Assessment 5

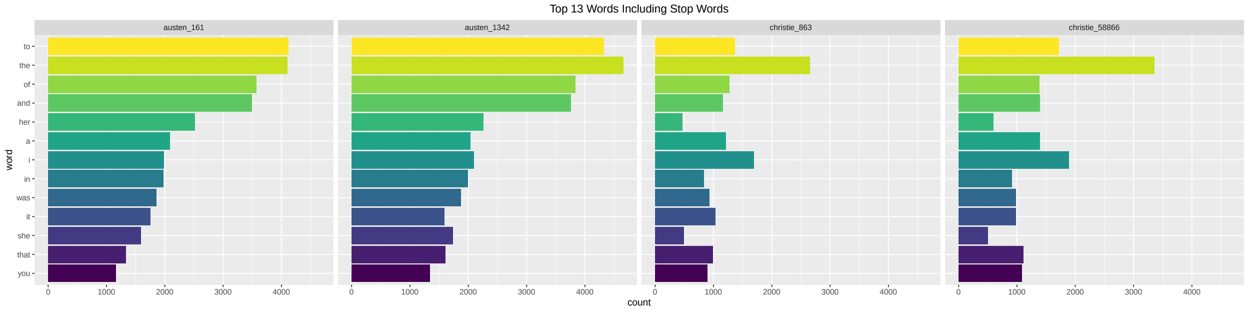
Joana Levtcheva, CID 01252821

Initial EDA

The EDA will be structured around two authors Jane Austen and Agatha Christie for two of their books, respectively Sense and Sensibility, Pride and Prejudice, and The Mysterious Affair at Styles, The Murder on the Links.

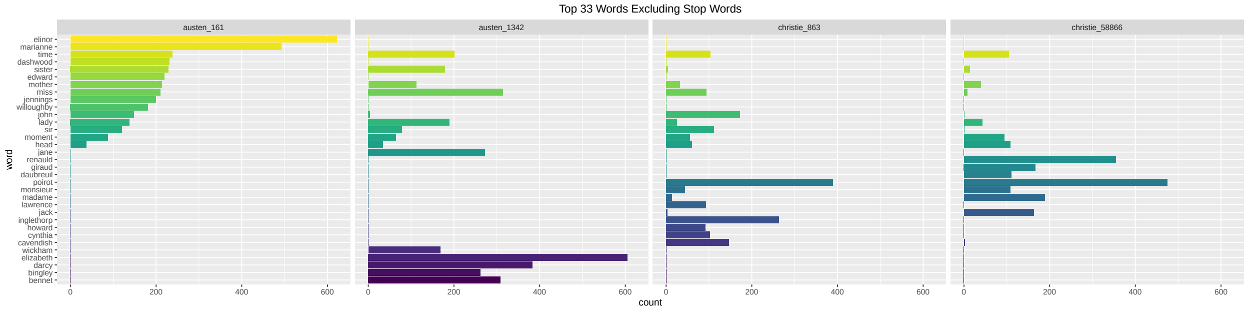
First, for each of the 4 books we are going to select the top 10 words based on their count in the book. (This produces almost the same results as when we are doing tf-idf). After that, we are going to make a union of the top 10 words for each book and that will define the set of top words fot the chosen books. We are going to do this analysis both with and without removing the stop words from the books.

For the case without removing the stop words, we get 13 top words among the books:



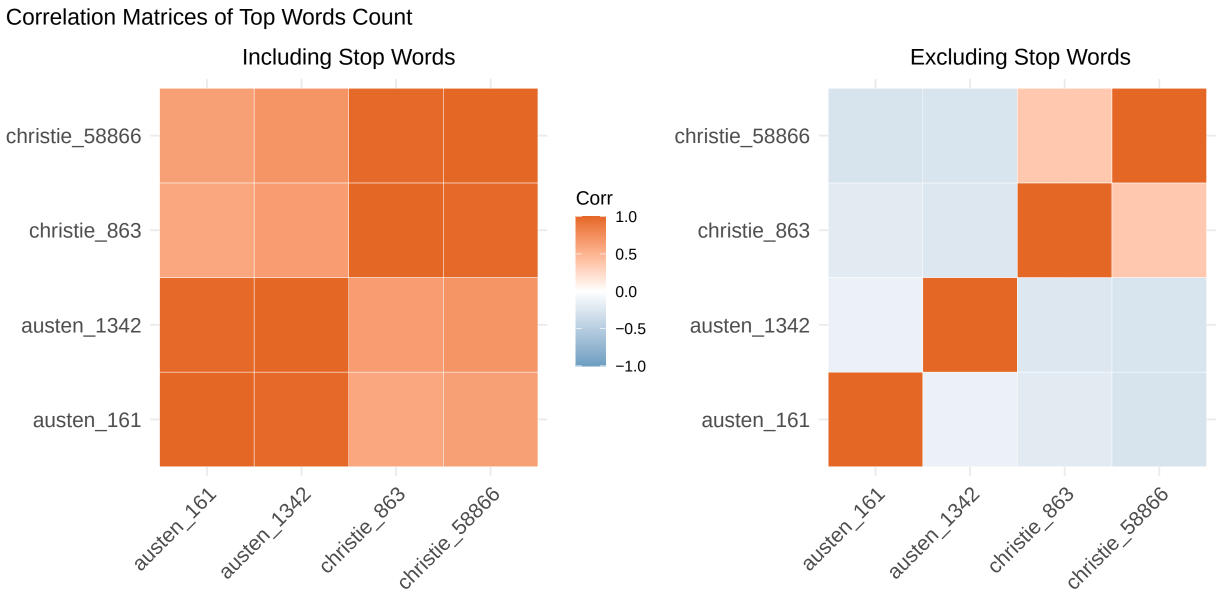
We can notice that between the two chosen books of Jane Austen and also of Agatha Christie there is a similar distribution of the top words. But at the same time we can see the difference in the words distribution between the two authors.

For the case with removing the stop words, we get 33 top words among the books:



Here, we can observe that there are similarities in the distributions of Agatha Christie books, whereas the words distributions in Jane Austen books are not similar. The most common words from the top words in Pride and Prejudice (austen_1342) don't even exist in her other book Sense and Sensibility (austen_161). This leaves us with the thought that we might not be able to determine that these two books have the same author.

To extend this observation, we are going to plot the correlation matrices for both of the cases for the counts of the top words in each book.



For the case with stop words, we can see that both of the authors' book are highly correlated with one another, and less correlated with the books from the other author. This leads to thinking that we might be able to define if two books are from the same author, as well as to identify if two books are from different authors.

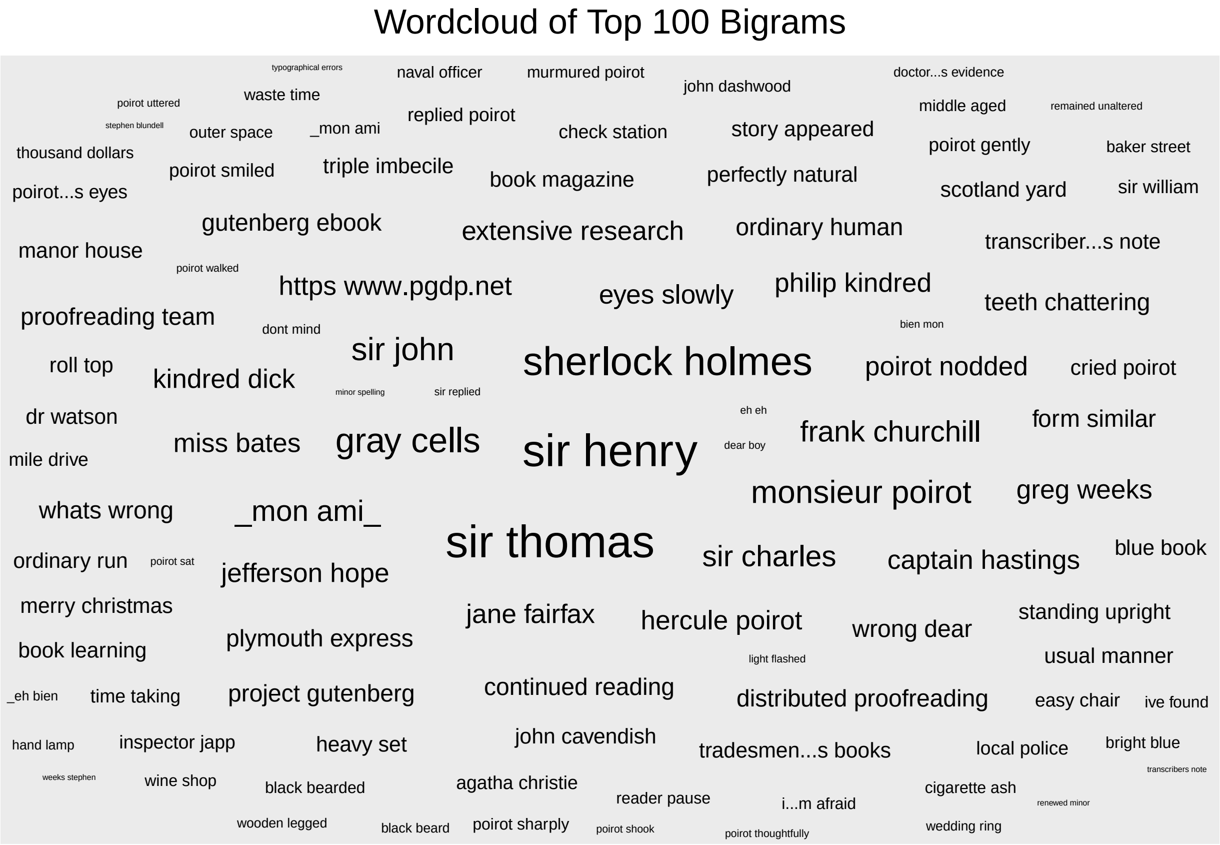
For the case without stop words, there is no correlation between Jane Austen books, as for Agatha Christie there is a little correlation. This leaves us with the conclusion that distinguishing authors in this case really depends on the type of books they write. If there are books based in the same series with the same characters as it is in Agatha Christie's case, then we might be abel to do it. But in the general case we do not have that guarantee so it might be useful to not exclude stop words in future analysis.

N-grams

We are going to do an analysis of the top 100 bigrams of all of the 26 books of the 6 authors. When determining the bigrams we have removed the bigrams with first or second word being in the list of stop words. We are defining the top 100 bigrams following the process:

- applying TF-IDF for the bigrams in every book
- grouping by bigram and for every bigram group taking the sum of the tf-idf values calculated for each book
- removing the bigrams which exist only in one document because they are of no value in assessing the similarity of language used across the corpus
- selecting the top 100 bigrams of the remaining bigrams based on the sum of their tf-idf values, this includes bigrams with repetaing sum of tf-idf values, hence we get 104 bigrams

Wordcloud of the chosen top bigrams:

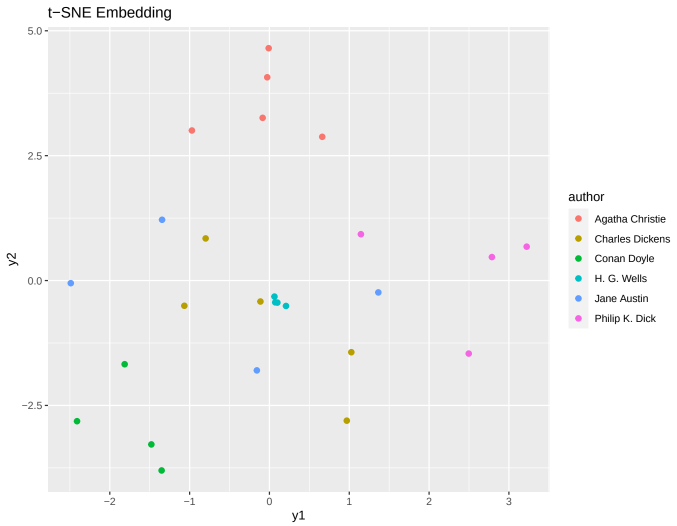


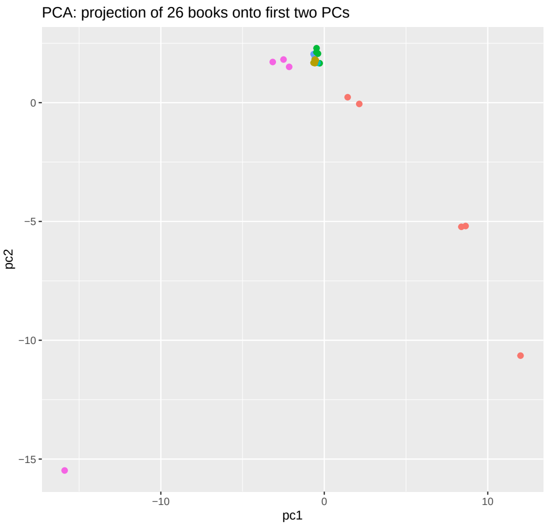
It is noticeable that the most popular bigrams are actually names. This means that they will probably be popular for only one book or if they are popular for more books this means that these books are probably part of one seris of an author. Thus, we can expect similarity between the bigram vectors (containing tf-idf bigram values) for the different books of one author. Especially if the books are part of a series, which is the case with Agatha Christie.

After that, for every book of the chosen top books we define the tf-idf for every value, and if a bigram doesn't exist in a book, we set the value to 0.

Dimensionality Reduction

Now, we are constructing a matrix with 26 rows consisting of the books, and 104 columns consisting of the chosen top bigrams, and values the previously defined tf-idf values of every bigram for every book. Performing PCA with 2 components and t-SNE leads to the following results:



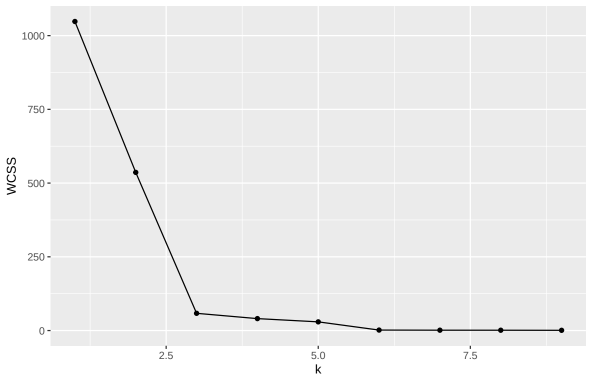


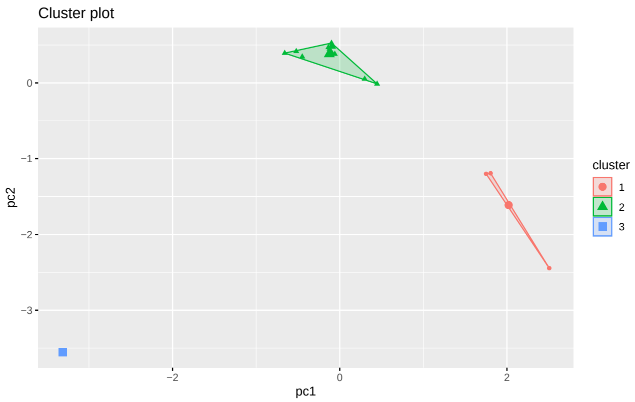
Grouping the books by author specified with distinct colours, we can state that t-SNE produces a relatively good visualisation of the books based on author. There are distinct groups for the books of the authors Sir Arthur Conan Doyle, Agatha Christie, Philip K. Dick, H. G. Wells. There is a bit of overlapping for Jane Austen and Charles Dickens, also noticing that the H. G. Wells group is really close to these two groups, or even inside of Jane Austin's group.

The PCA result shows a distinct group for Philip K. Dick (although with high inner group variation) and Agatha Christie. The other authors are all in one small group with really low inner group variation.

Clustering

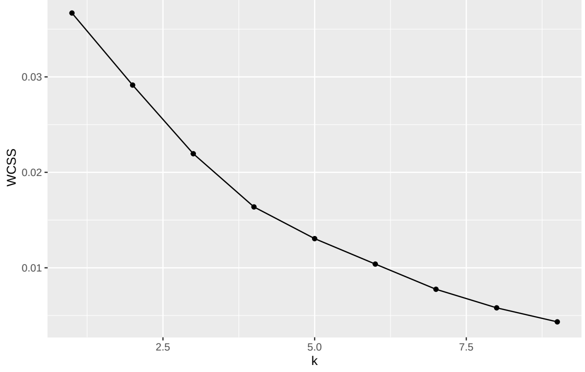
An attempt is made in applying k-means clustering after PCA with 2 components was applied. Experimenting with different number of clusters from 1 to 9 the elbow rule pointed to using 3 clusters.

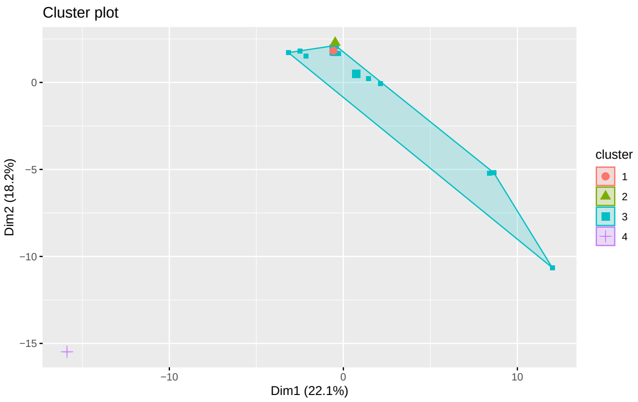




Given the bad PCA group distribution it is no surprise that the k-means clustering doesn't perform well as well. Almost all of the books are in one cluster, there is a cluster containing one book (The Eyes Have It by Philip K. Dick), and also a cluster with 3 books (all of them by Agatha Christie - The Plymouth Express Affair, The Hunter's Lodge Case, The Missing Will).

This time, let's apply k-means directly to the previosly defined matrix with 26 rows and 104 columns. Experimenting with different number of clusters from 1 to 9 the elbow rule pointed to using 4 clusters.





Again, most of the books seem to be in one cluster. There are three clusters with one single book Mansfield Park by Jane Austen, The Hound of the Baskervilles by Sir Arthur Conan Doyle, and The Eyes Have It by Philip K. Dick, where the last two books are not significantly distinct from the biggest cluster area.

In this case we experimented with TF-IDF of bigrams, but theoretically we could have chosen any value of n for our n-grams. By increasing the n we increase the number of n-grams and we decrease the number of books in wich each will appear.

Conclusion

Despite the fact that most examples of working with natural language texts exclude stop words, we actually showed that there could be value in using them. Because they are so popular working with them creates dense matrices that have many common features between different books. But their histograms can differ based on the writing style, which as the correlation matrix showed could be used for differentiating our authors. Other practical examples of that could be to find the real author of literature with an unknown one.

On the contrary, when working with word bigrams, especially after excluding stop words, we get very sparase matrices. This requires some carefulness when working with them, which in our case was removing bigrams specific only to single book in order to reduce sparseness. The plot produced from the t-SNE algorithm shows that there is some meaningful information in the book feature vectors we generated, because there are meaningful clusters for some authors. However, the way to use this information is not so obvious as both PCA and k-means failed to capture it properly and produced unusable results.