Dobre Bogdan-Mihai

Authorship Attribution

In this document I will present a way to do authorship attribution using hierarchical clustering. The algorithm is based on the approach presented by Liviu P. Dinu and Marius Popescu in the paper "Ordinal measures in authorship identification".
The goal is to use hierarchical clustering in such a way that texts written by the same author are clustered together. For that we would need to extract some stylistic features from each text that we can assume are different enough from author to author that texts written by the same author will most likely be closer to each other than texts written by different authors. In the approach used, function words rankings are used as stylistic features. Function words are a set of words that we choose and that we expect to appear in each text. The function words chosen in this case are those which have the biggest frequency in all texts. For every text, the apparitions of each function word is counted, and a ranking of function words is generated. These function word rankings tend to work well in authorship attribution because, while some words may be under the conscious control of the author and used differently between multiple literary works, function words tend to be the most common words in the language so they are very unlikely to be under the author's conscious control. Therefore, function words rankings tend to give good results in clustering algorithms intended to attribute authors to literary works such as novels. The optimal number of function words after many experiments turned out to be 13, so the 13 most frequent words among all texts were used as function words.
The hierarchical clustering algorithm sees each text as an initial small cluster and unites 2 clusters into a bigger one at each step. The 2 clusters chosen to form a bigger cluster are those with the smallest distance between them, and the distance between 2 clusters in this case is defined as the maximum distance between any pair of texts from the 2 clusters. At

the beginning of the clustering algorithm the smallest distances would be between texts with the same author.

The question that naturally rises is how to define a "distance" between 2 words rankings. There are many approaches discussed in the original paper: Pearson's correlation coefficient, Spearman's rank-order coefficient, Spearman's footrule, Goodman and Kruskal's gamma and Kendal's tau-b. After trying all the different distances on my set of literary works, I chose to use Kendal's tau-b as it gave the best results. Given 2 rankings, X and Y, Kendal's tau-b distance is defined as: (P - Q) / sqrt((P+Q+T) * (P+Q+U)) where P is the number of concordant pairs and Q the number of discordant pairs in the rankings. A pair (i,j) is said to be concordant if (X[i] - Y[i]) * (X[j] - Y[j]) > 0 and is said to be discordant if (X[i] - Y[i]) * (X[j] - Y[j]) < 0. T is the number of ties only in X, U is the number of ties only in Y and ties in both X and Y are not counted.

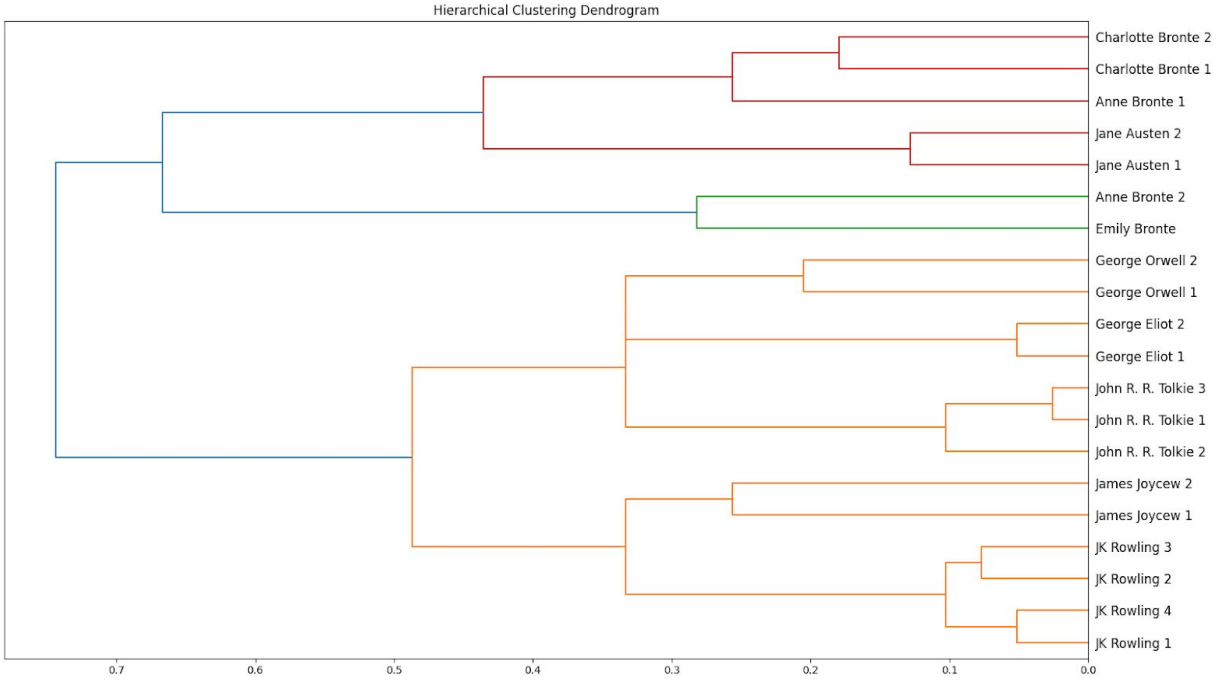
There were many literary works used in the experiment. Most of them were from classical literature, but there were also operas like "Harry Potter" and books from "The Lord of the Rings" lore. A total of 20 works were used. While the algorithm gave correct results in most cases, it seemed to struggle with differentiating between the Bronte sisters. The full dendrogram that the algorithm generated together with what exact literary works were used can be seen at the end of this document. The algorithm clustered together the 2 works of Charlotte Bronte, but it clusters one work of Anne Bronte with them and the other work of Anne Bronte with the "Wuthering Heights" of Emily Bronte. It should be mentioned, however, that one of Anne Bronte's works, "Agnes Gray", is among the shortest works used, so that might contribute to the error. Beyond that, it seems that the other works, both from classical literature and modern literature, are clustered together correctly. An article called ”Some Common Features in the Brontë Sisters’ Novels” published in the ”Bronte Studies” claims that there are similarities between Anne Bronte’s ”Agnes Gray” and Charlotte Bronte’s ”Jane Eyre” as well as between Anne Bronte’s ”The Tenant of Wildfell Hall” and Emily Bronte’s ”Wuthering Heights”. This corresponds to the results of the clustering algorithm.

Another interesting observation is that Jane Austen's works were clustered between Anne Bronte's works. A similarity between Anne Bronte’s ”Agnes

Grey” and Jane Austen’s works was noticed by George Moore. Another similarity between Anne Bronte’s ”The Tenant of Wildfell Hall” was noticed by ”The Examiner”, a leading intellectual journal of its time. This may confirm that the algorithm may cluster Jane Austen's works close to Anne Bronte's correctly.

In conclusion, using function words and hierarchical clustering as an approach in authorship attribution seems to be a very promising method which yields very good results.

Dendrogram:



Literary works used:

"Jane Austen 1" - Sense and Sensibility
"Jane Austen 2" - Pride and Prejudice
"George Orwell 1" - 1984
"George Orwell 2" - Homage to Catalonia
"James Joycew 1" - Dubliners
"James Joycew 2" - A Portrait of the Artist as a Young Man
"George Eliot 1" - Felix Holt, the Radical

"George Eliot 2" - Middlemarch
"Charlotte Bronte 1" - Jane Eyre
"Charlotte Bronte 2" - Shirley
"Emily Bronte" - Wuthering Heights
"Anne Bronte 1" - Agnes Gray
"Anne Bronte 2" - The Tenant of Wildfell Hall
"JK Rowling 1" - Harry Potter and the Philosopher's Stone
"JK Rowling 2" - Harry Potter and the Chamber of Secrets J. K. Rowling
"JK Rowling 3" - Harry Potter and the Prisoner of Azkaban J. K. Rowling
"JK Rowling 4" - Harry Potter and the Goblet of Fire
"John R. R. Tolkien 1" - The Two Towers
"John R. R. Tolkien 2" - The Fellowship of the Rings
"John R. R. Tolkien 3" - The Return of the King