Authorship Attribution

Dobre Bogdan-Mihai

January 7, 2021

Table of contents

1. Introduction

2. Methods used

3. The experiment

4. Results

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 data set 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

Distance between 2 clusters

- There are many ways to define distances between 2 clusters that contain multiple data points
- For example: distance between their closest data points, distance between their centers, distance between their farthest data points etc.
- In this case, ”the distance” between 2 clusters means the maximum distance between any pair of word rankings (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.
- They should ideally not be under the psychological control of the author.

Function words

- Function words are words that we expect to appear in every text and that we choose.
- Since they should not be willingly controlled by the author, it is optimal when the chosen functional words are just the most frequent words among all texts.
- This also ensures that they will have a high number of apparitions in each text.

Function words

- The most frequent 13 words are chosen as function words.
- For each 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

Different metrics

- The following 3 metrics are more commonly used but they do not yield goold resuts in our case
- Pearson’s correlation coefficient
- Spearman’s rank-order coefficient
- Spearman’s footrule

Different metrics

- The following 2 metrics yield the best results for our case as they are based on the idea of using the concordant and the discordant pairs to compute the distance for 2 rankings
- Goodman and Kruskal's gamma
- Kendal's tau-b.

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

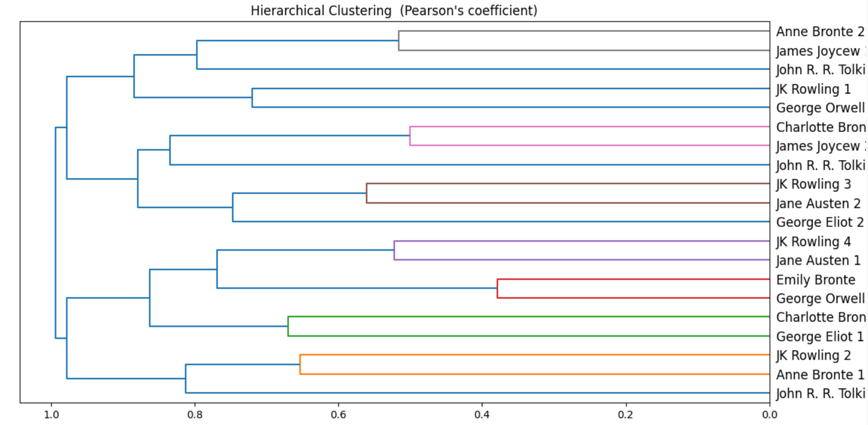
The experiment

- 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.

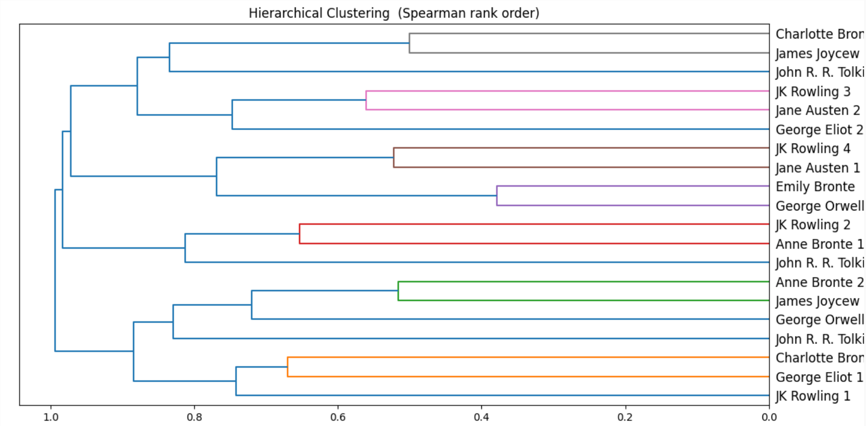
The experiment

- Distance metrics that were experimented with proved to have a huge impact on the results.
- Tau b is the only metric that provides very accurate results.
- All the other metrics produce results that are very inaccurate and seem almost random.

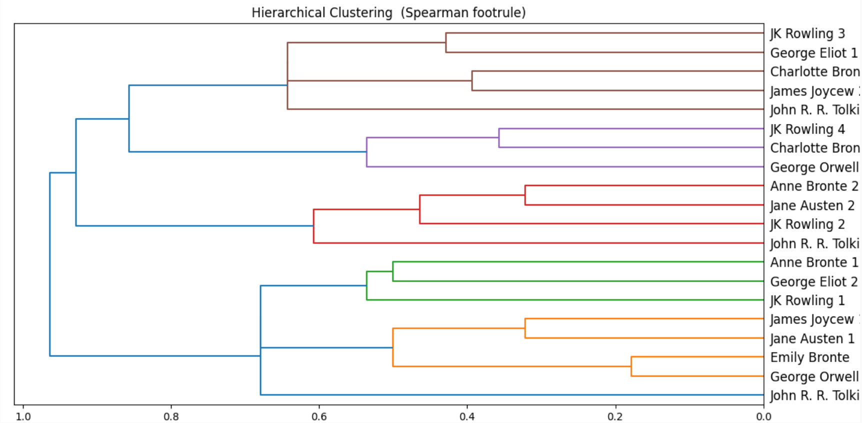
The dendrogram using Pearson’s correlation coefficient



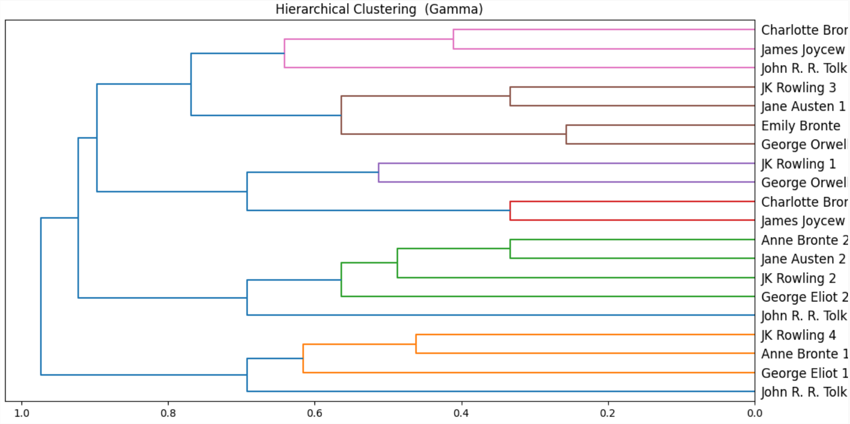
The dendrogram using Spearman's rank and order



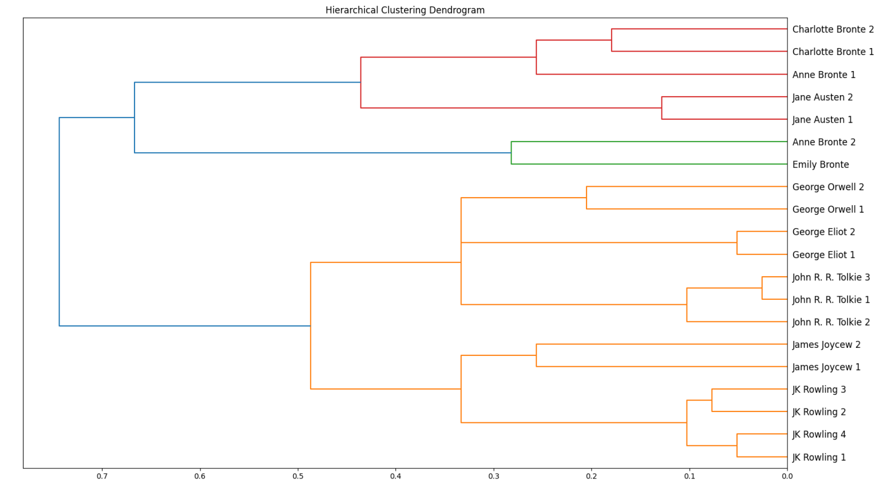
The dendrogram using Spearman's footrule



The dendrogram using the gamma metric



The dendrogram using Kendal’s tau b metric



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

Literary works used

- ”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

Literary works used

- ”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

- ”J.K. Rowling 1” - Harry Potter and the Philosopher’s Stone
- ”J.K. Rowling 2” - Harry Potter and the Chamber of Secrets J. K. Rowling
- ”J.K. Rowling 3” - Harry Potter and the Prisoner of Azkaban J. K. Rowling
- ”J.K. Rowling 4” - Harry Potter and the Goblet of Fire

Literary works used

- ”John R. R. Tolkien1” - The Two Towers
- ”John R. R. Tolkien 2” - The Fellowship of the Rings
- ”John R. R. Tolkien 3” - The Return of the King

Results

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.

The Bronte Sisters

- The algorithm clustered together the 2 works of Charlotte Bronte correctly.
- However, it clusters ”Agnes Gray” of Anne Bronte with them and ”The Tenant of Wildfell Hall” 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.

The Bronte sisters

- An article called ”Some Common Features in the Brontë Sisters' Novels” published in the ”Bronte Studies” provides some useful insights.
- It states that there are similarities between Anne Bronte's ”Agnes Gray” and Charlotte Bronte's ”Jane Eyre”.
- It also states that there are similarities between Anne Bronte's ”The Tenant of Wildfell Hall” and Emily Bronte's ”Wuthering Heights”
- That seems to accurately portray the result of the clustering.

Modern and classical authors

- J.K. Rowling and John R.R. Tolkie, the 2 modern authors, are very close to each other in the dendrogram.
- This suggests a bigger difference between author’s publishing dates might make some styles more distant and the other way around.

Jane Austen’s works

- Jane Austen’s works are clustered between Anne’s Bronte works.
- A similarity between Anne Bronte’s ”Agnes Grey” and Jane Austen’s works was noticed by George Moore.
- A similarity between Anne Bronte’s ”The Tenant of Wildfell Hall” and Jane Austen’s works was noticed by ”The Examiner”, a leading intellectual journal of its time.
- This suggests the algorithm is right to cluster Jane Austen’s works close to Anne Bronte’s.

The gender of the writers

- The female writers and male writers seem to be clustered separately.
- The only exception is J.K. Rowling, who wrote at a much later date than the other female authors.
- Even though she is not clustered with the other female authors, it is important to note that she is on ”the edge” of the male authors cluster.

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.