Part IV – Cluster Themes with Agglomerative Hierarchical Clustering

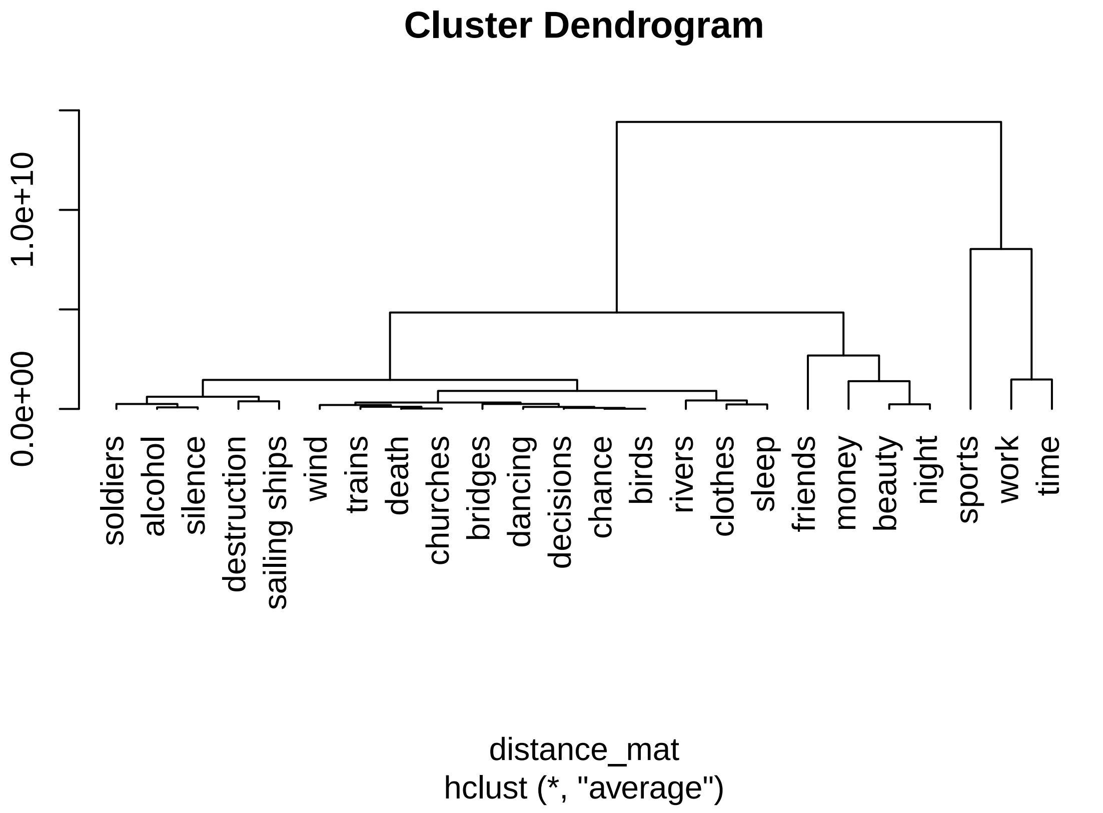
Rees Morrison

2022-01-29

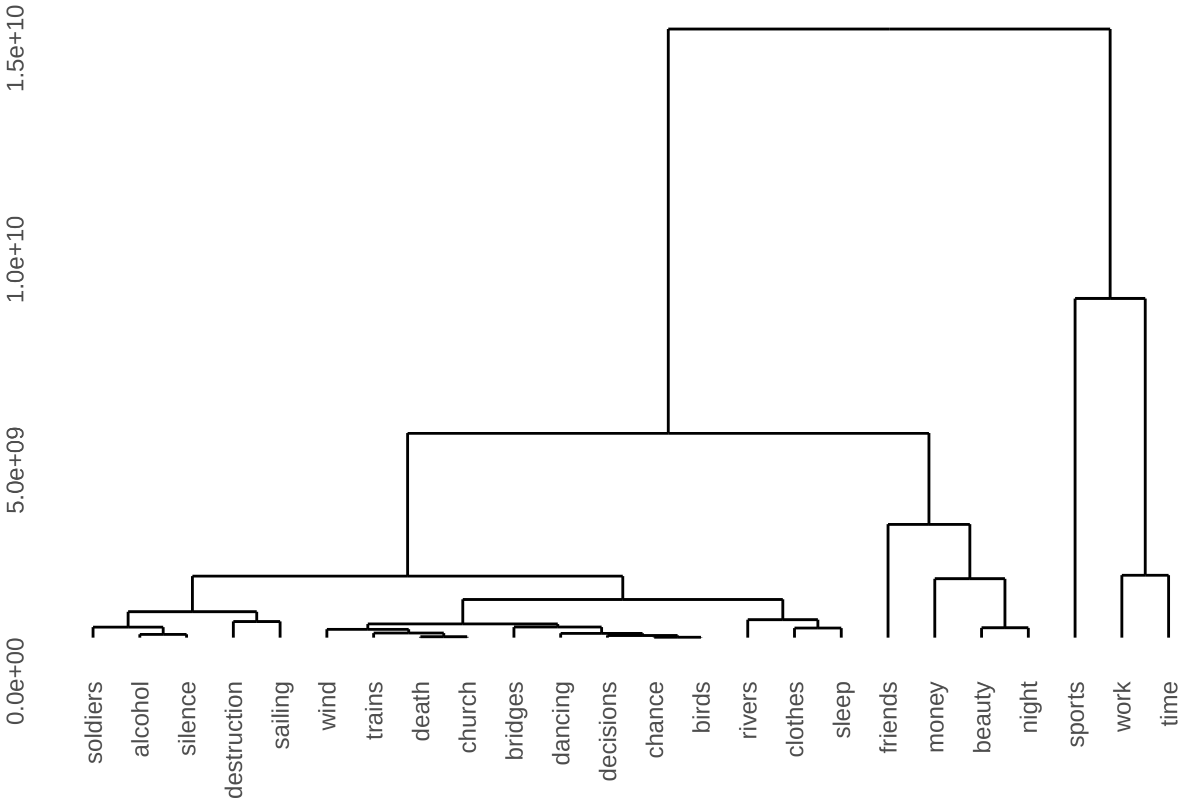
Part I of this working series introduced Themes, and Part II introduced the metrics for those Themes. Part III started the analysis of centrality and complexity by an unsupervised clustering algorithm, k-means clustering. It requires the analyst to specify the number of clusters, and figuring out the optimal number of clusters can often be hard. Part III side-stepped that challenge because the k-means algorithm was asked to find 12 clusters, ideally therefore identifying the closest, most-similar Themes. In this post we introduce a second algorithm for the same task. **Agglomerative hierarchical clustering** is an alternative unsupervised approach; it builds a hierarchy of closest Themes and doesn't require a pre-specified number of clusters. That orderly construction of clustering helps identify most-similar Themes, referred to as "closest pairs."

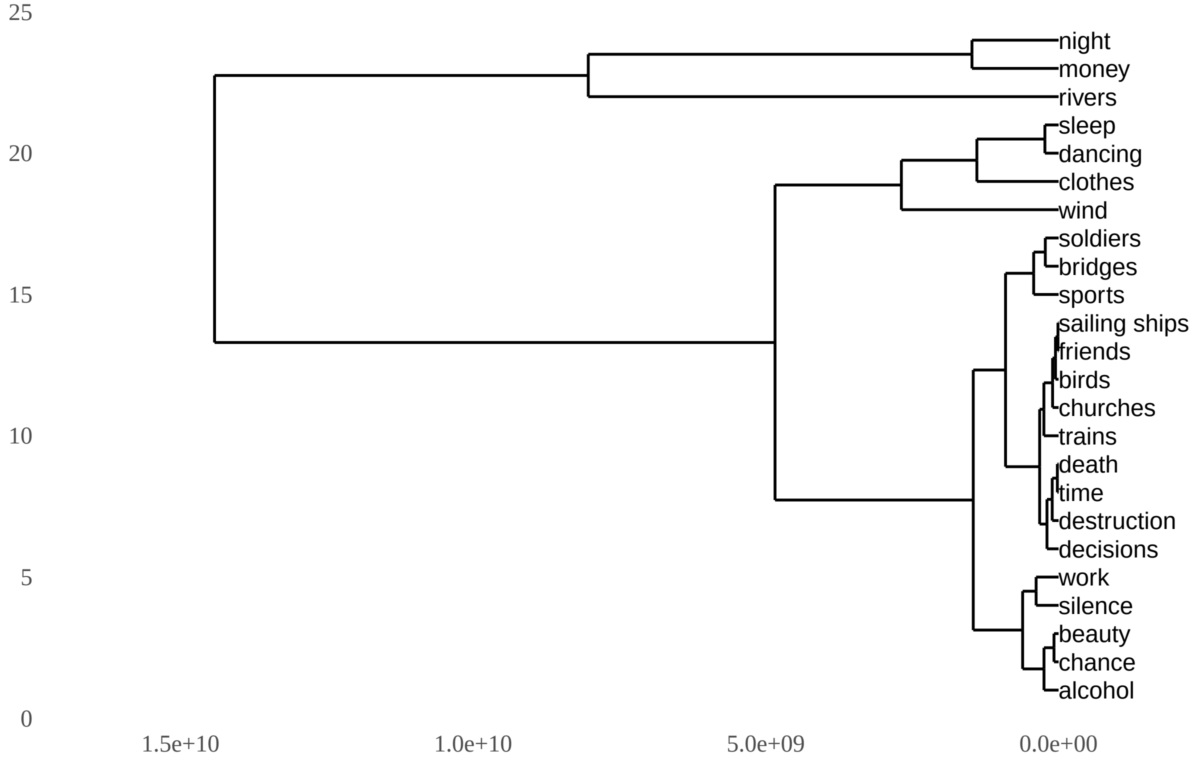
The algorithm starts by putting each Theme in its own unit. It then identifies the closest other Theme to a unit by the Euclidean distance between them in the hyperspace of metrics dimensions (each Theme is a vector in that 20-dimensional space). At first, a single Theme is clumped with another single Theme, each clumping thereby creating a two-Theme unit. Step-by-step the algorithm agglomerates each remaining Theme (or unit) to its closest Theme or unit. The algorithm keeps combining Themes with their closest Themes or units (based on the chosen **link method**, e.g., average linkage); if an already-agglomerated cluster is closest, it clusters the nearest Theme with that unit. Eventually, all the Themes rest in a single cluster.

Once the clustering ends, the results are usually visualized by a **dendrogram**. On the dendrogram, you can tell that two Themes are most similar because the height of the link that joins them together is short. That height on the y-axis is the value of the distance metric between the two Themes. so, Alcohol and Silence (second and third from the left on the x axis) are combined to form the first unit. The next-to-last agglomeration puts Sports with Work and Time (far right).



Let's use R's ggdendro package to plot the same data. You can see that Chance and Beauty are closest, so they cluster very low on the Y axis, and then Alcohol agglomerates into them.


1.5e+10
1.0e+10
5.0e+09
0.0e+00
soldiers
alcohol
silence
destruction
sailing
wind
trains
death
church
bridges
dancing
decisions
chance
birds
rivers
clothes
sleep
friends
money
beauty
night
sports
work
time



Dendrogram of Themes

Hierarchical agglomerative clustering of metrics