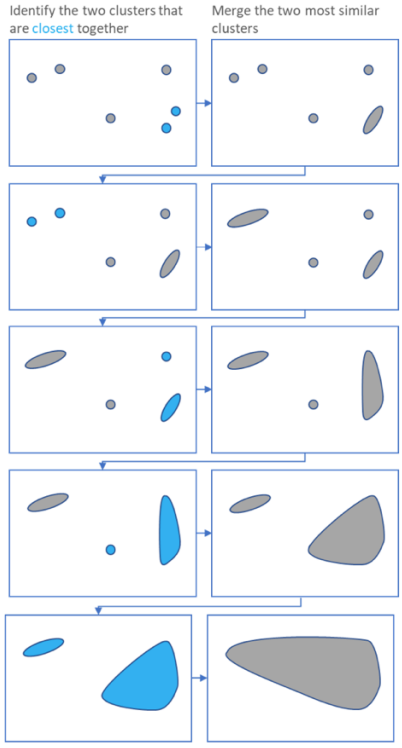
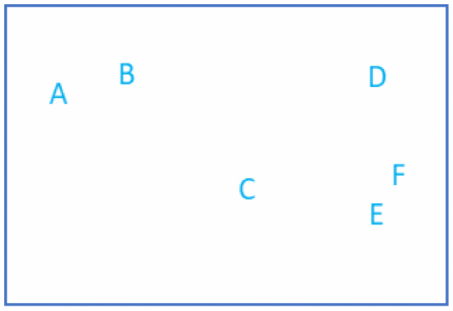
Hierarchical Clustering (HC)

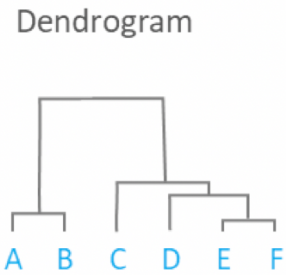
What is Hierarchical Clustering?

Hierarchical clustering is an algorithm that groups similar objects into clusters. In the NLP Suite in groups texts with similar Sentiment Analysis Score patterns together. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

The way the algorithm works is that it iteratively identify the 2 most similar clusters and merge them together until a tree is created. This tree is a called a Dendrogram.





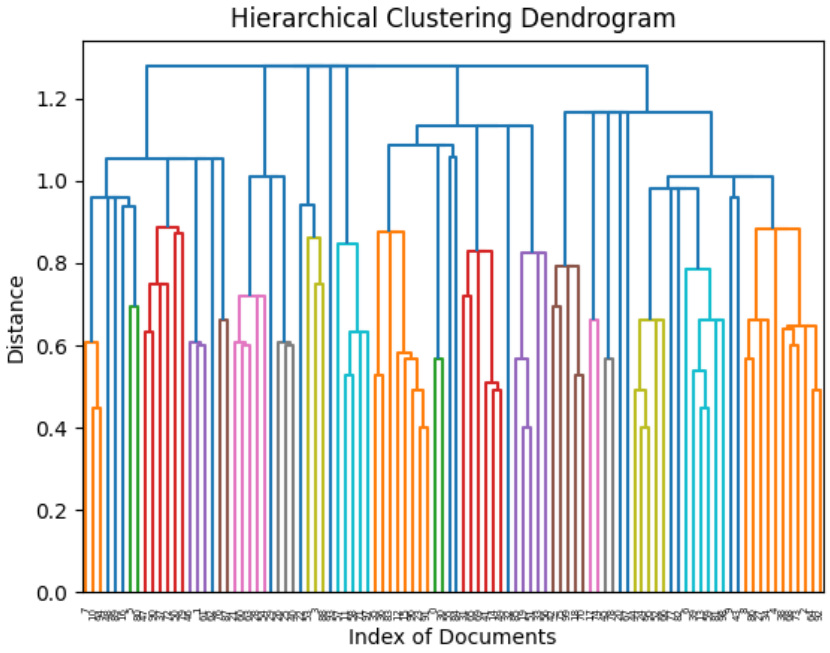


Images from DISPLAYR, Tim Block

Output

Dendrogram

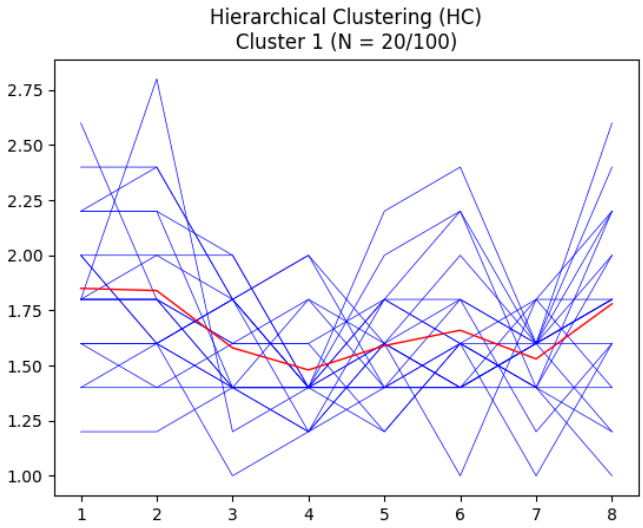
The first plot produced by the HC algorithm is the Dendrogram which shows hierarchies between clusters. It allows you to judge which clusters have strong links between themselves and which are far away in terms of sentiment analysis.



One issue with the dendrogram is that it is very hard to read when you have a large corpus.

HC Clusters graphs

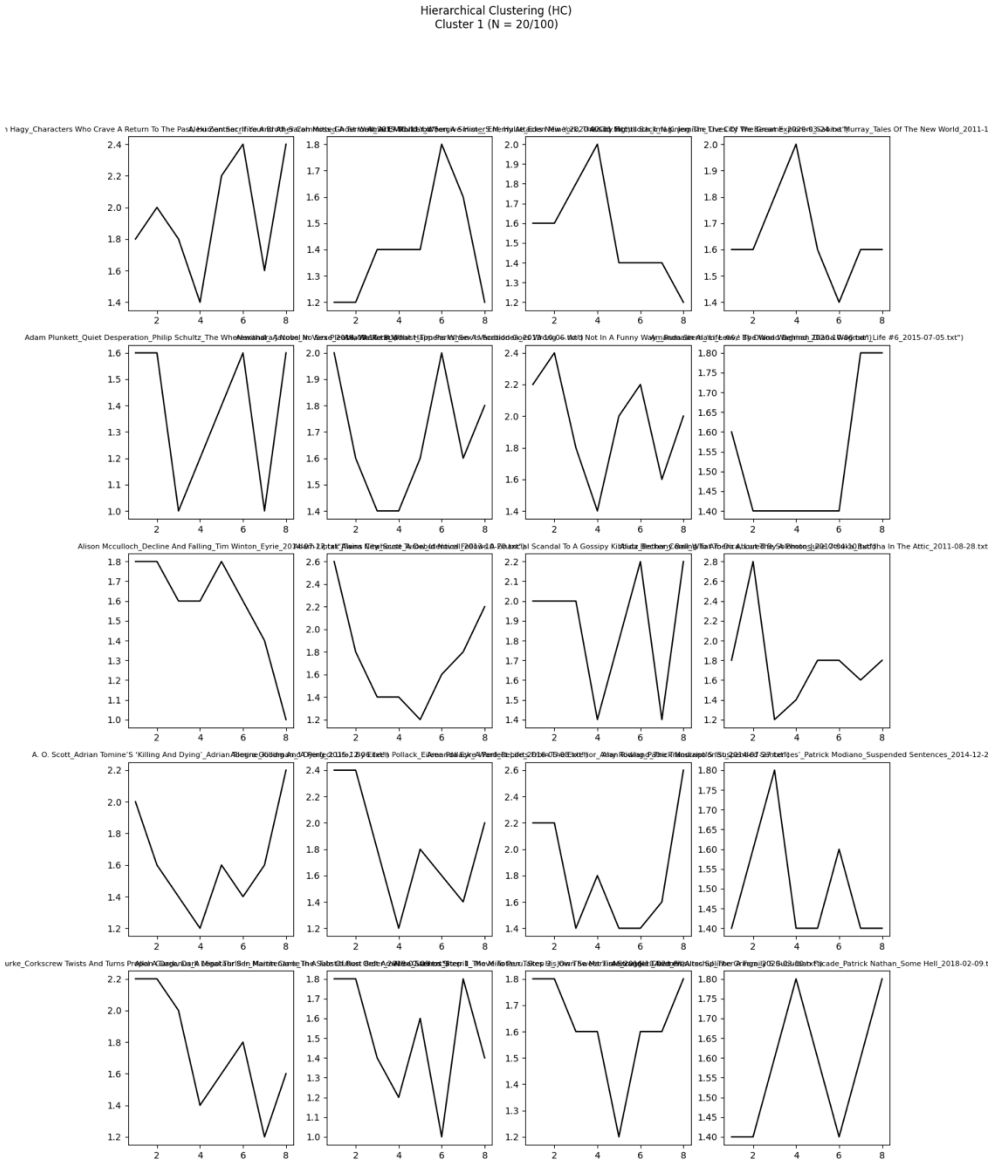
HC Clusters are displayed in the NLP Suite by line plots. Each Blue line represents a single text while the red line represents the general trend of the cluster.



It is important to notice that the x-axis corresponds to the vector size set up by the user before running the algorithm and the y-axis is the sentiment score. One graph will be created for each cluster.

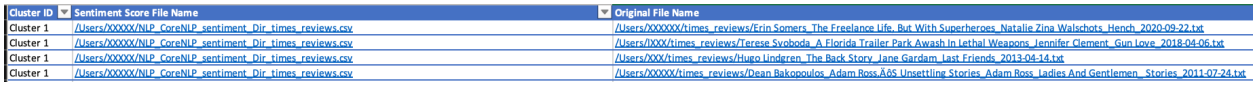
HC Clusters Subplots

The subplots generated let you have separated graphs for each text in your cluster. These subplots are ideal if you want to do closer reading onto a single text.



CSV

In addition to the 3 above graphs, you can use the csv file generated to find the content of each cluster in *Hierarchical Clustering Documents.csv*



Cluster ID	Sentiment Score File Name	Original File Name
Cluster 1	/Users/XXXXX/NLP_CoreNLP_sentiment_Dir_times_reviews.csv	/Users/XXXXXX/times_reviews/Erin Somers_The Freelance Life. But With Superheroes_Natalie Zina Walschots_Hench_2020-09-22.txt
Cluster 1	/Users/XXXXX/NLP_CoreNLP_sentiment_Dir_times_reviews.csv	/Users/XXX/times_reviews/Terese Svoboda_A Florida Trailer Park Awash in Lethal Weapons_Jennifer Clement_Gun Love_2018-04-06.txt
Cluster 1	/Users/XXXXX/NLP_CoreNLP_sentiment_Dir_times_reviews.csv	/Users/XXX/times_reviews/Hugo Lindgren_The Back Story_Jane Gardam_Last Friends_2013-04-14.txt
Cluster 1	/Users/XXXXX/NLP_CoreNLP_sentiment_Dir_times_reviews.csv	/Users/XXXXX/times_reviews/Dean Bakopoulos_Adam Ross.Â0S Unsettling Stories_Adam Ross_Ladies And Gentlemen_ Stories_2011-07-24.txt

The CSV file contains the Cluster ID of each text as well as the directory of the original .txt file and the directory of the Sentiment Scores of the .txt file. You can use the csv file generated to find the content of each cluster (the file name and cluster number) for the HC results.

References

Burrows, J.F. 1987. "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style." *Literary and Linguistic Computing*, Vol. 2, No. 2, pp. 61-70.

Franzosi, Roberto. 2004. *From Words to Numbers*. Cambridge, UK: Cambridge University Press.

Franzosi, Roberto. 2010. *Quantitative Narrative Analysis*. Thousand Oaks, CA: Sage.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". *EPJ Data Science*, Vol. 5, No. 31, pp. 1-12.

Vonnegut, Kurt. https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Vonnegut, Kurt. https://www.youtube.com/watch?v=GOGru_4z1Vc

Vonnegut, Kurt. 2005. "Here is a Lesson in Creative Writing." In: pp. 23-28, Kurt Vonnegut, *A Man Without a Country. Edited by Daniel Simon.* New York: Seven Stories Press.

Bock, T. (2022, January 17). *What is hierarchical clustering?* Displayr. Retrieved April 16, 2022, from https://www.displayr.com/what-is-hierarchical-clustering/