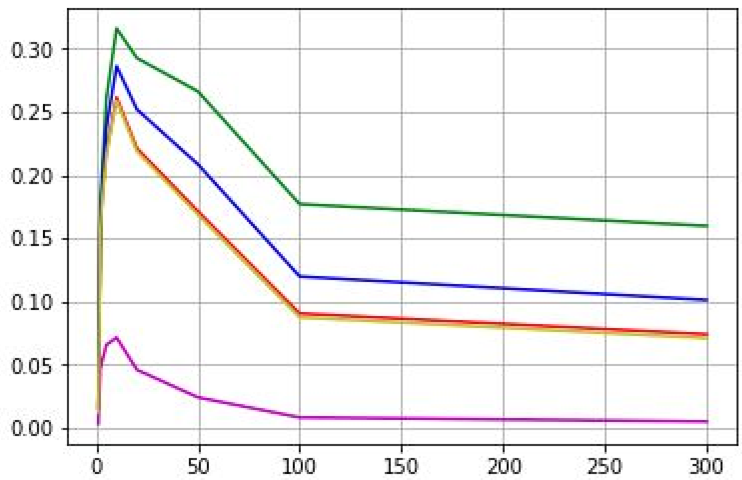
The best r-value for LSI was r=100. The best value for NMF was r=10.

The following plots show the performance for both LSI and NMF as r-value increases. The y-axis is units of whatever metric is being displayed on a scale from 0 to 1. The x-axis represents r-value. The legend is as follows:

Color	**Metric**
red	Homogeneity
green	Completeness score
blue	V-measure
magenta	Adjusted rand-index
yellow	Adjusted mutual information



(LSI) metric scores vs. r-value



(NMF) metric scores vs. r-value

The optimal r-values for LSI and NMF (100 and 10 respectively) were used in the transformation tests performed below. The same four methods were tested:

1) Normalizing the features such that each feature has unit variance.

   For LSI, applying normalization worsened all of the metrics. Normalization improved adjusted rand index for NMF, but worsened all of the other metrics.

LSI baseline (r=100)	LSI normalized (r=100)	NMF baseline (r=10)	NMF normalized (r=10)
homogeneity=0.29706 completeness=0.38877 v-measure=0.33679 adj rand index=0.06609 adj mutual info=0.29477	homogeneity=0.19519 completeness=0.27588 v-measure=0.22863 adj rand index=0.02468 adj mutual info=0.19257	homogeneity=0.26157 completeness=0.31635 v-measure=0.28636 adj rand index=0.07140 adj mutual info=0.25917	homogeneity=0.25016 completeness=0.29413 v-measure=0.27036 adj rand index=0.08328 adj mutual info=0.24772

2) Applying non-linear transformation (log transformation).

   Applying the log transform to LSI lowered all metric values. For NMF, homogeneity, v-measure, adjusted rand index, and adjusted mutual information improved while completeness marginally decreased.

LSI baseline (r=100)	LSI log (r=100)	NMF baseline (r=10)	NMF log (r=10)
homogeneity=0.29706 completeness=0.38877 v-measure=0.33679	homogeneity=0.28039 completeness=0.37745 v-measure=0.32176	homogeneity=0.26157 completeness=0.31635 v-measure=0.28636 adj rand index=0.07140	homogeneity=0.26394 completeness=0.31519 v-measure=0.28729 adj rand index=0.07274

adj rand index=0.06609 adj mutual info=0.29477	adj rand index=0.06485 adj mutual info=0.27803	adj mutual info=0.25917	adj mutual info=0.26154

3) Applying log transformation then normalization.

Applying log transformation and then normalization worsened all of the metrics across the board for both LSI and NMF, with the exception of adjusted rand index for NMF.

LSI baseline (r=100)	LSI log + scale (r=100)	NMF baseline (r=10)	NMF log + scale (r=10)
homogeneity=0.29706 completeness=0.38877 v-measure=0.33679 adj rand index=0.06609 adj mutual info=0.29477	homogeneity=0.20665 completeness=0.30849 v-measure=0.24750 adj rand index=0.03525 adj mutual info=0.20404	homogeneity=0.26157 completeness=0.31635 v-measure=0.28636 adj rand index=0.07140 adj mutual info=0.25917	homogeneity=0.24761 completeness=0.28635 v-measure=0.26558 adj rand index=0.07678 adj mutual info=0.24516

4) Applying normalization then log transformation.

Applying feature scaling before the log transformation resulted in score improvements for all metrics across the board. **Empirically, this was the best method, so we can conclude that feature scaling should be applied before any type of non-linear transformation.**

LSI baseline (r=100)	LSI scale + log (r=100)	NMF baseline (r=10)	NMF scale + log (r=10)
homogeneity=0.29706 completeness=0.38877 v-measure=0.33679 adj rand index=0.06609 adj mutual info=0.29477	homogeneity=0.30133 completeness=0.38911 v-measure=0.34348 adj rand index=0.07849 adj mutual info=0.30230	homogeneity=0.26157 completeness=0.31635 v-measure=0.28636 adj rand index=0.07140 adj mutual info=0.25917	homogeneity=0.26450 completeness=0.27809 v-measure=0.27113 adj rand index=0.12484 adj mutual info=0.26212