Part I - Concepts Quantitatively Compared and Explored

Rees Morrison

2022-01-26

This series of blog posts will serve as notes to myself on my long-term, evolving project. The motivation for my project is to explore whether and how concepts can be compared quantitatively to each other by their relative importance (centrality) or cognitive density (complexity). For example, can we say anything meaningful and empirically on those two dimensions to compare the concepts of "Silence" and "Destruction"? Stated differently, can data suggest one of those concepts is better known, more widely used, broader in associations – more central to English speakers, or deeper in thoughtfulness, more inter-connections, an abundance of associations – more complex?

Further, and more usefully, if metrics support such distinctions among concepts, how does that knowledge help us? It seems plausible that a provable distinction of centrality/complexity can help us reason more effectively, write more persuasively, understand differences between ideas more clearly, track and analyze word usage more insightfully, assess the readability of documents, improve Natural Language Processing (NLP) methods, or teach English more effectively.

This post, Part I of five, explains the source of the concepts (which I call "Themes") that are analyzed in the series. In short, a poem, an impressionist painting, a rock song, and a movie that share a Theme inspired associated ideas related to the Theme (which I call "Subthemes"). During 2021, as I identified Themes and fleshed out their Subthemes, I published groups of five posts for each Theme (four genre of art and one additional post) on my blog, Themes from Art.

Here are the 24 Themes included in this set of posts.

Table 1: 24 Themes

Theme	Theme	Theme	Theme
Alcohol	Clothes	Money	Soldiers
Beauty	Dancing	Night	Sports
Birds	Death	Rivers	Think
Bridges	Decisions	Sailing Ships	Time
Chance	Destruction	Silence	Trains
Churches	Friends	Sleep	Wind

This first series of five posts introduces the 24 Themes written about so far, the preliminary metrics I gathered to quantify and analyze those Themes (Part II), and three methods of clustering the Themes using the metrics: k-means, hiearchical agglomerative, and Reinert clustering (Parts III, IV, and V).

Part II describes the metrics that I have begun using to analyze the centrality and complexity of each Theme. A total of 20 metrics comprise the set for these first five posts; that set was whittled down from a larger group of candidate metrics based on my views of their reliability (is the data representative, broad, and thoughtful) and objectivity (has someone else curated the metrics). Each metric can be understood as a numeric value of counts or ranks for a Theme term. For example, the number of results returned by a search with Google for a Theme term constitutes one element of a metric. Eventually, in the terminology of the R programming language and having collected the relevant metrics, I created a 24-by-20 data frame with the

number of rows equal to the number of Themes (24), and the number of columns equal to the number of metrics (20). An Excel spreadsheet would have the same dimensions.

Part III introduces the first steps to analyze the Theme metrics to see how closely they associate with each other based on the metrics. Quantitative association between Themes takes a first step toward being able to classifying concepts along the dimensions of centrality and complexity; if the data indicates no clear associations, then this research program is scuppered. The approach uses unsupervised clustering, with the first algorithm applying *k-means clustering*. The results of identifying each Theme's closest other Theme starts what I refer to as the "pairings data set." The pairings data set for k-means clustering would (ideally) have a minimum of 12 pairs if each Theme were closest to another Theme, which was in turn closest to it. But it turns out that k-means produces more than 12 pairs, since A may be closest to B, but B may be closest to C.

With Part IV, we deploy another clustering methodology and add another set of closest-themes into the pairings data set. In that post we use *aggregative hierarchical clustering* to determine for each Theme which Theme is most similar to it, meaning which Theme (or combined themes) the hierarchical clustering algorithm merges each Theme with first.

Finally, in Part V we apply a third clustering methodology, *Rainert text clustering*. This clustering method relies on the text of the posts about Themes rather than the metrics data set. Once that algorithm has identified the closest Theme to another theme, because we have 24 Themes, our pairings data set has reached a minimum of a dozen "closest-Theme" pairs for each of the three techniques, and we can start drawing very preliminary conclusions about why Themes group the way they do.

As I add more Themes and modify the set of metrics (dropping, adding or modifying them), I plan to write more analytic posts that push the inquiry further. I primarily use the R programming language and would be pleased to provide my code and data to anyone who asks for them by emailing me: Rees(at)ReesMorrison(dot).com. Later posts will elaborate on the construct of Themes, such as whether they are representative or whether the terms chosen for the Themes make sense, as well as on the metrics, such as methodological challenges and decisions. Meanwhile, I will introduce additional software tools that can identify closest-Theme pairs. As that data set grows, more patterns should appear, and the preliminary conclusions should solidify.