Authorship Attribution

Dobre Bogdan-Mihai

January 4, 2021

Table of contents

Introduction

Introduction

- Authorship Attribution is an old problem and there are still many disputed literary works
- I will present an approach for authorship attribution using function words and hierarchical clustering based on the paper ”Ordinal measures in authorship identification” by Liviu P Dinu and Marius Popescu.
- The results of my own experiment on a set of literary works with this approach will also be presented.

Methods used

Hierarchical clustering

- Hierarchical clustering is an algorithm that separates a dataset into different clusters based on similarity.
- At the beginning every text (data point) is seen as a cluster. After that, at every step, 2 ”smaller clusters” are clustered together into a bigger one.
- The 2 clusters with the smallest distance between them are clustered together
- In this case, ”distance” between 2 clusters means the maximum distance between any pair of texts (data points) from the 2 clusters.

Stylistic features

- In order to differentiate between different literary works we need to extract stylistic features
- The features must be unique enough for each author and common among all the works written by the same author.
- Therefore, they shouldn't be under the psychological control of the author.

Function words

- Function words are some words that we expect to appear in every text and that we choose.
- The most frequent words among all texts are chosen.
- For every text, all the apparitions of function words are counted and a function word ranking is generated.
- Since the ranking for each text is numerical, the clustering algorithm can be used.

Distance between rankings

- The question that naturally rises is how to define an optimal ”distance” between 2 words rankings that is appropriate for this context.
- There are many approaches discussed in the original paper: Pearson’s correlation coefficient, Spearman’s rank-order coefficient, 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.

Kendall’s tau-b

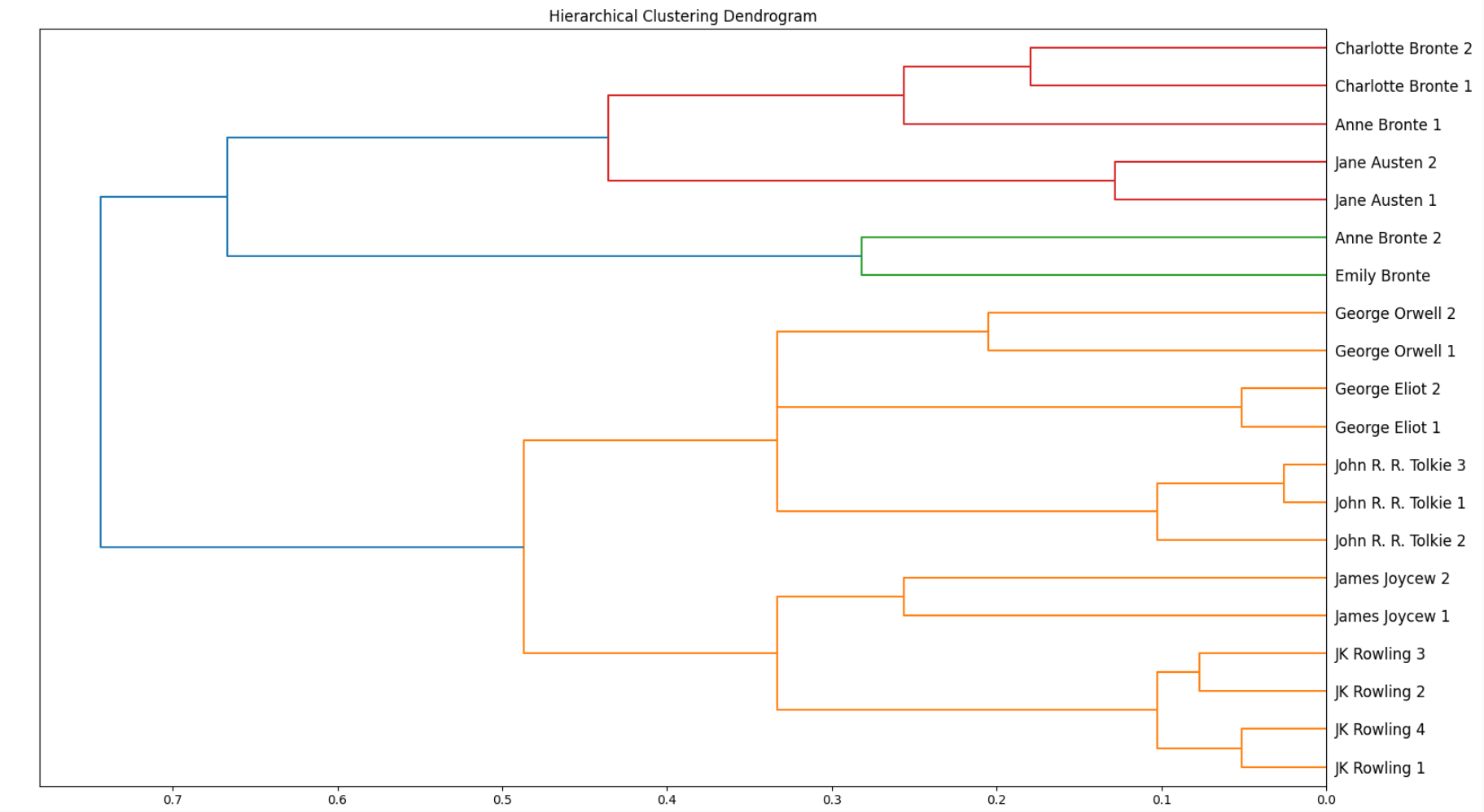
- Given 2 rankings, X and Y, Kendal’s tau-b distance is defined as: (P - Q) / sqrt((P+Q+T) * (P+Q+U)) where P is the number of concordant pairs and Q the number of discordant pairs in the rankings.
- A pair (i,j) is said to be concordant if (X[i] - Y[i]) * (X[j] - Y[j]) > 0 and is said to be discordant 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.

The experiment

The experiment

- 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.

The 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

Literary works used

- ”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

Literary works used

- ”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. Tolkie 1” - The Two Towers
- ”John R. R. Tolkie 2” - The Fellowship of the Rings
- ”John R. R. Tolkie 3” - The Return of the King

Results

- The algorithm seems to perform well on most literary works.
- Both classical and modern literary works are clustered together mostly correctly.
- A notable exception seems to be the Bronte sisters.

Results

- The algorithm clustered together the 2 works of Charlotte Bronte correctly.
- However, 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 that one of Anne Bronte’s works, ”Agnes Gray”, is among the shortest works used, so that might contribute to the error by not being as representative for the author’s style as the other works.

Conclusion

In conclusion, using function words and hierarchical clustering as an approach in authorship attribution seems to be a very promising method which, while not perfect, yields very good results.