**Homework 4: To be or not to be…the author – Report**

Bin Gao, Bin Yan

# Summary

Something like this: We split the complete works of William Shakespeare into different acts. 13 features are extracted from all these acts, and normalized for further clustering. We have applied which, which, which… clustering methods. We concluded which method is better in terms of (accuracy, performance). Based on the result, we deduced that which/those acts are most likely not written by William Shakespeare because of (some short reasons).

# Part I Data Wrangling

To collect the necessary data, we have different options initially. The complete works of William Shakespeare is in HTML format in <http://shakespeare.mit.edu/>. To avoid the problem of downloading data from different links or the time required doing it by hand, we located the complete works of William Shakespeare from Project Gutenberg at <http://www.gutenberg.org/cache/epub/100/pg100.txt>. Locally it’s stored as “pg100.txt”.

In DataWrangling.ipynb, we use each play’s year as the separator to divide the complete txt file into individual txt file. So in each play’s txt file, the first line is the year the play was written, and everything after and including string “THE END” is discarded. We also removed all blank single lines. After obtaining each play, we further divide each play into different acts and store them into a dictionary key’ed by act titles. For each act, it will have tokenized words, non-stemmed words (which are also cleaned by removing stop words and punctuations), sentences, full text and filename being stored. The dictionary is stored in the works.json file. Once we have all the basic building blocks of the text. The next step is to extract text-mining features.

# Part II Feature Construction

We constructed features of TFIDF, n-grams and additional features based on the papers of Stanko et al. (2013) [[1]](#footnote-1) and Selman & Husagic-Selman (2011)[[2]](#footnote-2) using both sklearn and NTLK. The features used in our study are listed in Table 1. Features 4~13 are ***normalized by the total number of words*** in an act. Therefore, more precisely, they are ***frequencies***.

Table List of features

|  |  |
| --- | --- |
| 1. TFIDF 2. 3~5-gram normalized by TFIDF 3. Average length of sentences in the text 4. Hapax Legomena  (number of words that occur exactly once) 5. Dis Legomena  (Number of words that occur exactly twice) 6. Number of unique words | 1. Number of nominative pronouns 2. Number of conjunctions 3. Number of commas 4. Number of periods 5. Number of “a”, “an”, “the” 6. Number of “in”, “on”, “to”, “of” 7. Number of “is”, “are”, “was”, “were” |

Figure 1 (needs to be updated based on new data) shows the scatter matrix plot of all features excluding TFIDF and n-grams. According to Figure 1, Hapax Legomena, Dis Legomena and number of unique words are strongly correlated, while others features are relatively independent.

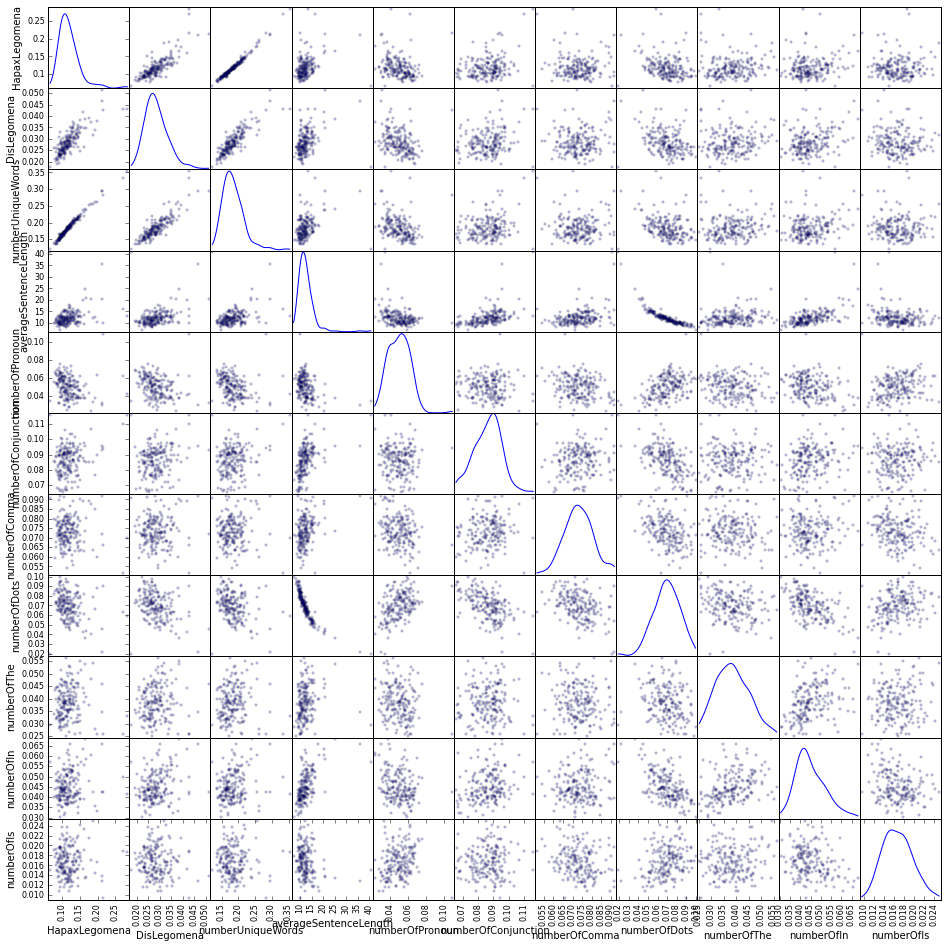


Figure Scatter matrix plot of 11 features

Do we need to normalize/standardize the feature matrix for clustering and PCA? After some research, the answer is yes. In this study, the average sentence length ranges from 10 to 30, while other features normalized by TFIDF or by total number of words are way below 0.1. As a result, if we run clustering based on Euclidean distance, the feature of sentence length dominates. Acts with long sentences become outliers and other features hardly play any role. Figure 2 shows covariance matrix of all features, non-standardized versus standardized. If the features are not standardized, the variance of average sentence length is on a different scale compared with other features. After standardization, the covariance matrix makes much more sense, as shown in Figure 2(b).

We also need to standardize the features when using PCA. Figure 3 shows the explained variance ratio of PCA components, non-standardized versus standardized. If the features are not standardized, the first PCA component, which is dominated by average sentence length, accounts for over 95% variance. By contrast, PCA after feature standardization makes more sense.

In this study, the features are standardized to , a zero mean and unit standard deviation.

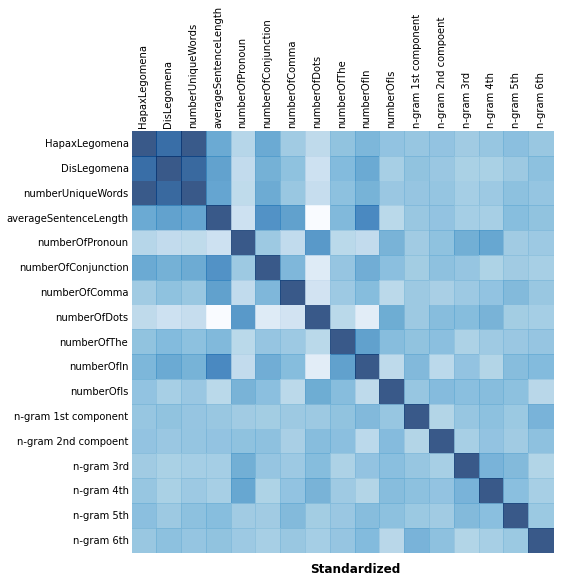
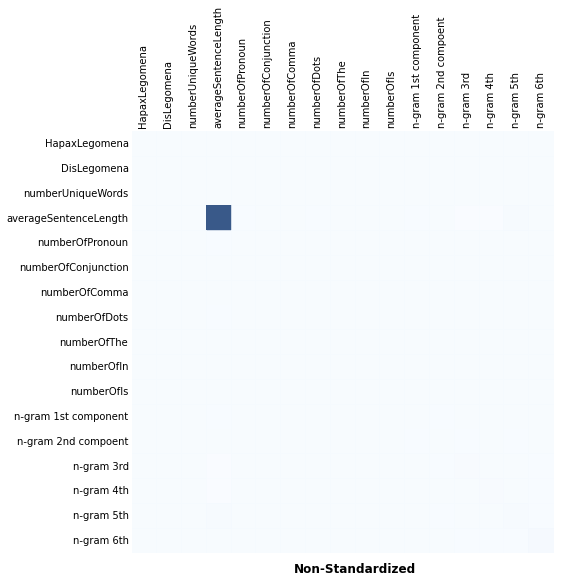


Figure Covariance matrix of all features (a) Non-Standardized features (b) Standardized features

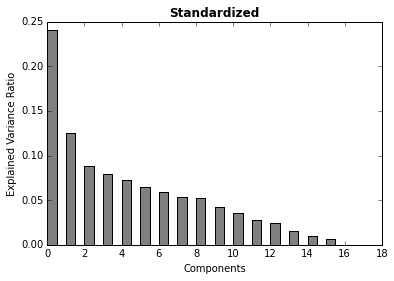
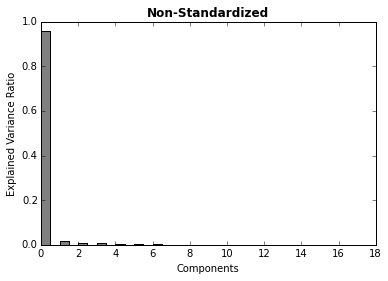


Figure Explained variance ratio of PCA components

# Part III Clustering Experiments

3.1 Clustering Methods

3.2 Feature Optimization

3.3 Parameter Optimization

3.4 Comparison of different methods

Bin Yan: Could you explain getTfidf() and getSVD() a little bit?

We use K-means as one of the clustering techniques. In our K-means, we use k-means++ to initialize the clusters, the max iterations is set to be 500, number of time the k-means algorithm will be run is 5.

We use Affinity Propagation as one of the clustering techniques. In our Affinity Propagation, the maximum iterations is set to be 1000.

# Part IV Discussion and Conclusion

Visualization: PCA vs. MDS

We use PCA reduction and 2D plot to draw the distribution and labeling of the clusters.

Features:

Clustering Methods:

4.2 Do any of the techniques give consistent results?

4.3 What did you conclude about the authorship?

1. Stanko, S., Lu, D., & Hsu, I. (2013). Whose Book is it Anyway? Using Machine Learning to Identify the Author of Unknown Texts. Machine Learning Final Projects. [↑](#footnote-ref-1)
2. Selman, S., & Husagic-Selman, A. (2011). Multilayered feedforward neural networks as a tool for distinction of the authors of texts. In Information, Communication and Automation Technologies (ICAT), 2011 XXIII International Symposium on (pp. 1-6). IEEE. [↑](#footnote-ref-2)