**Relation of Word Association Strength to Word Length**

**Miles Keppler, Maddie Fletcher, Hailey Smith**

**CS 424: Social Network Analysis**

**Summer 2017**

**1 Abstract**

This paper discusses the implications of words with high association strength. A word with high association strength is one that is thought of when presented with many different word as stimulus.The goal of this project was to find outhow centrality in a word association graph corresponds to other lexical properties.

**2 Background**

The dataset we have chosen to work with consists of simple nodes and directed edges. Each node is a word, and each edge has a whole number strength that represents the number of people who associated word *a* to word *b*. In other words, an edge represents a person who responded with word *b* when stimulated with word *a*.

It is important to note that the only attribute of the nodes in this dataset is the word itself. From there, we can discover and use lexical properties of the words, such as word length.

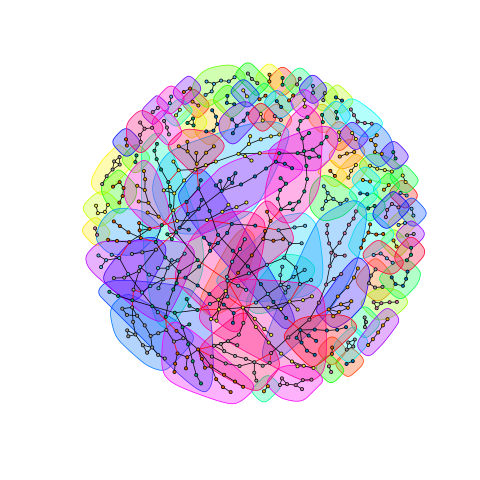
The dataset was originally collected between June 1968 and May 1971 by George Kiss, Christine Armstrong, Robert Milroy and J.R.I. Piper. Data was collected using a snowball technique, where an initial group was presented with a list of words (stimuli) in a nucleus set, and asked to write down the first word that came to mind after reading each word in the set. Those words were then added to the set of stimulus words, and the process was repeated. Data collection was terminated when the set of stimulus word reached 8,400.

The nucleus set was derived from “(a) the 200 stimuli used in the Palermo and Jenkins (1964) normq, (b) the 1,000 most frequent words of the Thorndike and Lorge (1944) word frequency count, and (c) the basic English vocabulary of Ogden (1954)” (EAT).

**3 Initial Analysis and Narrowing of**

**the Dataset**

At first look, the chosen dataset was too large to learn anything productive. At twenty-three thousand nodes and three-hundred twenty-five thousand edges, visual representations like Gephi produced a large blob, and numerical analysis tools like NetworkX could not handle the data. After further research, we found that the R iGraph library was able to produce meaningful results.

We decided at first to only examine the nodes and edges in the top five percent. After recursively eliminating nodes with zero in-degree (words no one thought of), we were left with a set of nodes and edges that was small enough to analyze efficiently. We were able to produce a visual representation of differing communities using the Louvain algorithm in R, including coloring the different communities within the graph, as well as partitioning the graph in a way that showed us the community with the largest amount of nodes. 

This subset of data is made up of the most connected nodes in the graph. Although this is only the beginning of our analysis, recent findings came to show that our hypothesis might be on the right track.

**4 Relevant Measures**

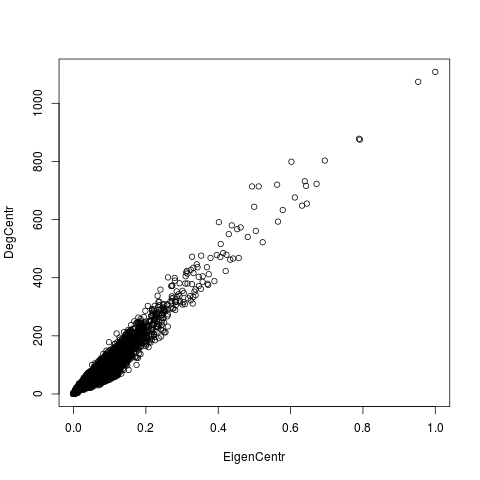
After learning about many different centrality measures, we decided the most relevant would be in-degree and eigenvector. In-degree of a node represents how many times that node was a response to any stimulus. Eigenvector centrality of a node represents how likely that word is to be thought of in response to any stimulus. In this context, these measures represent the prominence, or “importance” of a word in the English language.

We also explored closeness and betweenness centrality, but the data they presented did not make sense in the context. Other node-level measures did not make the cut, as they did not provide any meaningful information about the data. We also looked briefly at graph-level metrics, but these were also concluded irrelevant because we have no other graphs to compare to.

Identifying properties of edges and edge strengths would have been another good direction to take this research. Unfortunately, when calculating communities and clusters, we were unable to carry over the edge weight information in a meaningful and accessible way.

**5 Centrality Analysis**

Using the R iGraph library, we calculated both the eigenvector centrality and in-degree centrality of every node, and found that they were almost exactly proportional to each other, e.g. a node with high in-degree will have high eigenvector centrality. Here you can see, there are no outliers to this relationship.



These are the measures we used when comparing centrality to other lexical properties.

In our brief stint with Gephi, although it wasn’t able to provide a visual, it did allow us to find a few different graph-level metrics. First, the diameter is seven. This means that, starting with any word in the set, it would take an average of 7 stimulus-response pairs to reach a word chosen as the goal. We also found the average clustering coefficient to be 0.062 -- not very clustered, which was easy to see with the visual representation we produced of the smaller dataset. There were a lot of two-word pairs that weren’t connected to the rest of the graph. These pairs were eventually removed from the working dataset.

**6 Combining Centrality and Lexical**

**Analysis**

The main lexical property we looked at was word length. In this category, we hypothesized that length would be inversely proportional to centrality, i.e. shorter words would be more central. To a point, this was proven true.

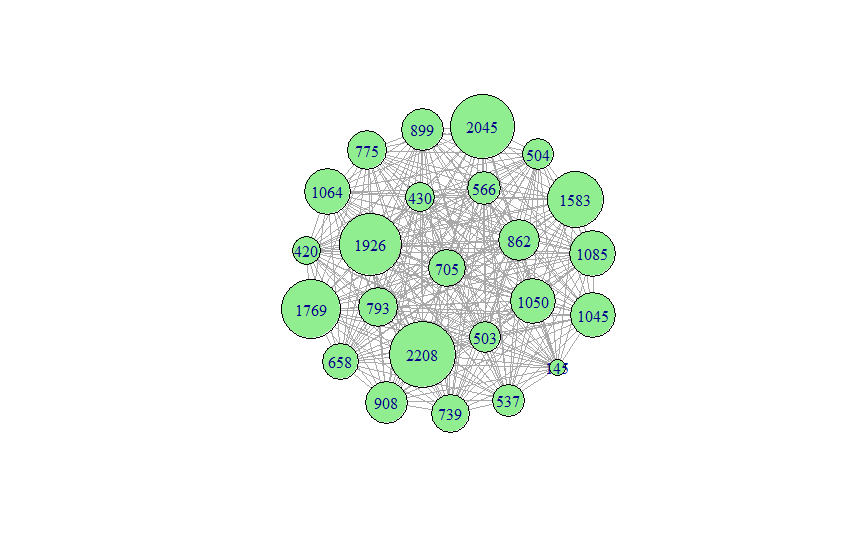
We found that 95% of the variation in the mean eigenvector centrality could be accounted for by looking at where are constants, and is word length.

There is a significant relationship between word length and centrality, as well as between the number of words in a phrase and the eigenvector centrality of that phrase. This was an especially good estimate with words of length between 5 and 15 characters. Unfortunately, The variance of words of a given length was high for each word length, so we were unable to find a method of detecting specific words with high or low eigenvector centrality.

An average degree of about 28 tells us that on average, there are likely to be near 28 words that stimulate a response of, or are a response to a random given word.

Most of the words that resided in the largest community were shorter in length, while, looking at the smaller communities, we found that their word length was greater than those that were more central to the graph. The most central nodes were also in the community that had the most nodes.

**7 Communities**

We were able to uncover 24 different communities in the entire graph. There was little that we could find that would cause the communities to form the way they did. The communities were fully connected, however we suspect the edges between communities were relatively low strength compared to the edges within the communities. There was large variation in size of communities, ranging from 140 to 2200 nodes, with little to explain the variation in size.****

When we examined the top 24 central nodes, we were able to extract 6 fully connected communities; 3 contained 2 words, 2 contained 4, and the final community contained 10 words. The most central word in our data set “me”. Within this elite set of words, we were unable to detect relations to sentiment analysis and general lexical frequency from our communities and high centrality words. There were, however, no phrases, and only two words of 6 or more characters. Those words being “nothing” and “people”.

**8 Conclusion**

There is a significant correlation between word length and word centrality in the word association graph. This means that shorter words are more likely to be thought of. Phrases, or responses with more than one word, are even less likely to be thought of than single words of the same length as phrases.

The main use of the data we found is in marketing: use shorter words. Make brand names short, make slogans short, use short words. This idea is already widely in use, which can be seen in the easily remembered brand names like *Google* and *Sharpie*, and the easily remembered slogan “Got milk?”. Another application, also in marketing, is to use commonly associated words to suggest an idea in an advertisement without stating it directly.

**References:**

1. The Edinburgh Associative Thesaurus, http://vlado.fmf.uni-lj.si/pub/networks/data/dic/eat/Eat.htm
2. Kiss, G.R., Armstrong, C., Milroy, R., and Piper, J. (1973) An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), *The Computer and Literary Studies*. Edinburgh: University Press.
3. The present version of [The Edinburgh Associative Thesaurus](http://monkey.cis.rl.ac.uk/Eat/htdocs/eat.zip) (EAT)(ZIP, 2.7M)
4. Coltheart, M. (1981) MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology, 3A, 497-505.
5. Download [MRC Psycholinguistic Database 2](http://ota.ahds.ac.uk/texts/1054.html)