**Data Parsing and Pre-process**

Done..scraping.py, tokenStemRemove.py

**Feature Extraction**

**The following types of features have been applied as style markers can be distinguished:**

1. **Term Frequency- Inverse Document Frequency**

We choose **bag of words**  which is a sparse vector of occurrence counts of **words as a representation of works, and we use tf–idf**, short for **term frequency–inverse document frequency**, is a numerical statistic, intended to reflect how important a word is to a [document](https://en.wikipedia.org/wiki/Document) in a collection or [corpus](https://en.wikipedia.org/wiki/Text_corpus).We extracted features by counting unigram, 2-grams, 3-grams, 4-grams this way we capture phrases and multi-word expressions. The extracted features are exported as arrays of dimension ( n x m ), n is the number of plays or scenes, or poems, the m is the number of features, and we limit the number of features to 5000. The multiple vectors are saved into txt files for further processes.

Below is an excerpt of the 5000 4-grams vocabulary vectorized from Shakespear’s poems:

{'pin herenmistress quicklyo word': 1105,

'post henc seriou matter': 4521,

'physician enemi peac troop': 684,

'play thou strive poor': 1917,

'play game thou art': 1783,

'pleasant princ mock hath': 2448,

'pound yearrobertgraciou lieg father': 4735,

Latent Semantic Analysis is applied then to reduce the dimensionality from 5000 to 2, however the explained variance of the LSA reduced model is about 5% to 10% only.

1. **Type Token Ratio( Richness of vocabulary)**

We believe different authors are likely to have different level of vocabulary richness reflected in the works, and therefore we extracted the type-token ratio as a measure of vocabulary richness from Shakespear’s plays. Type-token ratio is calcuated as V/N, V representing the size of the vocabulary(unique tokens) of the sample, and N the number of tokens.

Referrering: <http://www.lexically.net/downloads/version5/HTML/index.html?type_token_ratio_proc.htm>

1. **Syntactic Features (Readability)**

Syntactic features have been proposed as more reliable style markers since they are not under the conscious control of the author (Baayen et al., 1996; Diederich et al., 2000; Khmelev and Tweedie, 2001; Kukushkina et al., 2001; Stamatatos et al., 1999).

We counted the number of sentences, words, letters, avg lengths of words, avg lengths of setences, for each act, scene and play of Shakespear’s plays. Besides them, we also calculated Coleman–Liau index (CLI), which is a readability test designed by Meri Coleman and T. L. Liau to gauge the understandability of a text (<https://en.wikipedia.org/wiki/Coleman%E2%80%93Liau_index>)

The Coleman–Liau index is calculated with the following formula:

CLI = 0.0588{L} - 0.296{S} - 15.8\,\!

*L* is the average number of letters per 100 words and *S* is the average number of sentences per 100 words.

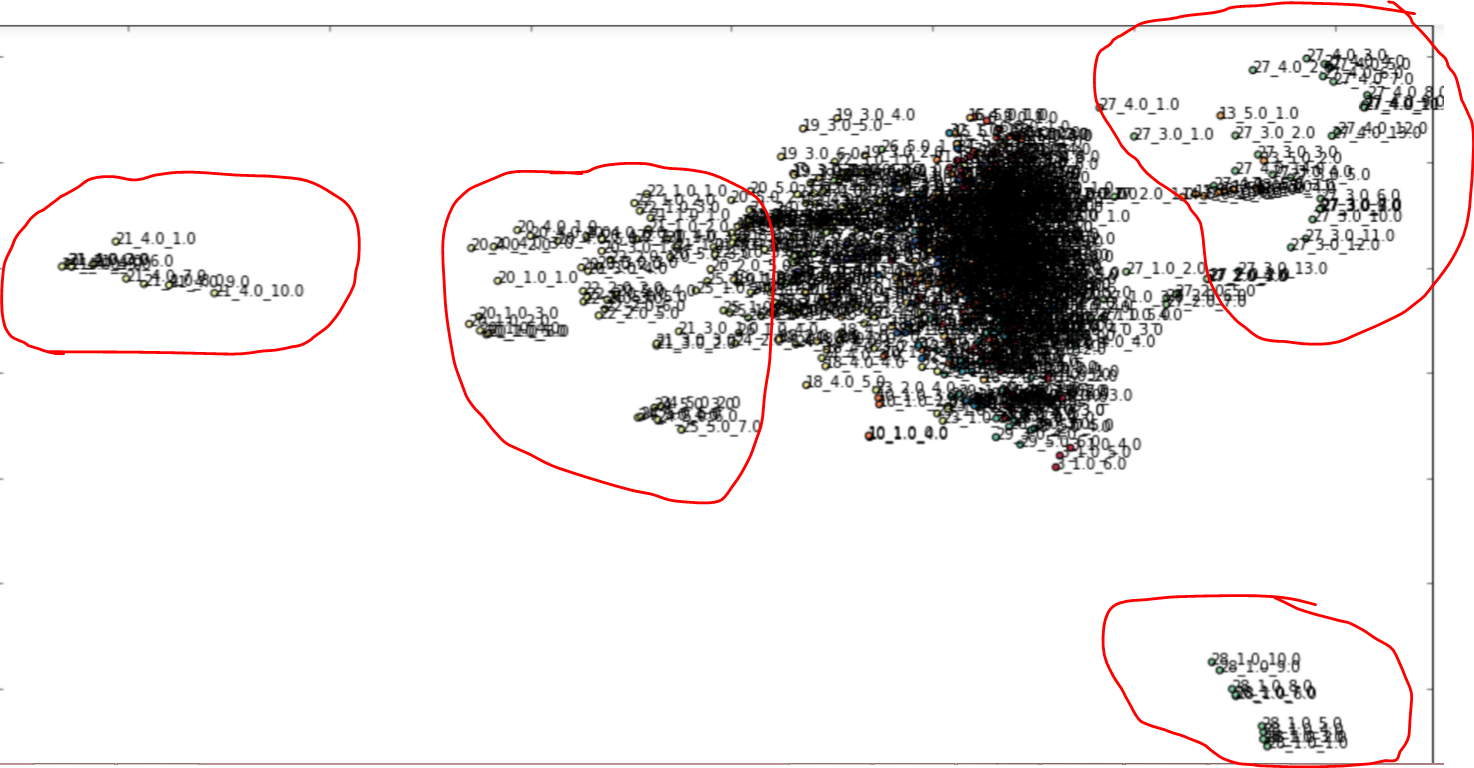
Those features above are exported as .txt files in Readability\_Features folder, and there sure will be some multicollinearity among them, and we will use PCA to remove the multicollinearity.

**Visualization**

1. **PCA projection**

We first use PCA to reduce the dimensions of n-grams vectorization and readability features, and project the first two components, and here are some observations from the three types of features we extracted .

In the PCA processed unigram vectorization of Shakespears’ plays in each scene, we see that scenes from play 20,21,22,25,27,28 clusters in their own and seperated from the rest of scenes of plays.



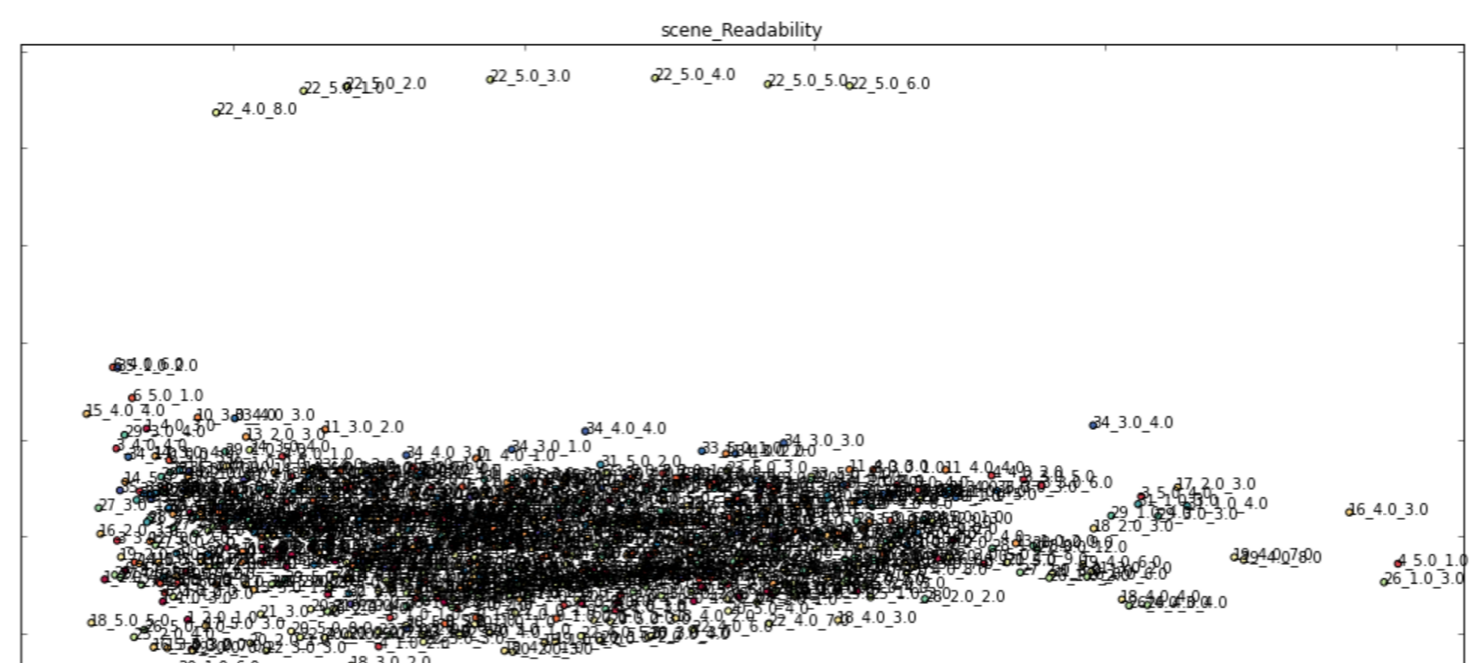
Some scenes from play 27

Some scenes from play 28

Some scenes from play 21

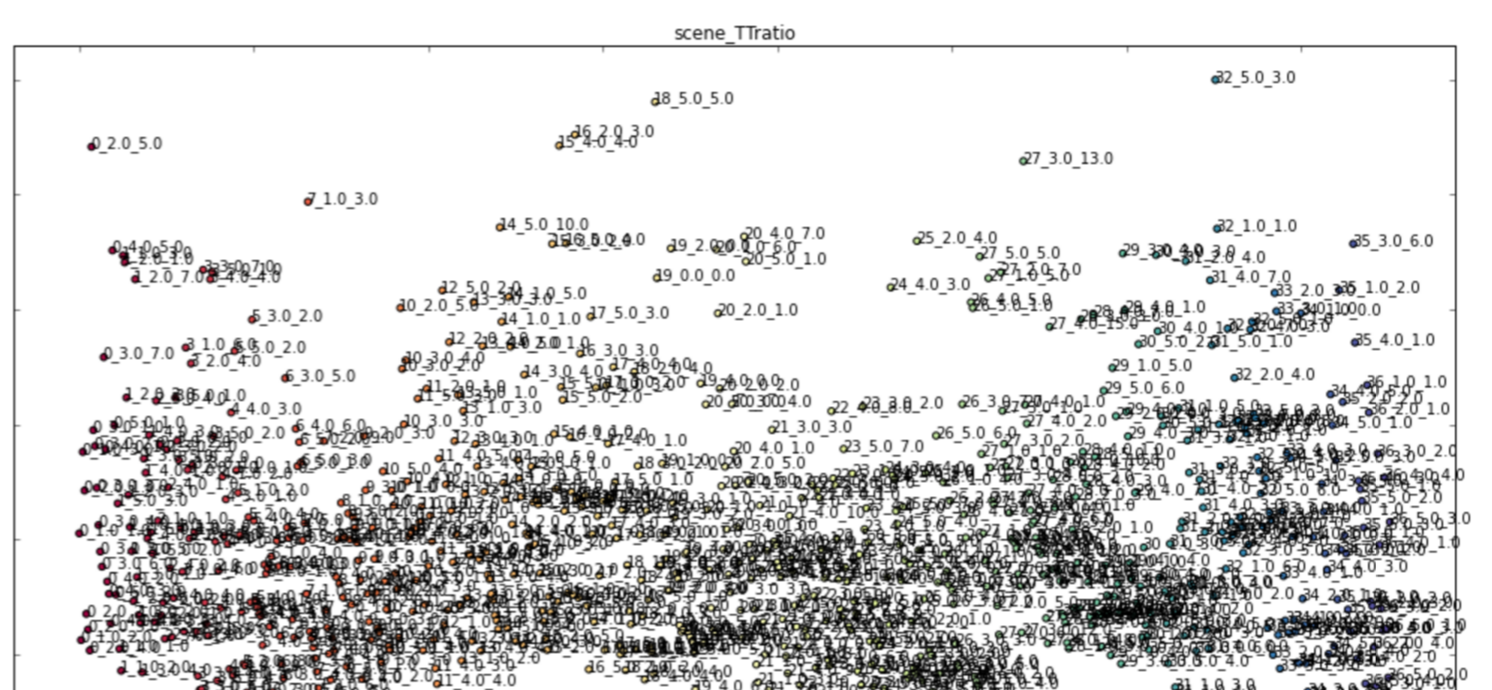
Some scenes from play 20,21,22,25

In the PCA processed readability features of Shakespears’ plays in each scene, we see that scenes from paly 22 have a distinct level of readabiliy(syntactic features) than other plays.



Some scenes from play 22

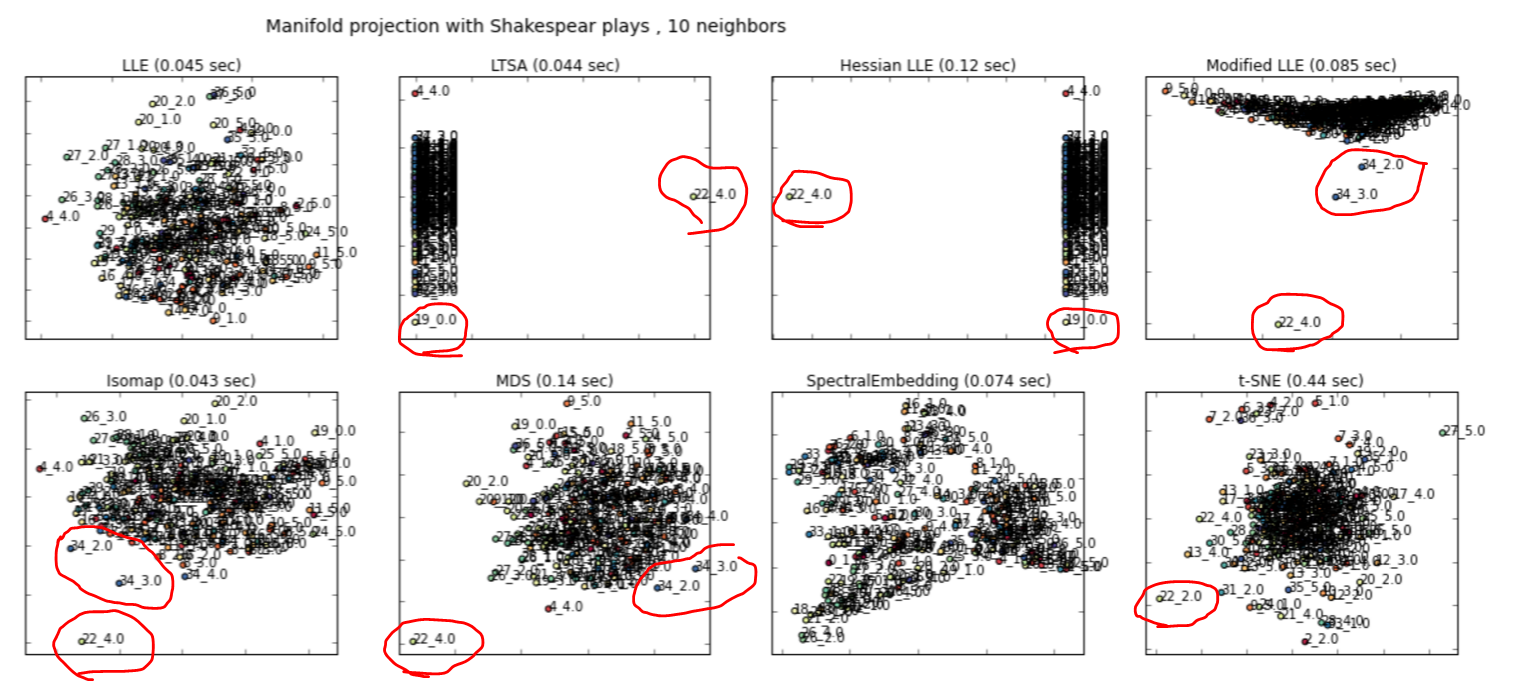
The Type Token Ratio feature is an one column vector, and thus there is no need to use PCA to reduce dimensions to visualize, here the x-axis is the play number, and the Y- axis shows the Type Token Ratio, and we can see some scenes from plays 0,7,17,15,16,18,22,27,32 have higher Type Token Ratio, indicating more vocabulary richness than the rest of plays.



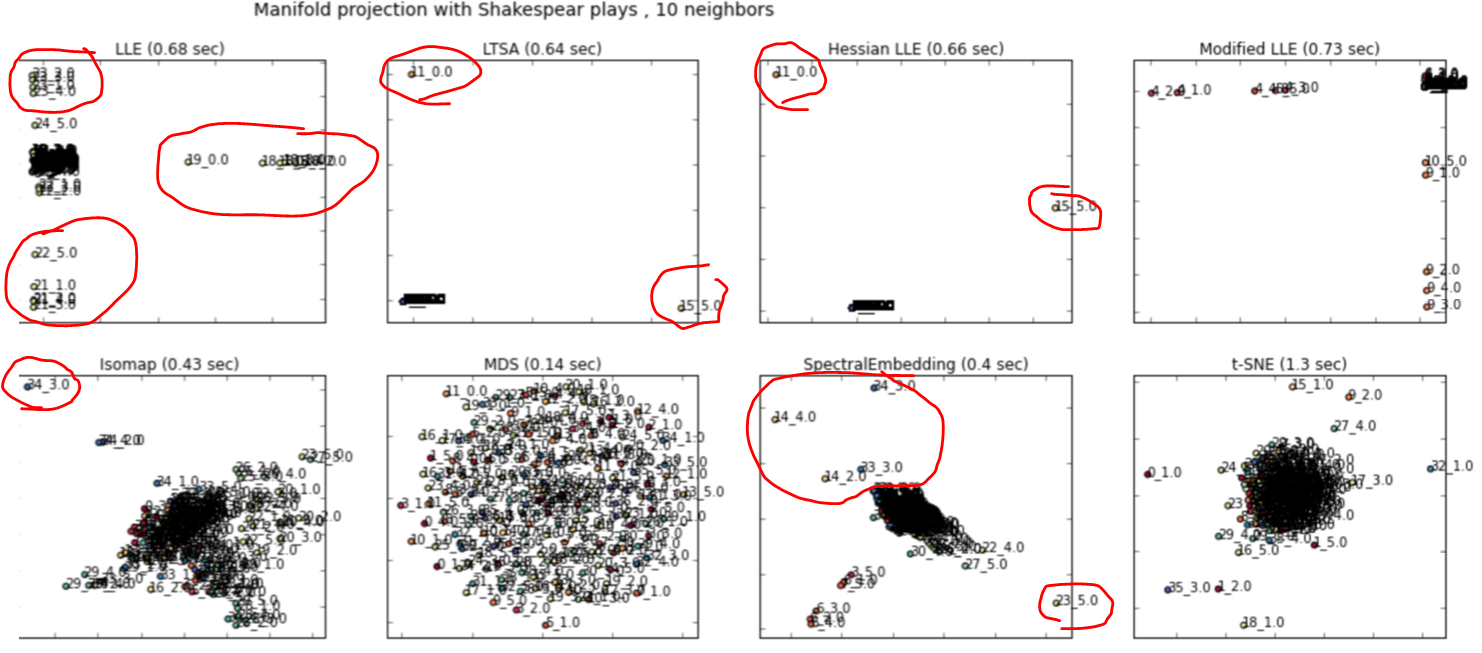
1. **Manifold projection on non-linear plane**

We also tried manifold projection methods including LLE,LTSA,Hessian LLE, Modified LLE, Isomap, MDS, SpectralEmbedding, t-SNE dimension reduction methods and visualize those proejctions.

In the scene readability manifold projection graphs, we can see that scenes 3 and 4 from play 22 , scene 0 from play 19 and scene 2 and 3 from play 34 appear quiet a few times as outliers than the rest plays:



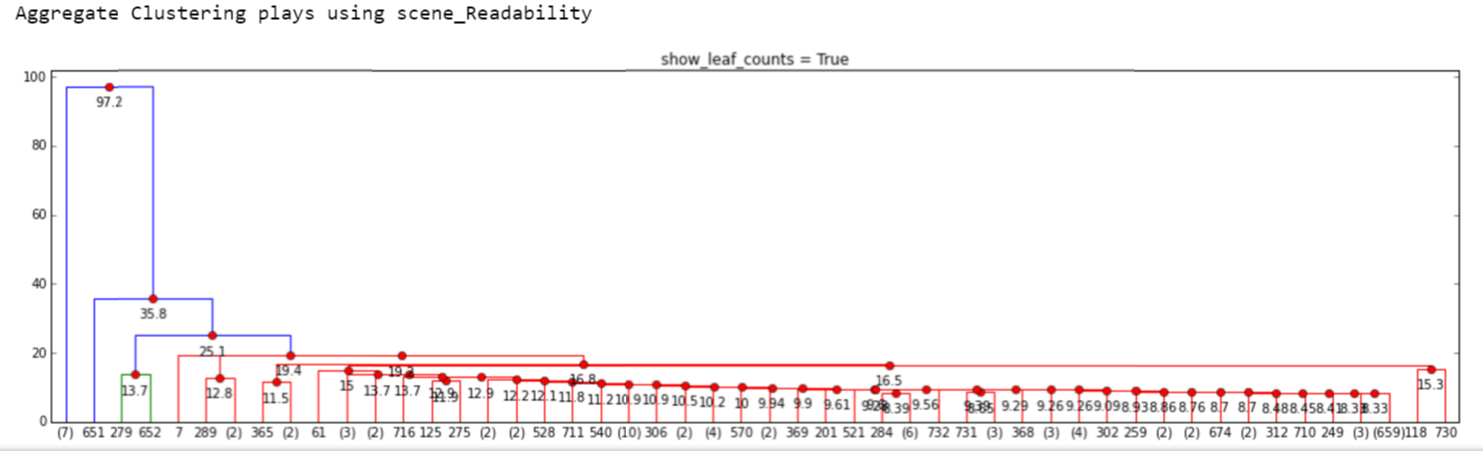
In the scene unigram vectorization manifold projection graphs, we can see that some scenes from plays 11,15,19,22,23,24,34 quiet a few times as outliers than the rest plays:

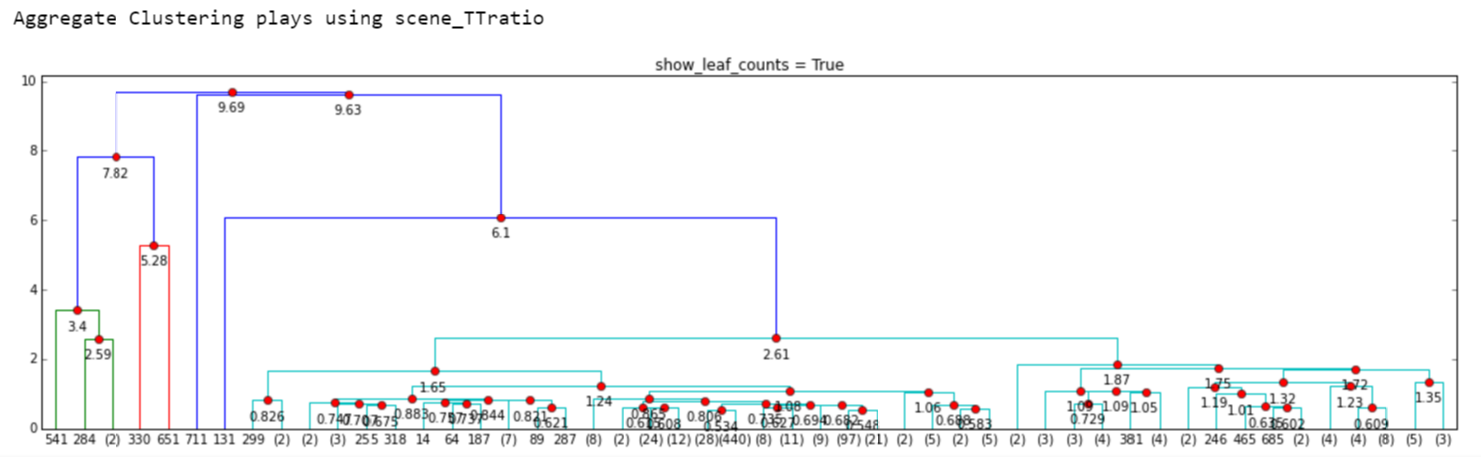


**Space seems not enough: need polish some ipyrs and point grader to read those for details of visualization such n-grams PCA and Manifold projection**

**Clustering**

1. **KM**
2. **GMM**
3. **DBSCAN**
4. **Hierachical Clustering**





**Conclusion**

K means clustered Bag of words features extract( vectorized and LSA,PCA dimensionality reduced) generate the following Outliers plays( when set cluster numebr to 2), and the scenes, acts outliers are also mostly coming from the following plays, which validates consistency

['the first part of king henry the fourth']

['the second part of king henry the fourth']

['the life of king henry the fifth']

['the first part of king henry the sixth']

['the second part of king henry the sixth']

['the third part of king henry the sixth']

['the life of king henry the eighth']

['the life and death of king john']

['the life and death of richard the second']

['the life and death of richard the third']