IST707 – Data Analytics

Old Writings and New Tactics: Analyzing Shakespeare with Text Mining

Ally Peyton

awpeyton@syr.edu

Abstract

This project is an experiment in digital humanities, one of my first forays into the field. For the paper, I will analyze the Elizabethan playwright William Shakespeare’s works using text mining in R. The analysis will provide a deeper understanding of two facets of Shakespeare’s works: why they are categorized the way they are, and whether the playwright authored more texts than are usually attributed to him. Using three types of classification, this paper will explore the categorization of two highly disputed plays by Shakespeare, *The Tempest* and *Pericles, Prince of Tyre,* finding that based on word frequency of the playwright’s other works, these two plays should be considered comedies as well. In addition to this analysis, this paper will also present a word-frequency-based analysis of two works that have previously disputed authorship: *Arden of Faversham* and the *Additional Passages of the Spanish Tragedy, 1602.* Both works have been said to be written by either William Shakespeare or his contemporary Thomas Kyd. This analysis is inconclusive, as both Shakespeare and Kyd were returned as possible authors of *Arden,* but it does point to Shakespeare as the likely author of the *Additional Passages.*

Old Writings and New Tactics: Analyzing Shakespeare with Text Mining

As a student of the liberal arts and of data science, I have always been fascinated by the field of digital humanities, and for this project, I decided to take a first foray into examining Shakespeare’s works using text mining. William Shakespeare is one of the best-known playwrights of all time, famous for works such as *Romeo and Juliet* and *Hamlet.* A consensus exists among modern Shakespeare scholars that there are thirty-eight works that can be definitively attributed to Shakespeare, either because of historical documents proving authorship or new analysis providing certainty on the topic. However, as with any historical figure, there are disputes surrounding various aspects of Shakespeare’s work. One is the authorship dispute, which this paper will examine through a narrow lens, looking at whether Shakespeare or his contemporary and possible collaborator Thomas Kyd authored two works, *Arden of Faversham* and the *Additional Passages of the Spanish Tragedy, 1602.* The second dispute surrounds categorization and purpose of his works: some modern scholars argue that certain works should be categorized certain ways based on their contents, others believe that the historical records of performance and context are more important. This analysis will focus on the categorization of *The Tempest* and *Pericles, Prince of Tyre,* two works which could be considered tragedies or comedies based on the view one takes.

Literature Review

Shakespeare’s body of works have been exhaustively analyzed in the past, and with the recent rise of digital humanities and programming-driven solutions to difficult problems, many researchers have undertaken efforts to analyze his works with technological solutions. For example, Nicholas Ruta at Harvard College performed an analysis on Shakespeare’s body of work to determine whether portions of the works attributed to the playwright may have been authored by students or contemporaries of his (Ruta, n.d., p.1). Ruta’s analysis focused on Thomas Middleton’s possible contributions to Shakespeare’s work and used k-means clustering to analyze his input, coming to the conclusion that based on word frequency, it is likely that Thomas Middleton may have had a part in writing thirteen of the works commonly contributed to Shakespeare (Ruta, n.d., p. 4). Ruta’s analysis used clustering techniques to look at Middleton and Shakespeare’s writings, showing word frequency as a valid indicator of potential authorship (Ruta, n.d., p. 3). Similarly, Eric Kammers, a data scientist from the University of Washington, used clustering to perform a genre analysis of Shakespeare’s thirty-eight known works. Kammers finds that when using additional genres such as “late romances” and “problem plays” (eg. plays with no specific genre), *The Tempest, The Winter’s Tale, Cymbeline, and Pericles, Prince of Tyre* can all be categorized as late romances as opposed to comedies (Kammers, 2017). He also notes that this analysis displays a relationship between romance plays by Shakespeare, especially the later ones, and comedies by the playwright – they share many similar words and with further analysis may share similar textual features (Kammers, 2017). Finally, as each of these prior analyses relies on clustering for their conclusions, I also researched the use of different text mining techniques for the digital humanities, finding a paper by Matthew G. Kirschenbaum at the University of Maryland most useful. Kirschenbaum examines the use of classification techniques such as SVM models and Naïve Bayes models within the digital humanities, concluding that while the results of classification on non-standard data such as written texts are not as certain as the conclusions drawn from data with known ground truths, classification still has value as a system which can indicate possible new truths for nontraditional data (Kirschenbaum, 2007, p. 5). The prior literature around Shakespeare, text mining, and the digital humanities indicates that there is space to attempt to analyze Shakespeare’s body of works using other text mining techniques.

Research Questions or Hypothesis

In this paper, I will use classification methods to attempt to determine new insights on William Shakespeare’s body of work. There will be two paths of inquiry in this research: first, looking at the categorization of known problem plays, and second, attempting to determine whether Shakespeare authored two works with disputed authorship.

*Hypothesis 1*: Shakespeare’s plays *The Tempest* and *Pericles, Prince of Tyre* will be categorized as comedies due to their similarities to other comedies by the playwright.

*Hypothesis 2:* The disputed works *Arden of Faversham* and the *Additional Passages of the Spanish Tragedy, 1602* will be attributed to William Shakespeare due to their similarities to his other works.

Method

This analysis was performed using RStudio to create machine learning algorithms for classification of Shakespeare’s works. The data used originates from several different sources. The main dataset, which contains the complete works of Shakespeare sourced from Project Gutenberg, is from an R package called “bardr” developed by Zane Billings. Other data were gathered from several online sources: a list of Shakespeare’s characters, used to clean the character names out of the text corpus, was sourced from PlayShakespeare.com. The text of the disputed authorship plays *Arden of Faversham* was retrieved from an Oxford repository as text files and manipulated into usable data. The full text of the *Spanish Tragedy,* used as the sample text of Thomas Kyd’s writing, and including the *Additional Passages of the Spanish Tragedy, 1602* were sourced from Project Gutenberg via a Wordpress site and also manipulated into usable data, including splitting out the additional passages from the full text. Each data source was loaded into an R file and prepared for text mining using R’s tm and dplyr packages. Some wrangling was necessary for each datasource, including reshaping the data to be useful in a corpus of words. Once the corpuses were built, sparse document-term matrices containing the corpuses were created and used for Naïve Bayes model generation using the e1071 package. Following the creation of Naïve Bayes models, the data were again transformed into dataframes to be used for kSVM models and decision trees, which were then built using the kernlab and C50 packages, respectively. Analysis of the results was performed using gmodels for crosstabling.

Results

The data used contains, in all, several hundred thousand observations for both Shakespeare and Kyd’s works. After wrangling, two dataframes – one named plays\_genre and one named authors\_plays were developed, containing 38 observations of 3 variables and 29 observations of 4 variables, respectively. The structures of the original data and the manipulated datasets are shown in Tables 1 through 3.

Table 1. Original data

|  |  |  |  |
| --- | --- | --- | --- |
| name | content | full\_name | genre |
| All's Well that Ends Well | "ALLS WELL THAT ENDS WELL" | All's Well that Ends Well | Comedy |
| All's Well that Ends Well | “Contents” | All's Well that Ends Well | Comedy |
| All's Well that Ends Well | “Scene I.” | All's Well that Ends Well | Comedy |

*Continues for 134,767 more rows*

Table 2. Manipulated data – plays\_genre & authors\_plays

|  |  |  |
| --- | --- | --- |
| name | genre | content |
| All's Well that Ends Well | Comedy | *full text of All’s Well* |
| Antony and Cleopatra | Tragedy | *full text of Antony and Cleopatra* |
| As You Like It | Comedy | *full text of As You Like It* |

*Continues for 35 more rows, authors\_plays has the same structure but replaces the genre column with an author column containing the authors’ last names*

The analysis of the genres of *Pericles, Prince of Tyre* and *The Tempest* resulted in the plays being categorized as comedies by each method – Naïve Bayes, SVM models, and decision trees. The Naïve Bayes model and SVM model both immediately categorized the plays as comedies. The decision tree model did as well, on the basis of the presence of certain words such as king and prince.

The analysis of the authorship of *Arden of Faversham* and the *Additional Passages of the Spanish Tragedy, 1602* was more complex. The Naïve Bayes model and the decision tree model predicted that both works were written by Shakespeare, based on the presence of words that frequently reoccurred in both plays. The SVM model, however, predicted that both works were written by Thomas Kyd.

Discussion

The analysis produced for this paper used classification to determine the genre of two plays by Shakespeare and the authorship of two disputed works. The rules used by the decision tree in the case of genre classification were quite interesting. Rules the tree used include that if the word king occurred less than 90 times in a play, but the word prince occurred more than 6 times, then 75% of the time it could be considered a comedy, and that if the word prince occurred less than 6 times, the play is a history. This is interesting in that a king and a prince co-occurring as characters in a play about royalty is highly likely, but in historical plays, the word king can be used frequently without a prince being a relevant character as well.

This analysis shows that the digital humanities can be a useful tool to apply to historical texts as a precursor to further analysis. While the results produced by this analysis are certainly interesting, further analysis would be useful. For example, further analysis of separate facets of Shakespearean comedies could provide more information about the categorization of *Pericles* and *The Tempest.*

References

Billings, Z. (2021, March 24). Package 'bardr'. Retrieved May 16, 2021, from https://cran.r-project.org/web/packages/bardr/bardr.pdf

Kammers, E. (2018, December 09). Text mining Shakespeare's First Folio. Retrieved May 16, 2021, from https://mydatasci.wordpress.com/2017/12/11/text-mining-shakespeares-first-folio/

Kirschenbaum, M. G. (n.d.). The Remaking of Reading: Data Mining and the Digital Humanities. Retrieved from https://d1wqtxts1xzle7.cloudfront.net/55515713/MKirschenbaum.pdf?1515728276=&response-content-disposition=inline%3B+filename%3DThe\_remaking\_of\_reading\_Data\_mining\_and.pdf&Expires=1621190606&Signature=dpam2NaVleglMPV9cOLVbult8HHY1F2P-zwGAsAJjEaD3RM~THU2E5d1BNk2OTKDS2cr67aO8BFj1YCMWv~8A4BnegBpd~9U45bVXnV9XagWGXW76BbpoOSbuqlT44naGmSejoEmhCebld0lmL4KRHlqyoNye-VFx6hgAACVRKi6kePvs~VuN4HpRDGWzFt183ZKQuuq2i8xaO-K1~zJ1hxtnXq5aYlTg0DbmW1DX3OkEhomCV0PKTmE9FZT5MbjV5VDubIRIAc3nW6qJfhbprlweUpk5jVweuaRWSzqWxvN46WSBvwprA8q4DeGC-0kRhXxnTL~H9Zi3K01EdNYDw\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Ruta, N. J. (n.d.). Text Mining for the Analysis of Shakespeare. Retrieved May 16, 2021, from http://nickruta.com/shakespeare.pdf

Shakespeare, W. (n.d.). Complete List of Shakespeare's Characters. Retrieved May 16, 2021, from https://www.playshakespeare.com/study/complete-shakespeare-character-list

Spanish Tragedy Full Text. (n.d.). Retrieved from https://sourcetext.files.wordpress.com/2018/01/spanish\_tragedy\_flues.pdf-

Thomas Kyd. (n.d.). Retrieved May 16, 2021, from https://www.britannica.com/biography/Thomas-Kyd

Unknown. (1592, January 01). Arden of Faversham Full Text. Retrieved May 16, 2021, from https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/3001

William Shakespeare BIOGRAPHY. (n.d.). Retrieved May 16, 2021, from https://www.shakespeare.org.uk/explore-shakespeare/shakespedia/william-shakespeare/william-shakespeare-biography/