Final report

2022-10-28

Introduction

Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works, as well as to “encourage the creation and distribution of eBooks.” It was founded in 1971 by American writer Michael S. Hart and is the oldest digital library. Most of the items in its collection are the full texts of books or individual stories in the public domain. All files can be accessed for free under an open format layout, available on almost any computer. As of 3 October 2015, Project Gutenberg had reached 50,000 items in its collection of free eBooks.

The releases are available in plain text as well as other formats, such as HTML, PDF, EPUB, MOBI, and Plucker wherever possible. Most releases are in the English language, but many non-English works are also available. There are multiple affiliated projects that provide additional content, including region- and language-specific works. Project Gutenberg is closely affiliated with Distributed Proofreaders, an Internet-based community for proofreading scanned texts.

In this project, I used the gutenbergr package in R to download and process public domain works from the Project Gutenberg collection and tried to answer the following questions:

1. How is the word usage different among different authors?
2. Is it possible to train a model that can predict the author of a work based on its word frequencies?
3. If the answer to question 2 is “yes,” what method should we use to achieve a better performance?

Data Exploration

To answer the above questions, we first explore the data a little bit by generating some summary statistics and plots.

Works and Authors

Figure 1 shows the distribution of languages used by the works in Project Gutenberg, where the number of books is mapped in a log scale. As we can see from the histogram, English is the dominant language, with over 10,000 works in total. The second and third most frequently used languages are French and German, respectively.

Figure 2 shows the distributions of the birth years, death years, and ages of the authors. Most of the authors were active from 1850-1950 and lived to the age of 70 to 80 years. In addition, the distributions of these three characteristics are all left skewed, implying that a few authors are from ancient times, far away from today.

Figure 3 shows the relationship between the books and subjects. Most subjects correspond to only 1 book, i.e., most books only have 1 subject. There are some subjects corresponding to 2-4 books, but subjects corresponding to more than five books are quite rare.

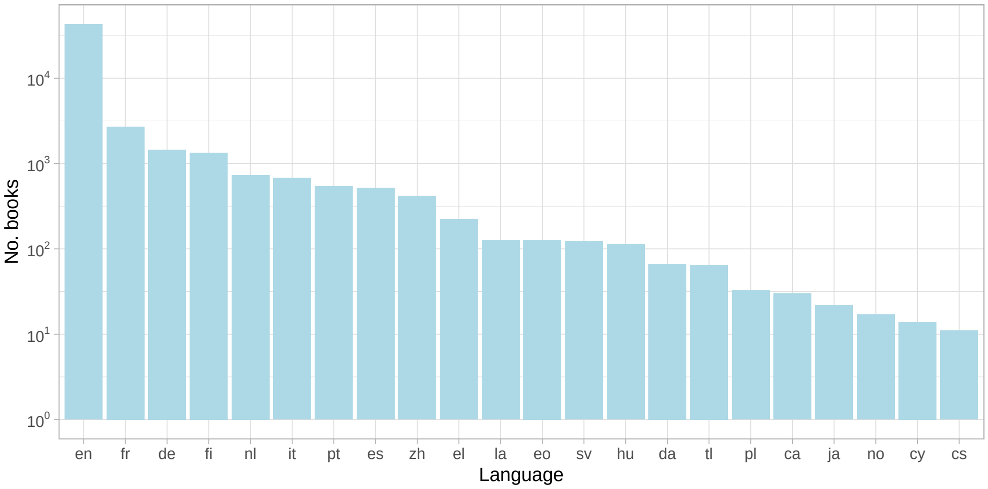

Figure 1: Language of the works in Project Gutenberg

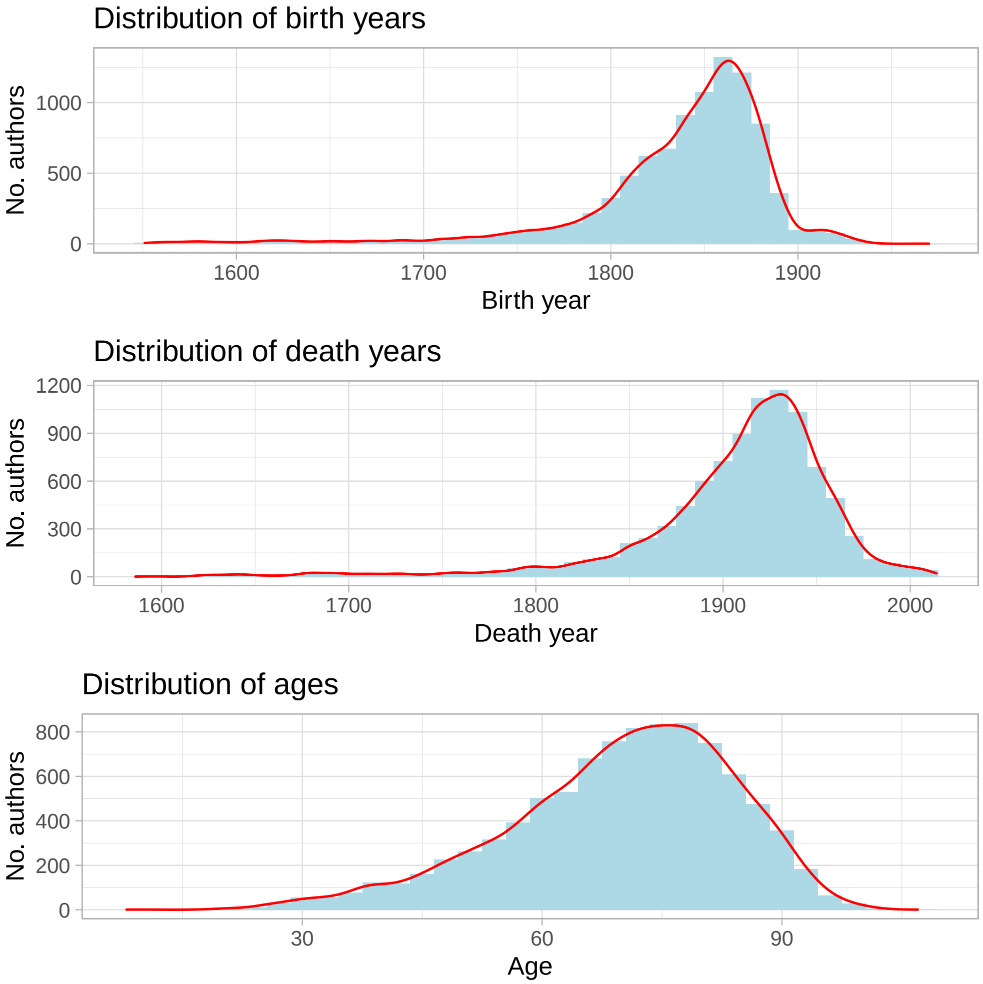

Figure 2: Birth years, death years and ages

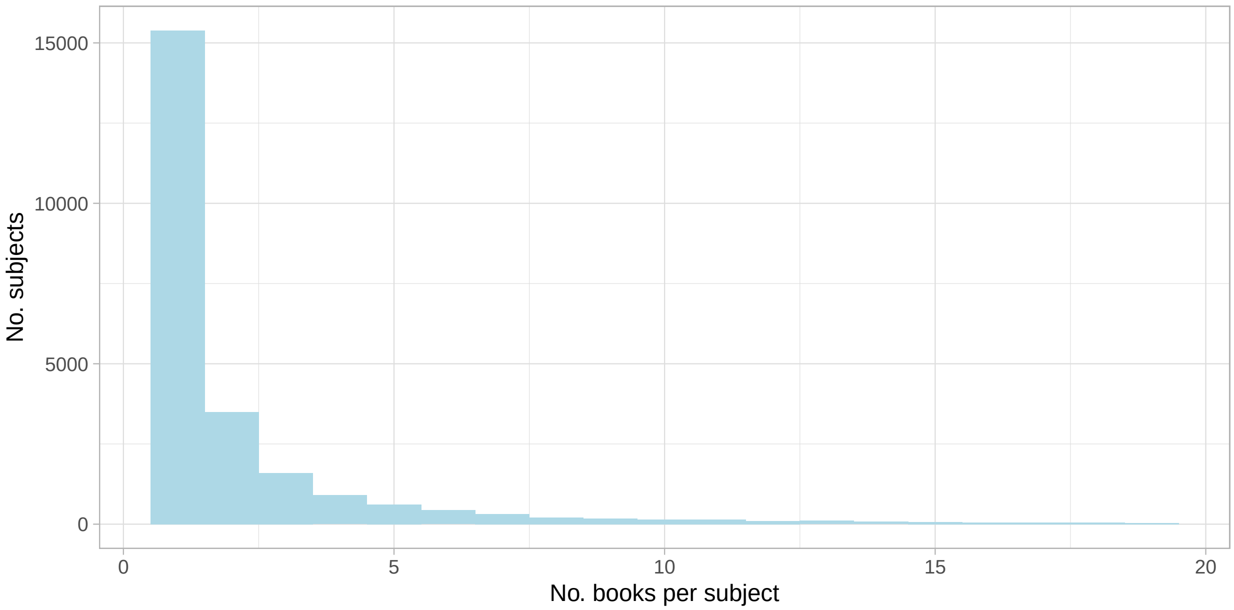

Figure 3: Books and subjects

Figure 4 shows the top 20 authors with the most works. Lytton is the top 1 with more than 200 books. Mark Twain, William Shakespeare, and Charles Dickens are three well-known authors with many works, which we will focus on in the sequel analysis.

Word Clouds

To visualize the word usage of different authors, we generated word cloud graphs for Mark Twain, William Shakespeare and Charles Dickens, and displayed the top 100 most frequently used words in their works after removing the meaningless stopping words. We also prepared a rectangular data set that recorded the frequencies of words that occurred in the three authors' works for later analysis.

Mark Twain

Samuel Langhorne Clemens (November 30, 1835 – April 21, 1910), known by his pen name Mark Twain, was an American writer, humorist, entrepreneur, publisher, and lecturer. He was praised as the "greatest humorist the United States has produced", and William Faulkner called him "the father of American literature". His novels include The Adventures of Tom Sawyer (1876) and its sequel, Adventures of Huckleberry Finn (1884), the latter of which has often been called the "Great American Novel". Twain also wrote A Connecticut Yankee in King Arthur's Court (1889) and Pudd'nhead Wilson (1894), and co-wrote The Gilded Age: A Tale of Today (1873) with Charles Dudley Warner. Figure 5 show the top most frequently used words in Mark Twain's works.

William Shakespeare

William Shakespeare (bapt. 26 April 1564 – 23 April 1616) was an English playwright, poet and actor. He is widely regarded as the greatest writer in the English language and the world's pre-eminent dramatist. He is often called England's national poet and the "Bard of Avon" (or simply "the Bard"). His extant works, including collaborations, consist of some 39 plays, 154 sonnets, three long narrative poems, and a few other verses, some of uncertain authorship. His plays have been translated into every major living language and

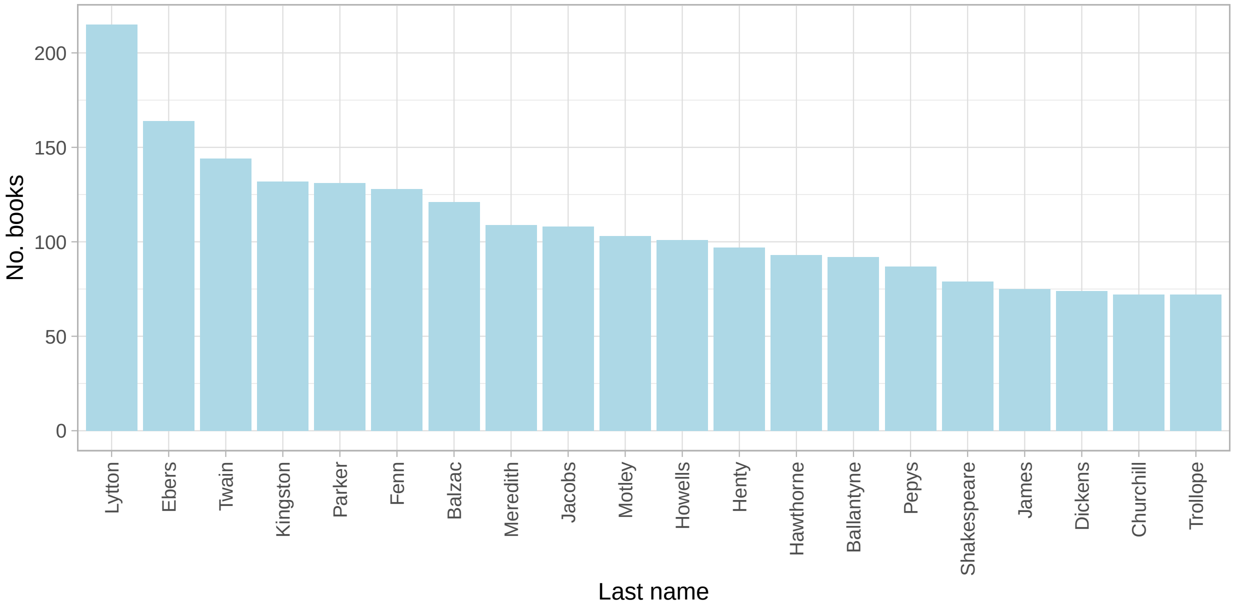

Figure 4: Top 20 authos with the most works

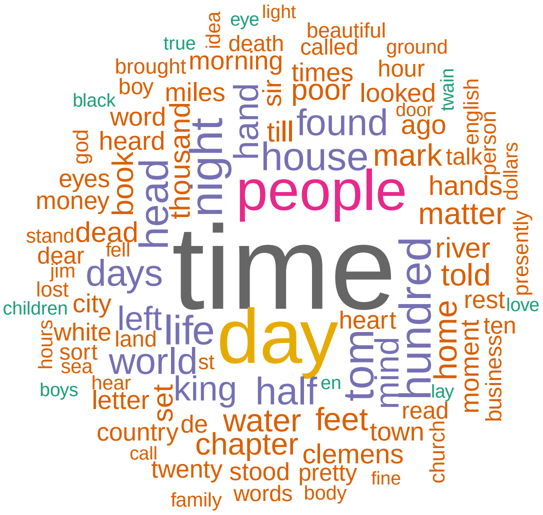

Figure 5: Mark Twain

are performed more often than those of any other playwright. He remains arguably the most influential writer in the English language, and his works continue to be studied and reinterpreted. Figure 6 show the top most frequently used words in William Shakespeare's works.

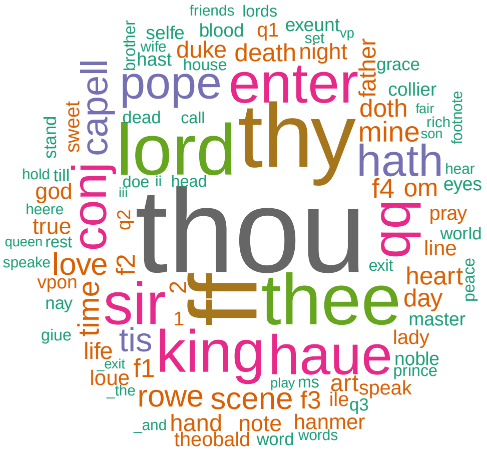

Figure 6: William Shakespeare

Charles Dickens

Charles John Huffam Dickens (7 February 1812 – 9 June 1870) was an English writer and social critic. He created some of the world's best-known fictional characters and is regarded by many as the greatest novelist of the Victorian era. His works enjoyed unprecedented popularity during his lifetime and, by the 20th century, critics and scholars had recognized him as a literary genius. His novels and short stories are widely read today. Figure 7 show the top most frequently used words in Charles Dickens' works.

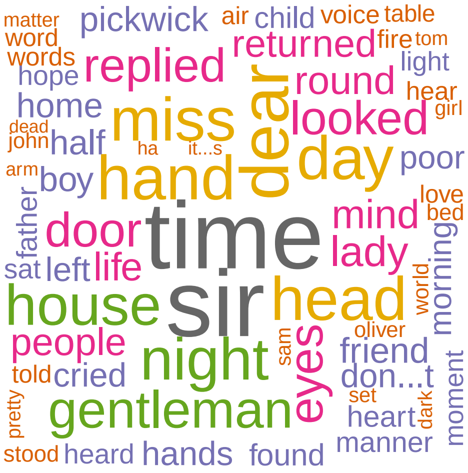

Figure 7: Charles Dickens

Dimension Reduction

Since the date we prepared is very high-dimensional with more than 10,000 words (variables) and only 290 works (observations), we need to perform dimension reduction via PCA or tSNE for later analysis. We first perform tSNE to check whether the works of the same author will be clustered. This will give us some idea whether it is possible to train a machine learning model to predict the author.

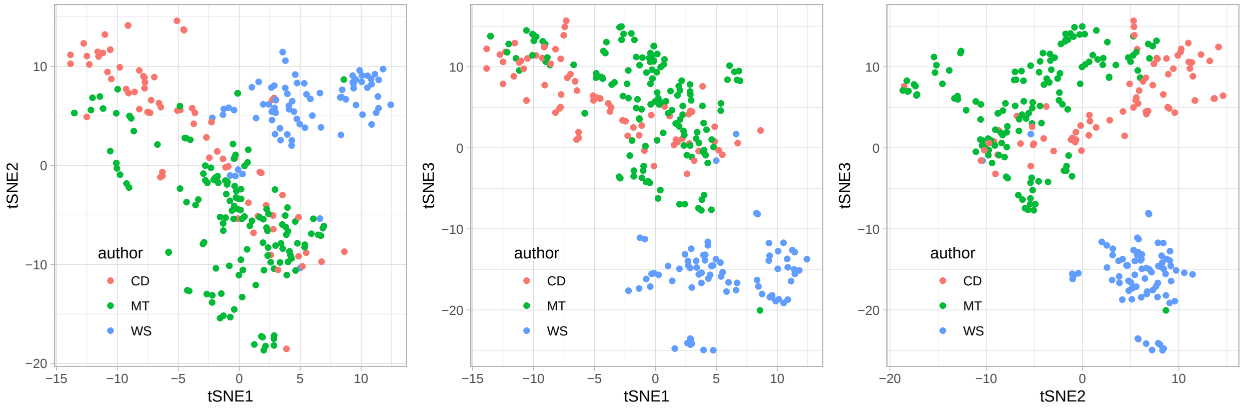

Figure 8: tSNE dimension reduction

Figure 8 displays the results of tSNE, reducing the dimension to 3. As we can see, the works of the three authors have a natural tendency to cluster together. It is especially obvious for the works of William Shakespeare, due to the old English he used. The works of Mark Twain and Charles Dickens are more close to each other because they lived in times closer to each other. The results of tSNE suggest that it is hopeful to build a machine learning algorithm to predict the authos based on the pattern of word usage.

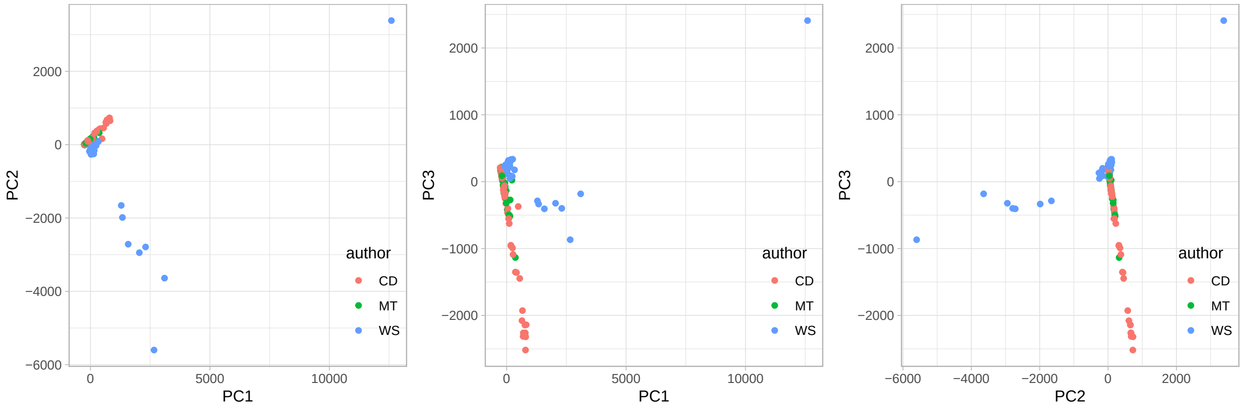

Figure 9: PCA dimension redcution

Next, we used PCA to perform dimension reduction and the results are displayed in Figure 9. Since tSNE can capture the local structure of the data while PCA captures the global structure, the visual performance of PCA is not as good as tSNE. However, PCA can provide much more dimension components than tSNE, which is the base of building a classification machine learning method. In our case, PCA can provide up to 290 principle components while tSNE can only provide up to 3.

Classification

Before implementing machine learning algorithms, we need to pick the number of PCs that are included as predictors in the data. The most easy way is to check the cumulative proportion of variance explained by the PCs. As shown in Figure 10, 30 PCs will account for more 95% variance of the data, so it is reasonable to use up to 30 PCs as predictors.

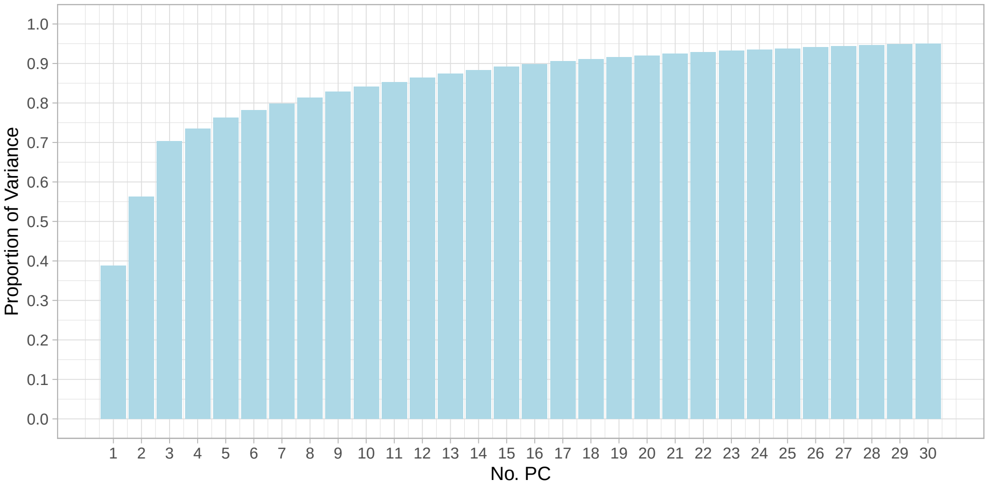

Figure 10: Cumulative proportion of variance explained by PCs

In the sequel, we will used generalized boosted model and support vector machine to build two classification models and compare their performance.

Generalized Boosted Models (GBM)

Generalized boosted models (GBM) are a combination of two techniques: decision tree algorithms and boosting methods. Generalized Boosting Models repeatedly fit many decision trees to improve the accuracy of the model. For each new tree in the model, a random subset of all the data is selected using the boosting method. For each new tree in the model the input data are weighted in such a way that data that was poorly modeled by previous trees has a higher probability of being selected in the new tree. This means that after the first tree is fitted the model will take into account the error in the prediction of that tree to fit the next tree, and so on. By taking into account the fit of previous trees that are built, the model continuously tries to improve its accuracy. This sequential approach is unique to boosting.

Figure 11 shows the prediction accuracy of GBM using different numbers of PCs as predictors. For a fixed number of PCs, we used 4/5 of the data to train the model and 1/5 of the data for testing. Repeat this procedure 100 times and the average prediction accuracy is reported. As we can see, the prediction accuracy increases as the number of PCs included in the data increases. With 30 PCs, the prediction accuracy is more than 90%.

Figure 12 shows the relative influence of each PC. Only 9 PCs play a relatively important role in prediction, which is consistent with the fact that the prediction accuracy achieves around 87.5% when the number of PCs is 9.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM

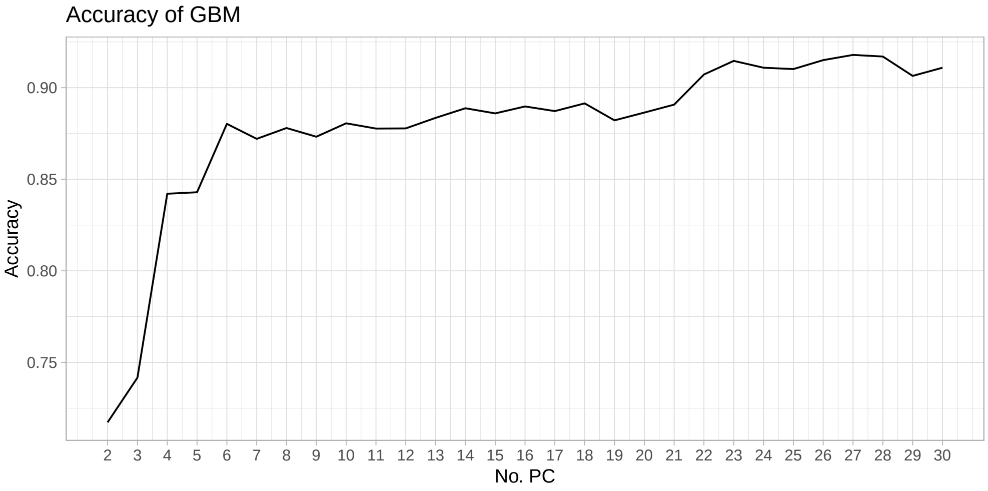

Figure 11: Prediction accuracy of generalized boosted models

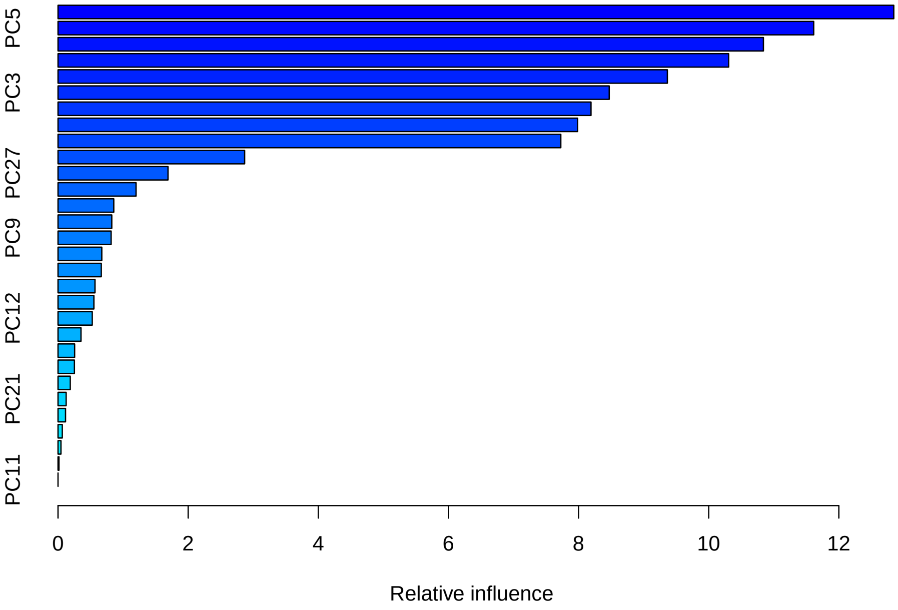

Figure 12: Relative influence of each PC

algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Figure 13 shows the prediction accuracy of SVM using different numbers of PCs as predictors. For a fixed number of PCs, we used 4/5 of the data to train the model and 1/5 of the data for testing. Repeat this procedure 100 times and the average prediction accuracy is reported. Different from GBM, the increment in the prediction accuracy is not significant when the number of PCs increases: it is always around 75%.

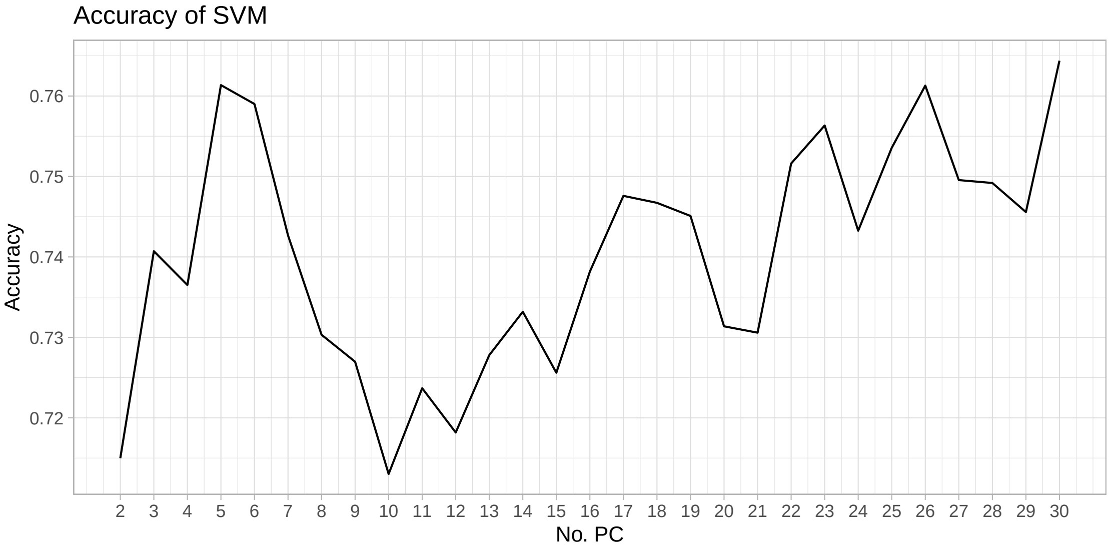

Figure 13: Prediction accuracy of support vector machine

We put the two accuracy curves together in Figure 14 for comparison. As we can see, the performance of GBM is uniformly better than SVM.

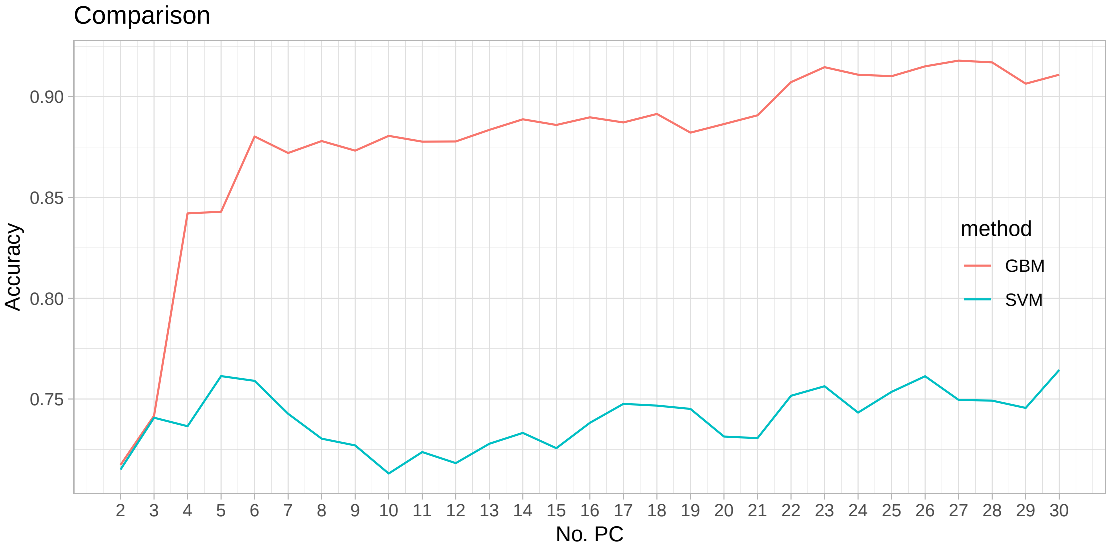

Figure 14: Comparison of GBM and SVM

Conclusion

Now we are able to answer the three questions we raised at the beginning. We visualized the difference of the word usage in different authors by generating the word cloud graphs. We performed tSNE dimension reduction to visualize the structures of the works from the same author and the results revealed a promising possibility of building a machine learning classification algorithm to predict the author. Finally, we compared the performance of GBM and SVM on the task of predicting authors based on the words used in the works. The GBM is uniformly better than SVM.

There are several limitations of this project. First, we only considered three authors, which is a small number. If the number of authors increases, the performance of the classification algorithm is expected the be worse. It is worthwhile to include all the authors in Project Gutenberg and perform the same analysis. Second, we only compared two classification methods: GBM and SVM. Comparing more methods will provide a more comprehensive understanding of the performance of machine learning methods on this special classification task.