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

Bin Gao, Bin Yan

# Part I: How you approached the problem

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 “t8.shakespeare.txt”. In Shakespeare.ipynb, we use each play’s year as the separator to divide the complete txt file into txt files from 1.txt to 38.txt. So in each play’s txt file, the first line is the year the play was written. In Generate json-v1025.ipynb, 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\_v1028.json file. Once we have all the basic building blocks of the text. The next step is to extract text-mining features.

# Part II: What text mining features you tried

We selected 10 features based on theory foundations mentioned from two papers[[1]](#footnote-1)[[2]](#footnote-2): 1. Hapax Legomena, number of words that occur exactly once; 2. Dis Legomena, Number of words that occur exactly twice; 3. Total number of unique words; 4. Average length of sentences in the text; 5. Number of nominative pronouns; 6. Number of conjunctions; 7. Number of commas; 8. Number of periods; 9. Number of “a”, “an”, “the”; 10. Number of “in”, “on”, “to”, “of”; 11. Number of “is”, “are”, “was”, “were”.

These 10 features are mostly constructed by using the NTLK package. On top of it, we also constructed n-grams by using sklearn.

# Part III: What clustering approaches you used and show the parameter optimization

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: What visualization methods you found most helpful

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

# Part V: Do any of the techniques give consistent results?

To be added.

# Part VI: What did you conclude about the authorship?

1. Stanko, S., Lu, D., Hsu, I. Whose Book is it Anyway? Using Machine Learning to Identify the Author of Unknown Texts [↑](#footnote-ref-1)
2. Selman, S., Husagic-Selman, A. Multilayered Feedforward Neural Networks as a Tool for Distinction of the Authors of Texts [↑](#footnote-ref-2)