

will the real shakespeare please stand up?



September 16, 2018

ist 585

LUECK

## Abstract:

While most Shakespeare scholars have little doubt as to the authorship of William Shakespeare’s works, a minority view posits that either Sir Francis Bacon or Edward de Vere are the real authors. Implementing several data mining tools, this paper attempts to examine whether the works attributed to William Shakespeare have any overlap with the other two writers.

## Introduction:

Most Shakespeare scholars belong to what’s known as the Stratfordian school of thought which says that William Shakespeare wrote all his own poems and plays. However, since the mid-nineteenth century, the Baconian theory has argued that Sir Francis Bacon, a distinguished philosopher and scientist wrote all of Shakespeare’s poems and plays under a pen name. Stemming from this idea, a third school of thought came about in the early 1900s. This new idea, known as the Oxfordian theory, thinks that Edward de Vere, the Earl of Oxford during Shakespeare’s lifetime, was the original author. Both Baconians and Oxfordians have similar arguments as to why William Shakespeare couldn’t have written his plays and poems. The main argument is that William Shakespeare’s family didn’t have enough money to provide him with the education necessary to have a full comprehension of the subjects that appear in his works. The proof of these theories relies on ties to similarities in language and phrases used by either Sir Francis Bacon and Edward de Vere as compared to William Shakespeare.

Stratfordians such as Dr. Patrick Cheney of Penn State University dismiss these ideas as conspiracy theories (Stevenson). Although there isn’t much about followers of the Baconian theory anymore, plenty of respected scholars and writers alike cling to the idea that William Shakespeare was not the original author of his works. One major supporter of the Oxfordian theory, The Shakespeare Oxford Fellowship, is a non-profit organization formed in 2014 dedicated to “investigating the Shakespeare authorship question and disseminating the evidence that Edward de Vere, the 17th Earl of Oxford (1550 – 1604), is the true author of the poems and plays published under the pseudonym ‘William Shakespeare’(Shakespeare Oxford Fellowship Board of Trustees).” This organization publishes several newsletters and journals as well as maintains a website dedicated to this debate.

## Method:

This paper explores all three theories using decision trees (DT), k-means, k-nearest neighbor (KNN), and support vector machine (SVM) to view the similarities and differences between the three writers. To normalize the amount of writing per author, only the first 100 sonnets by William Shakespeare are included. The included sonnets were found at gutenberg.org. Sir Francis Bacon’s 19 essays were downloaded from gutenberg.org as well. For Edward de Vere, the data set includes 20 poems from shakespeareoxfordfellowship.org and 15 letters collected from and oxford-shakespeare.com.

Once all works were collected, any English stop words such as “the” or “a”, and Elizabethan stop words such as “thou” or “thy” were removed. Some of the Elizabethan stop words were compiled by Bryan Bumgardner and were copied from his website http://bryanbumgardner.com. Below is a full record of the additional stop words:

art, can, doth, dost, ere, hast, hath, hence, hither, nigh, oft, should’st, thither, tither, thee, thou, thine, thy, tis, t’was, wast, whence, wherefore, whereto, will, withal, would’st, ye, yon, yonder, lordship, come, hamlet, cause, also, use, said, see, other, say, exeunt, ham, man, men, shall

After removing all stop words and cleaning the data, the texts were put into a corpus for data analysis.

## Data Exploration:

Upon completion of the preprocessing stage, each author’s word count was examined to minimize any bias toward one author. The word count per author and per document are as follows:

|  |  |
| --- | --- |
| Author | Word Count |
| Shakespeare | 5438 |
| Bacon | 5348 |
| De Vere | 4556 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 22 | 52 | 58 | 98.98 | 107 | 508 |

The maximum in the above line is an outlier. However, that outlier represents Shakespeare’s play Julius Caesar. Without this outlier, the data would be solely based on Shakespeare’s Sonnets

In addition to viewing the word count, most reviews of these theories emphasize the similarity between words and phrases used by either Bacon or de Vere, and Shakespeare. The top five words across the various collected words were:

**“good”, “love”, “make”, “time”, “yet”**

The following word cloud demonstrates the seventy most frequently used words found in the collected works of these three authors.



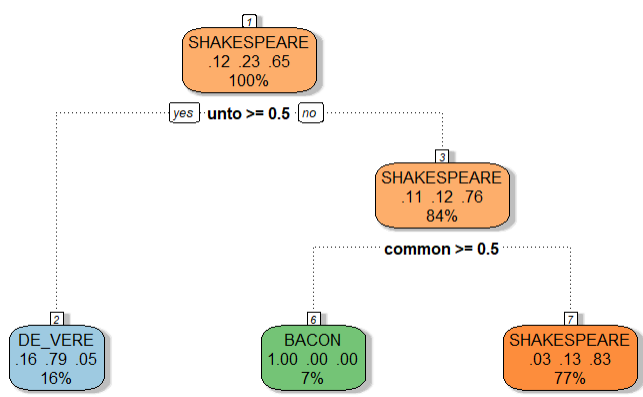
Additionally, the following table illustrates the top five most common word associations for the top three words along with how frequently these words are paired:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Time** | | **Love** | | **Good** | |
| rejected | 0.65 | absurdly | 0.68 | otherwise | 0.68 |
| course | 0.60 | admit | 0.68 | thing | 0.61 |
| continuance | 0.58 | alteri | 0.68 | citizen | 0.60 |
| cutting | 0.54 | amorous | 0.68 | things | 0.60 |
| speeches | 0.52 | antonius | 0.68 | affect | 0.55 |

A more comprehensive list of frequently used words can be found by running the code against the included data set. The code appendix also includes a method for viewing the number of words per document.

## Decision Tree Method:

Statisticians and data scientists have typically used a Bayesian approach to identify authorship in documents for natural language processing. One of the most famous instances of one of these methods in use to identify an author can be found in Inference and Disputed Authorship by Frederick Mosteller and David L. Wallace wherein they used Bayesian statistics to identify that Madison authored the disputed Federalist Papers written under the pseudonym “Publius” (Nerbonne 6). This study was able to predict that James Madison was the author of the disputed documents based heavily on his use, or lack thereof, of the word “upon” (Nerbonne 7). The following model attempts to identify authorship based on the word selection per document.



This method does identify the word “unto” as a word frequently used by both Shakespeare and de Vere, and the word “common” as frequently used by Shakespeare and Bacon. However, the probability of choosing the correct author per document using this model only returns the low accuracy level of ~70% with a confidence interval (CI) of between ~52% and ~83%. The Kappa prediction rate is much lower than any of those figures with an accuracy level of ~29% leading to the belief that this model may be just slightly better than randomly guessing which of the three authors wrote each document. The two tables listed below compare Shakespeare with each author individually to test the CI on a smaller scale.

## k-Means Method:

The k-means method was integrated in two ways. The first attempt includes three centers to try to account for the words used by each author. The second way attempts to divide the documents by author also using three centers. Both instances of the k-means test allow for a maximum of 50 iterations with 25 starting points. They also normalize the data by incorporating the tf-idf method.

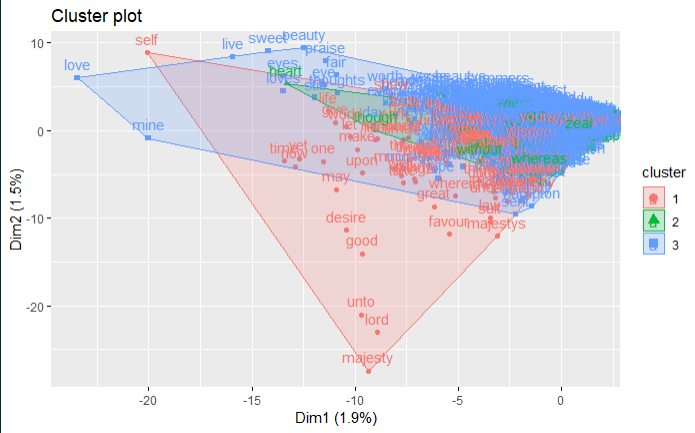


Figure 1 k-means analysis words per author

The above figure shows data from Shakespeare in blue, de Vere in red, and Bacon in green. Attempting to separate the words by author shows that, while very few words stick out, most of the vocabulary found in the corpus overlaps significantly for all three authors.

Additionally, the below figure illustrates that there’s lots of overlap in most of the documents for each author with only a couple of Bacon’s works standing out.

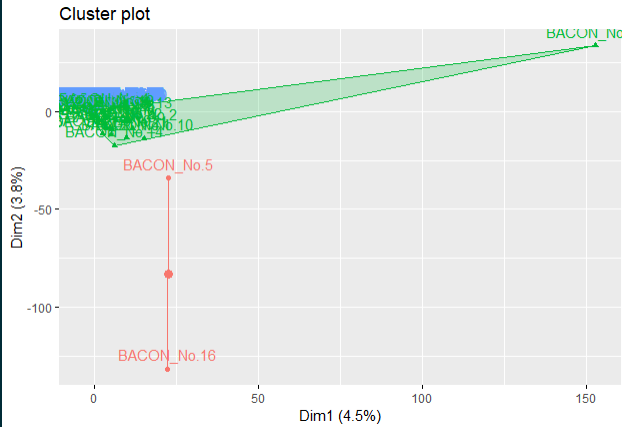
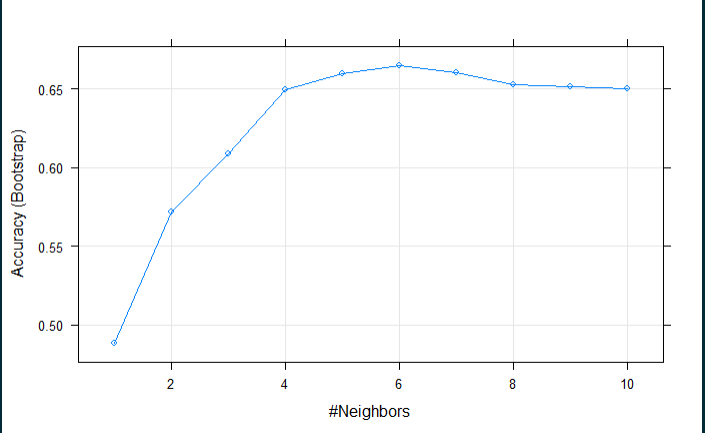


Figure 2 k-means- author per document

## k-Nearest Neighbor (KNN) Method:

Typically, if the k-means model has difficulty in separating data the k-nearest neighbor is often used to split the data further by grouping each data point to its nearest points as specified by the user. Using KNN on this data set results in a reading of ~64% accuracy with a confidence interval between ~47% and ~79%. For the highest accuracy, the model recommends incorporating 6 neighbors when using the bootstrap method according to the plotted data below:



## Support Vector Machine (SVM) Method:

Support vector machines can work with both regression and categorical problems. In the case of categorical data, the support vector machine will try to find a hyperplane of separation between different categories and exploit that point of separation to categorize the data. Since this method is considered an assisted learning algorithm, the user provides some guidance as to how to separate the data. In this model, a cost of 25 was assigned to the model which produces about 100 support vectors on average. The outcome has an average accuracy of ~92% with a confidence interval between ~76% and ~97%. The kappa value is ~79%.

## Conclusion:

While most models don’t find enough information in the given documents to provide a solid confidence interval to debunk any of the given theories, a well-tuned support vector machine provides the most support for Stratfordians. More investigation into how well these authors correlate with Shakespeare’s plays, and further research into how their word associations match from a contextual point of view would help to either prove or disprove these theories. However, given the time constraints and scope of this project, those data will have to be explored at another time.

## References:

Nerbonne, John. “The Exact Analysis of Text.” semanticscholar.org. https://pdfs.semanticscholar.org/4c15/7f110dfd1d90f9931205003719f4c9856d01.pdf

Shakespeare Oxford Fellowship Board of Trustees. “Mission Statement.” Shakespeare Oxford Fellowship. https://shakespeareoxfordfellowship.org/about-us/mission-statement/

Stevenson, Alexa. “Probing Question: Did Shakespeare really write all those plays?” Penn State News. https://news.psu.edu/story/141315/2008/12/01/research/probing-question-did-shakespeare-really-write-all-those-plays (Accessed 8/24/2018)