Reproducing “An authorship analysis of the Jack the Ripper letters” by Nini (2018)

Aylin Karapanar

10.03.2024

Reproducing “An authorship analysis of the Jack the Ripper letters” by Nini (2018)

This paper attempts to reproduce the paper by Nini (2018) titled “An authorship analysis of the Jack the Ripper letters” per the steps described by the author. Due to the lack

of specific pre-processing steps, upon investigating the files, I have decided to put everything in lowercase and filter out expressions stating that some parts of the letters are illegible and for the words, some parts are completed later, gather both parts.

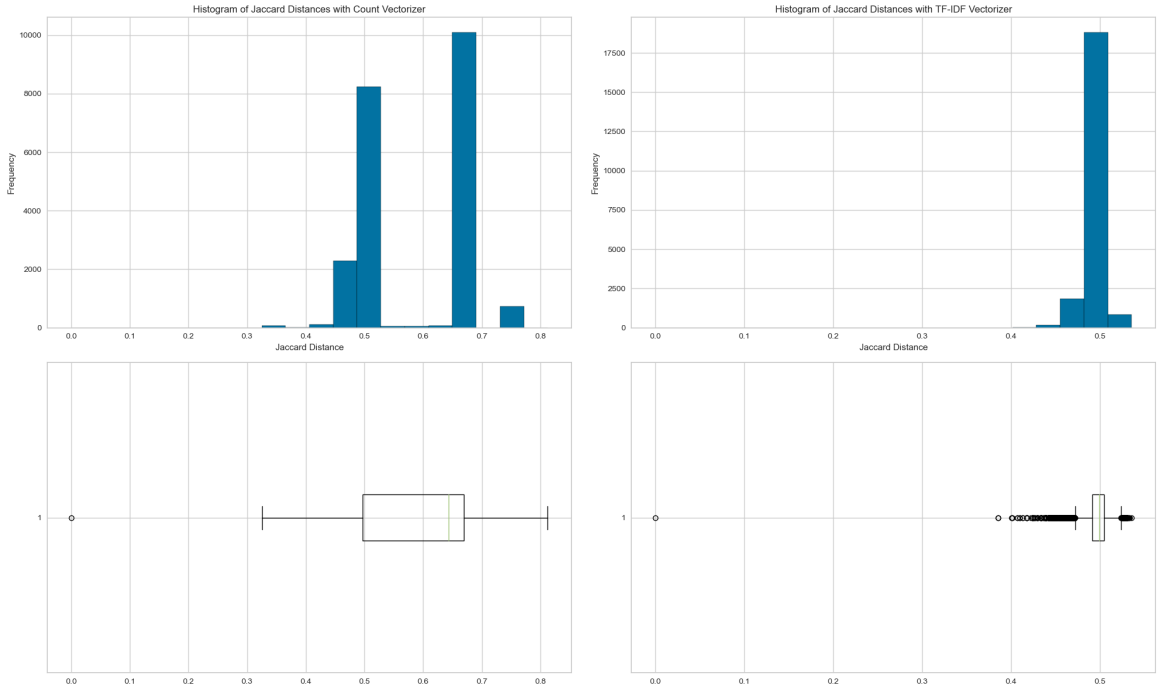
Afterwards, the bigram distances with Euclidian distances were calculated both using TF-IDF and count vectorizer since it was not specified by the author. When the ngram range was set to only bigrams, for TF-IDF vectorizer KMeans clustered 4 groups and hierarchical clustering clustered 2 groups; for count vectorizer, there were 3 clusters in KMeans and 2 in hierarchical. However, in this scenario, due to the algorithmic nature of KMeans which involves randomly selecting points as the centroids of the clusters, the number of clusters keeps changing in every run and is stabilized by setting a seed. The resulting dendrograms can be found in Appendix A. When the ngram range was set to include both bigrams and unigrams, for TF-IDF vectorizer both KMeans and clustering divided the letters into 4 groups; for count vectorizer, these were 3 clusters in both clustering. The aim of looking at both ngram range (2,2) and (1,2) was to determine which one was used by the author and whether it can provide additional information in clustering considering the words chosen can also signal a specific characteristic of an author.

The next step was to calculate Jaccard distance as done by Nini (2018), again due to lack of specification this step was executed by using both count and TF-IDF vectorizer for both ngram ranges of 2, 2 and 1, 2. As described eight bigrams were excluded. With count vectorizer, the Jaccard distances of only bigrams yielded 2 clusters using Ward's linkage which can be seen in Appendix B figure B1 and Jaccard distances of ngram range (1,2) yielded 3 clusters. Using TF-IDF posed a challenge as the values obtained were continuous which was not suitable for calculating Jaccard distances. Therefore, several approaches were tested: rounding the decimals, scaling by multiplication, scaling by using MaxAbsScaler, turning the values into strings and deleting the remaining decimals after the first four decimals. These all failed apart from using Binarizer. The bigram Jaccard distances were clustered into 5 clusters as shown in Appendix B figure B2, bigram and unigrams were clustered into 3.

All these different attempts showcase how different interpretations can lead to vastly different results. The combination of techniques that comes close to the number of clusters as obtained by Nini (2018) is the Jaccard distances with TF-IDF vectorizer. Even though it can be considered close there are some considerable differences; first would be the frequency and the distribution of the Jaccard distance values as displayed in Figure 1, the values change drastically with the use of different vectorizers, and the values obtained by the author who found that most of the values to be approximately 1 and only 25% to be smaller than 0.98.

Figure 1

Frequency of Jaccard distances for Count and TF-IDF Vectorizer



Additionally, Nini (2018) stated that the letter named ‘From Hell’ was linguistically different from other famous letters, but even with the use of two vectorizers this was not the case in my reproduction. The same thing also appears in terms of the ‘Midian’, ‘Saucy Jacky’, and ‘Dear Boss’ letters, they were clustered together in the paper, but in the reproduction, they were in two different clusters with no repeating pattern.

In conclusion, this article due to lack of analysis steps disclosure could not be replicated and the fact that the authors of these letters are unknown, it cannot be known for sure. However, still, this does not mean there were no similarities between the letters as bigram patterns are commonly used in plagiarism and authorship analysis.

Appendix A

Figure A1

Dendrogram Ward's Linkage (Bigram distances with TF-IDF Vectorizer)

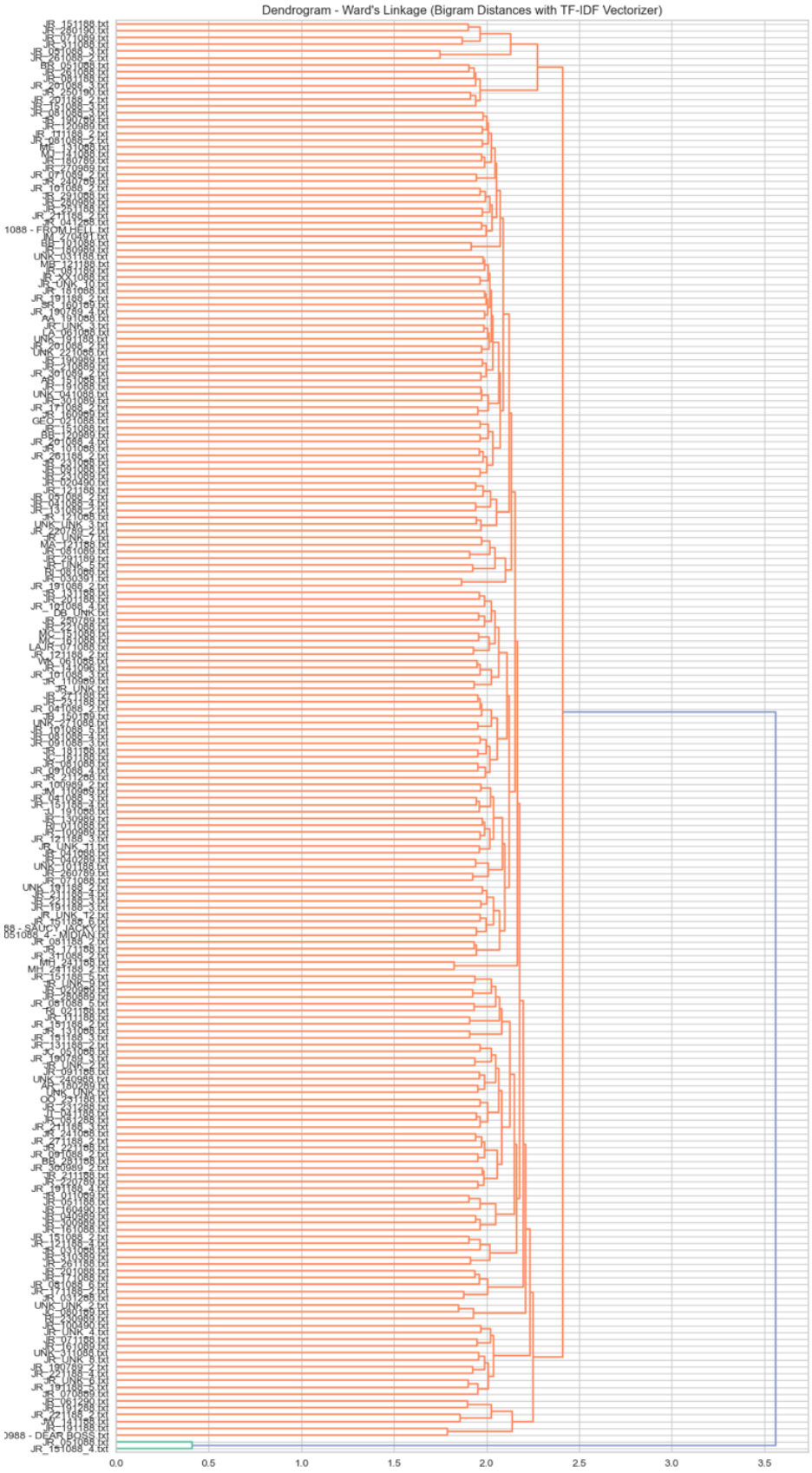
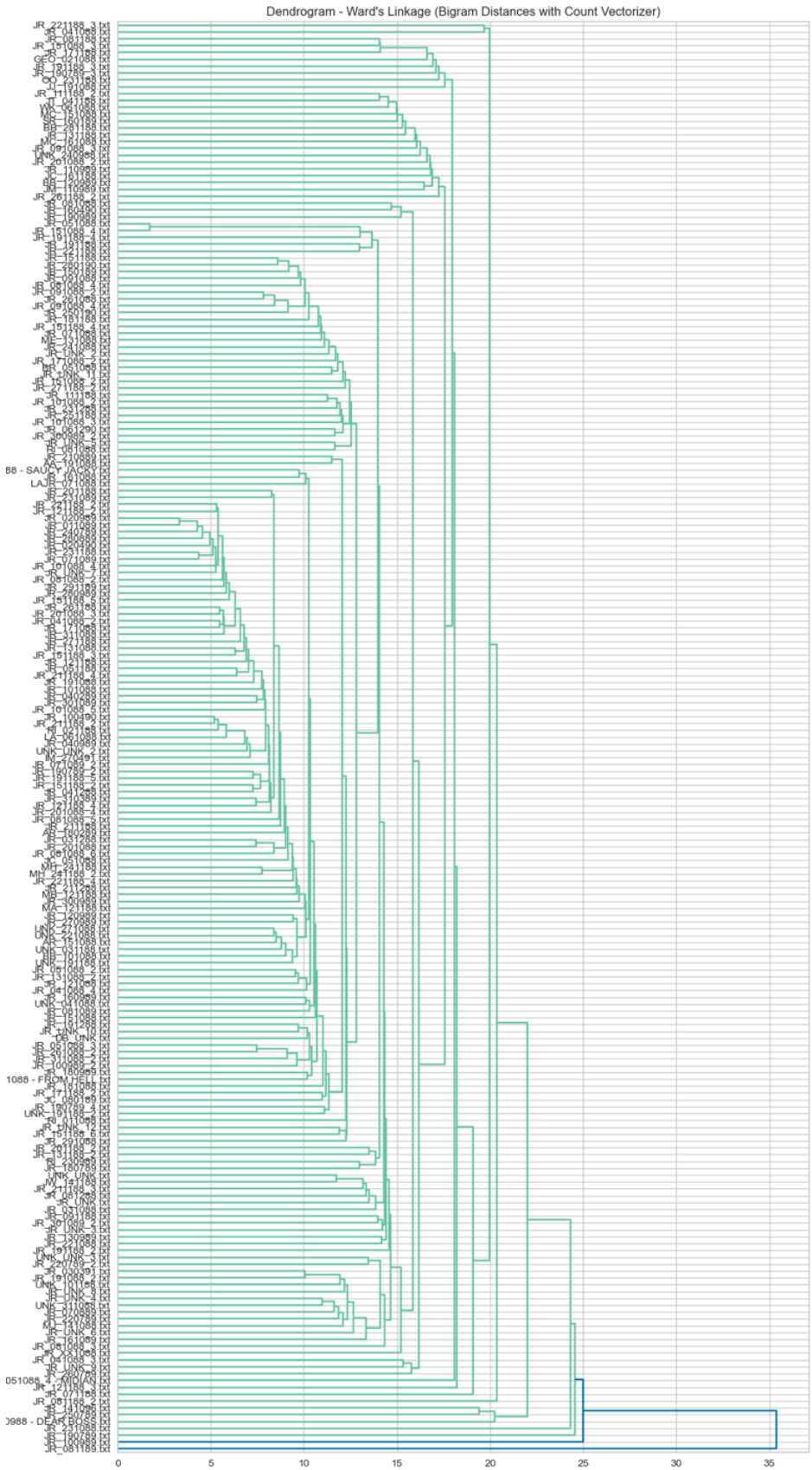


Figure A2

Dendrogram Ward's Linkage (Bigram distances with Count Vectorizer)



Appendix B

Figure B1

Dendrogram Ward's Linkage (Jaccard distances with Count Vectorizer)

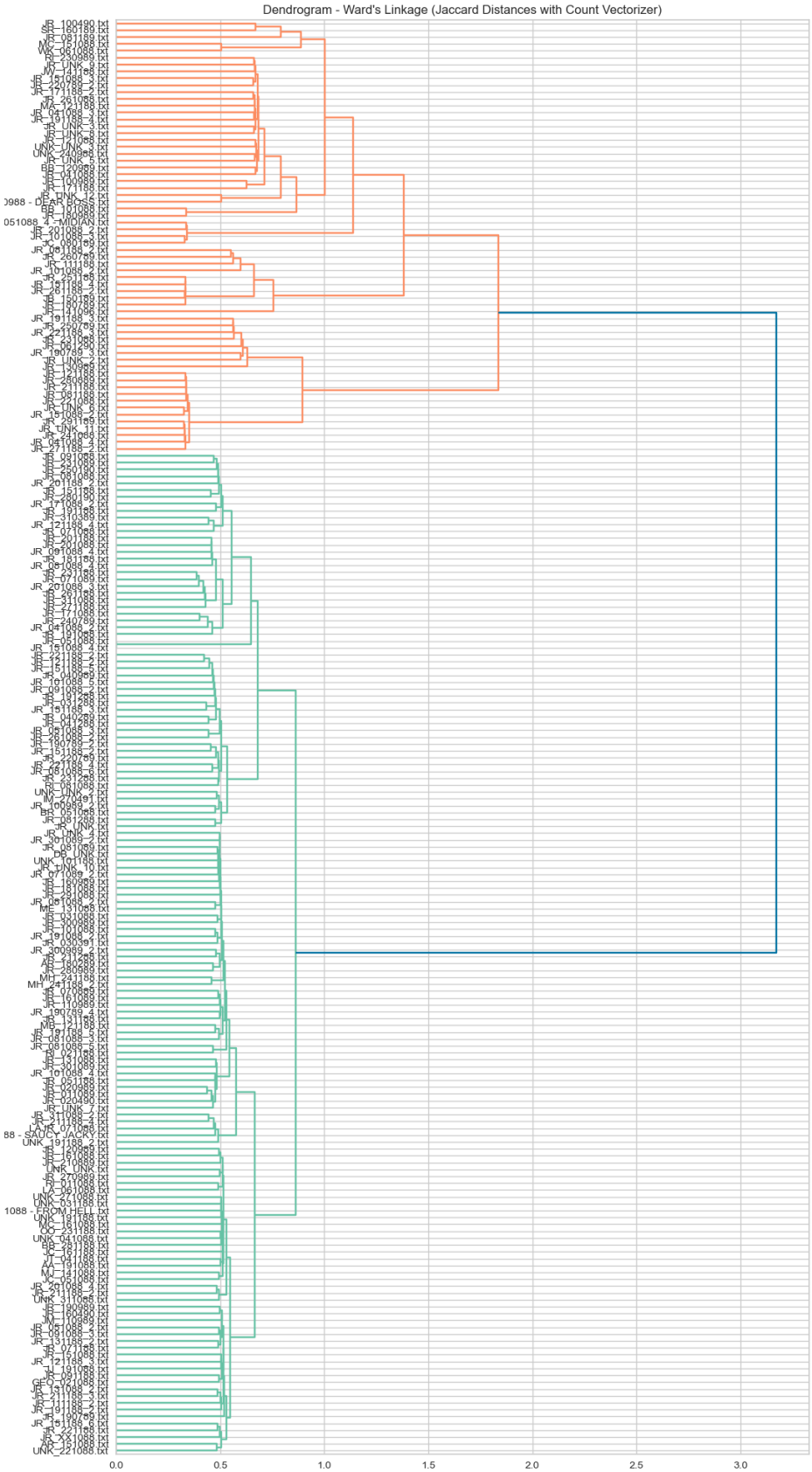
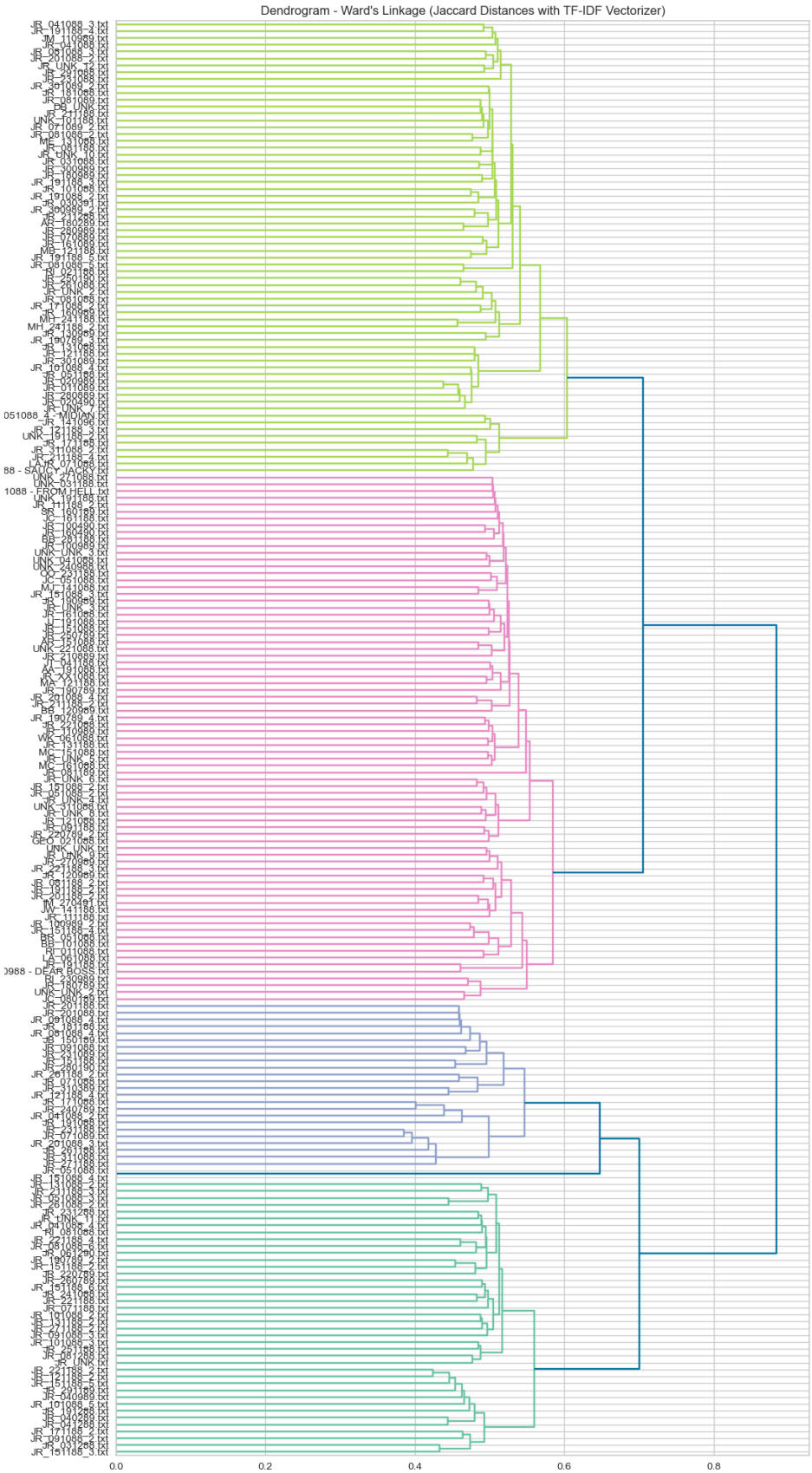


Figure B2

Dendrogram Ward's Linkage (Jaccard distances with TF-IDF Vectorizer)


Dendrogram - Ward's Linkage (Jaccard Distances with TF-IDF Vectorizer)
0.0
0.2
0.4
0.6
0.8