Project Title: Cues to Use

Team Members :

Raghav Chegu Shyam Kumar: racheg@iu.edu
Shefali Mahendra Luley: sluley@iu.edu
Tanay Sainath Kulkarni: tankulk@iu.edu

Abstract:

It's fascinating to see what words a person remembers when given a certain term, since recalling certain words gets progressively difficult as the context or scenario in which the person is in changes. We will aim to understand how a person remembers specific words based on cues to other words in this project, as well as explore certain patterns in how a human memorizes and correlates words with one another.

Introduction:

How do we humans normally map some words to other words that may or may not have the same meaning but are related in some way? We don't consciously consider these words, but we inadvertently deduce or recollect them through other words. The name "Cues to Use" implies that there are some cues that serve as pointers for memorizing specific words. This provides an intriguing application case for research into how people think and correlate words with one another. This project's dataset includes not just word associations, but also information on other properties of these words. The scope of this project goes beyond just understanding association patterns; there are many verticals of network science, such as centralities and community detection, that can help us understand how different cues might stick together, or which words attract the most usage amongst all cues or recall words. Finally, the project is an application that may be mapped to several fields of psychology, linguistics, and neuroscience in addition to network science.

Proposal:

We aim to utilize the data and deep dive into the connections between the cued words and the recalled words. To understand this, we are employing network concepts such as centralities and communities to detect any connections formed or the reasons behind them. Moreover, classifying these words under distinct topics, sub-topics, or parts of speech and getting their links will help us understand the broad connections between them.

Reference:

Nelson, McEvoy & Schreiber: **http://w3.usf.edu/FreeAssociation/AppendixA/index.html**

Acknowledgment:

We would like to express our sincere gratitude to Prof. Yong Yeol Ahn and all associate teachers for their guidance and timely feedback on our work.

Final Paper content

Introduction:

Words are one of the most fundamental components of today's world, and it's fascinating to see what words individuals recall when given a specific term. People are more likely to recall a word that has a memory or an experience associated with it or a meaning mapping to the word's context. In this study, we will aim to figure out how people recall words based on cues from other words, as well as look for patterns in how people memorize and connect words.

The fundamental goal of this research is to learn more about the human cognitive process and to figure out why people remember specific words when given a cue word in a specific context.

The question that has piqued our interest is:: **Can we analyze which words will be remembered when given a word, and can we know what qualities those words have?**

Even though this is a highly subjective topic, we can divide it into several sections to investigate various elements that influence word associations. It's always been fascinating to see what a person would recall in different situations using varying analyses. There are indeed a variety of models that do this now, and we want to be a part of breaking down these barriers by framing these issues as a network problem. Using this data to better understand human behavior and thinking in a certain setting is one of the reasons for this endeavor, among others. By producing communities, shortest paths, and clustering, among other aspects, it becomes easier to evaluate the structure and relationships of words as a network problem.

In this Project "Cues to Use", the dataset contains word associations as well as other properties of the cues and target words. Properties about the associations such as the number of people responding to the cue, forward and backward strength of the cue, and target words provide us with good insights into the relations between the words. We have used these properties to set appropriate filters to work on a smaller subset of data. The filters used in this study are: records with no target to cue backward strength (which says that there is no impact of the target word on the cue word), a minimum of 3 people responding to the cue word, having only nouns, adjectives and verbs as the cue words and considering valid data for analysis which, according to the dataset, means only normed words). We also use the cue and target words part of speech tags to dig deeper into their relationship.

Furthermore, due to a large number of cue-target terms, we associate the words with their hypernyms to obtain their subtopics and topics, from which we can deduce insights into the association from a higher perspective.

Methods:

We employ multiple methods to understand the associations of what words are reminded given certain cue words as shown below:

1. **Part-of-Speech Tag**
2. **Vowels Association**
3. **Subtopics Association**
4. **Topics Association**

Part-of-Speech Tag:

In the dataset, the cue and target words are tagged with the respective Parts of Speech. Using these tags, we build a cue-target network and visualize them in Gephi software.

During pre-processing, it could be seen that the part-of-speech tags such as Noun(N), Verb(V), and Adjective(ADJ) were the most prominent tags and the others were negligible, hence, removed from the dataset. Apart from this, the mean of the Forward Strength seen between them is taken as the weight for the edges in the network, thus, showing the choice of the tags on others.

The network is loaded as a directed graph in Gephi and the following attributes are used:
Node Size: Ranked by Degree
Edge Color: Weight(FSG)

Vowels Association:

Using the dataset and filtering the cue words to start only with vowels. This is done as the dataset is too large and would like to see certain associations of the words.

After preprocessing the data, along with the cue-target words, we are also taking the Forward Strength observed between them as the weights for the edges in the network, thus, showing the choice of the words on others.

The network is loaded as a directed graph in Gephi and the following attributes are used:
Node Size: Ranked by Betweenness centrality
Node Color: Community formation using Modularity
Edge Color: Weight(FSG)

Subtopics Association:

Since the number of cue-target words is quite high, we get the hypernyms of these words and use them in place of the cue-target words. These words being used are the subtopics that can be used to generalize the words.

After initial pre-processing of the data, we found that there were a lot of words that did not have a subtopic associated with them, thus, these records were removed from the dataset. Apart from this, the mean of the Forward Strength seen between them is taken as the weight for the edges in the network, thus, showing the choice of the subtopics on others.

The network is loaded as a directed graph in Gephi and the following attributes are used:
Node Size: Ranked by Betweenness centrality
Node Color: Community formation using Modularity
Edge Color: Weight(FSG)

Topics Association:

Since the number of cue-target words is quite high, we get the hypernyms of these words and use them in place of the cue-target words. To further generalize them, we take the hypernyms of these words which are the topics of the given words.

After initial pre-processing of the data, we found that there were a lot of words that did not have a topic associated with them, thus, these records were removed from the dataset. Apart from this, the mean of the Forward Strength seen between them is taken as the weight for the edges in the network, thus, showing the choice of the topics on others.

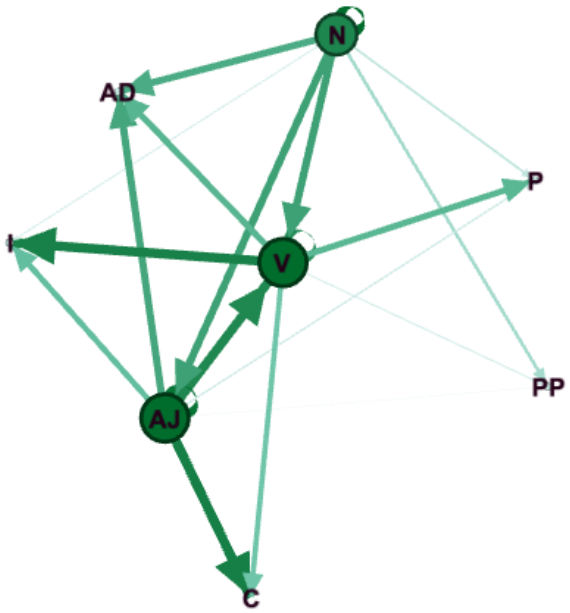
The network is loaded as a directed graph in Gephi and the following attributes are used:
Node Size: Ranked by Betweenness centrality
Node Color: Community formation using Modularity
Edge Color: Weight(FSG)

Results:

The above four methodologies are implemented to get further insights into the characteristics or choice of words being chosen as shown below:

Part-of-Speech Tag:

The network visualized for the Part-of-Speech is shown below:



In the above network, we can observe the following characteristics:

- **Nodes:** 8; **Edges:** 20; **Avg Degree:** 2.5; **Avg Weighted Degree:** 0.124

The following insights can be derived from the network -

- The prominent Parts-of-Speech are Noun(N), Verb(V), and Adjective(AJ).
- Adverb(AD) is the most recalled word given any prominent Part-of-Speech.
- Verbs are recalled given Adjective or noun but not vice versa.
- As the dataset consists mostly of the prominent Parts-of-Speech, the insights could be slightly biased, and better observations can be made with a balanced dataset/network.

Vowels Association:

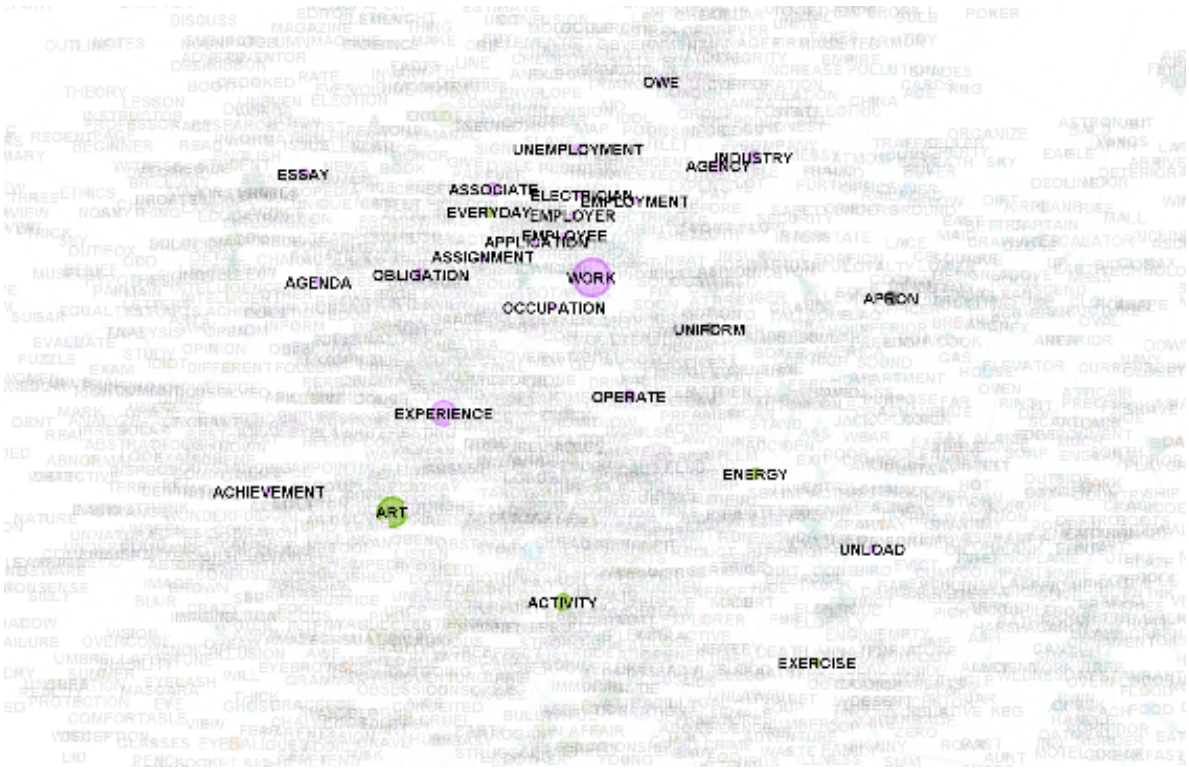
The network visualized for Vowels Association is shown below:

In the above network, we can observe the following characteristics:

- **Nodes:** 2126; **Edges:** 4208; **Avg Degree:** 1.979; **Avg Weighted Degree:** 0.122
- **Avg Clustering Coeff:** 0.014; **Avg Path Length:** 4.031; **Communities:** 8 (Res=5.0)

We can see that the network is sparse and does not have many nodes with strong associations with their neighbors. The large nodes indicate the betweenness centers and we will explore through some examples to find further insights in this network.

Vowel Association: Work

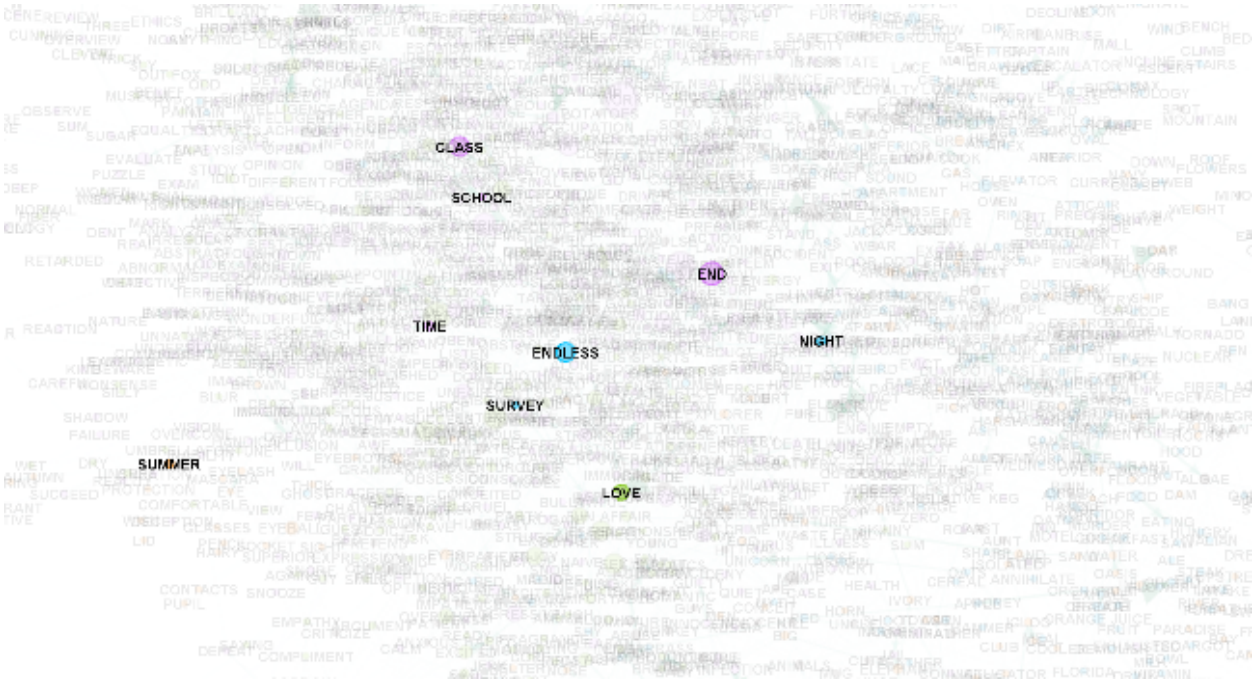


In the network, the node with the label Work has been taken into consideration.

Observations:

- It can be seen that it is one of the betweenness centers and is well connected across different communities.
- The number of neighbors is also quite high.
- Another important thing to be noted is that even though the Cue words were filtered for words starting with vowels, the node Word starting with a consonant is one of the biggest betweenness centers.
- This could indicate that there could be consonant words that are more important to connecting other words and can be verified with more examples.

Vowel Association: Endless



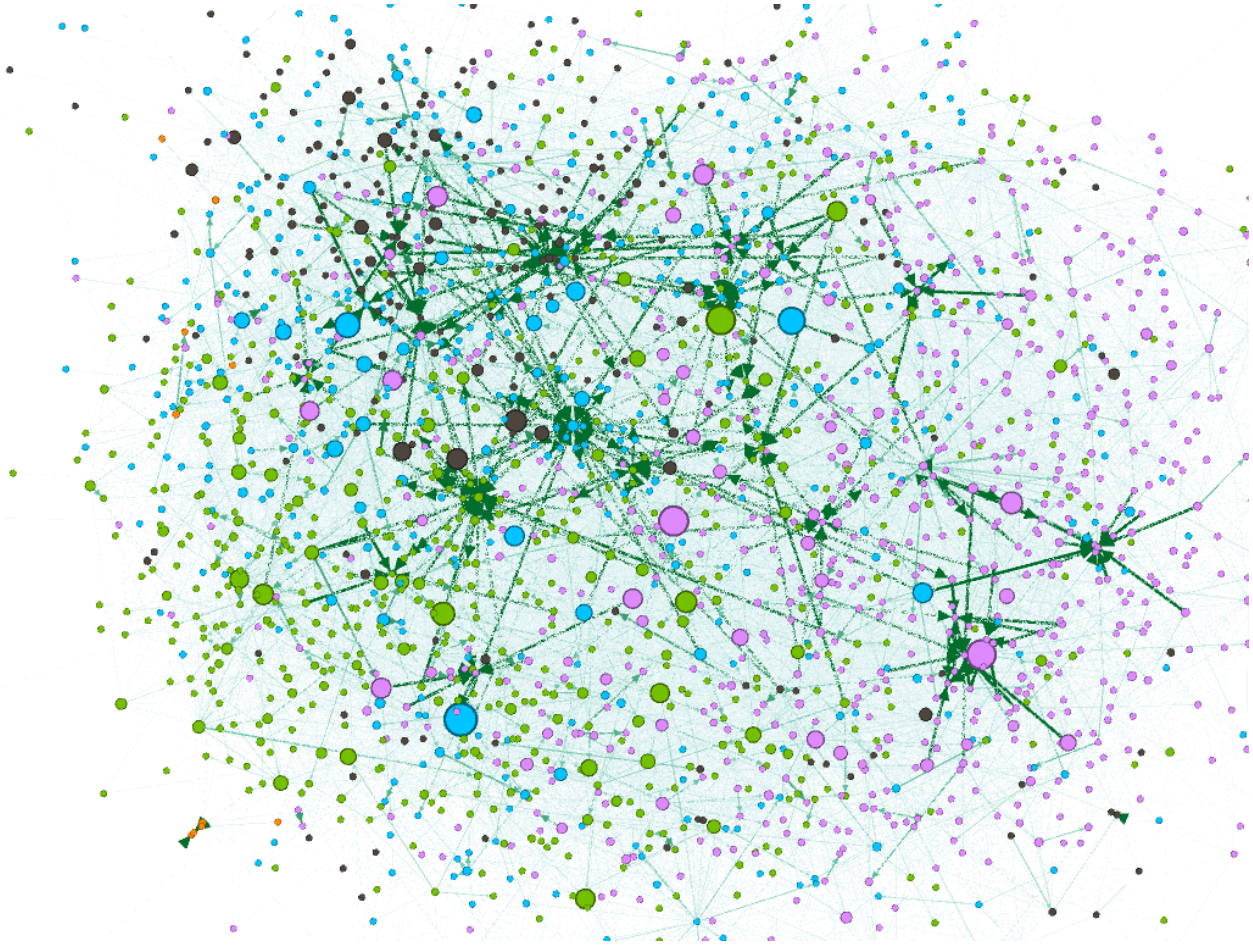
In the network, the node with the label Endless has been taken into consideration.

Observations:

- It can be seen that it is one of the smaller betweenness centers and is not well connected across different communities.
- The number of neighbors is also quite low.
- This could show a slight indication that vowel words may not be as important when bridging between multiple words.

Subtopics Association:

The network visualized for the Subtopics Association is shown below:

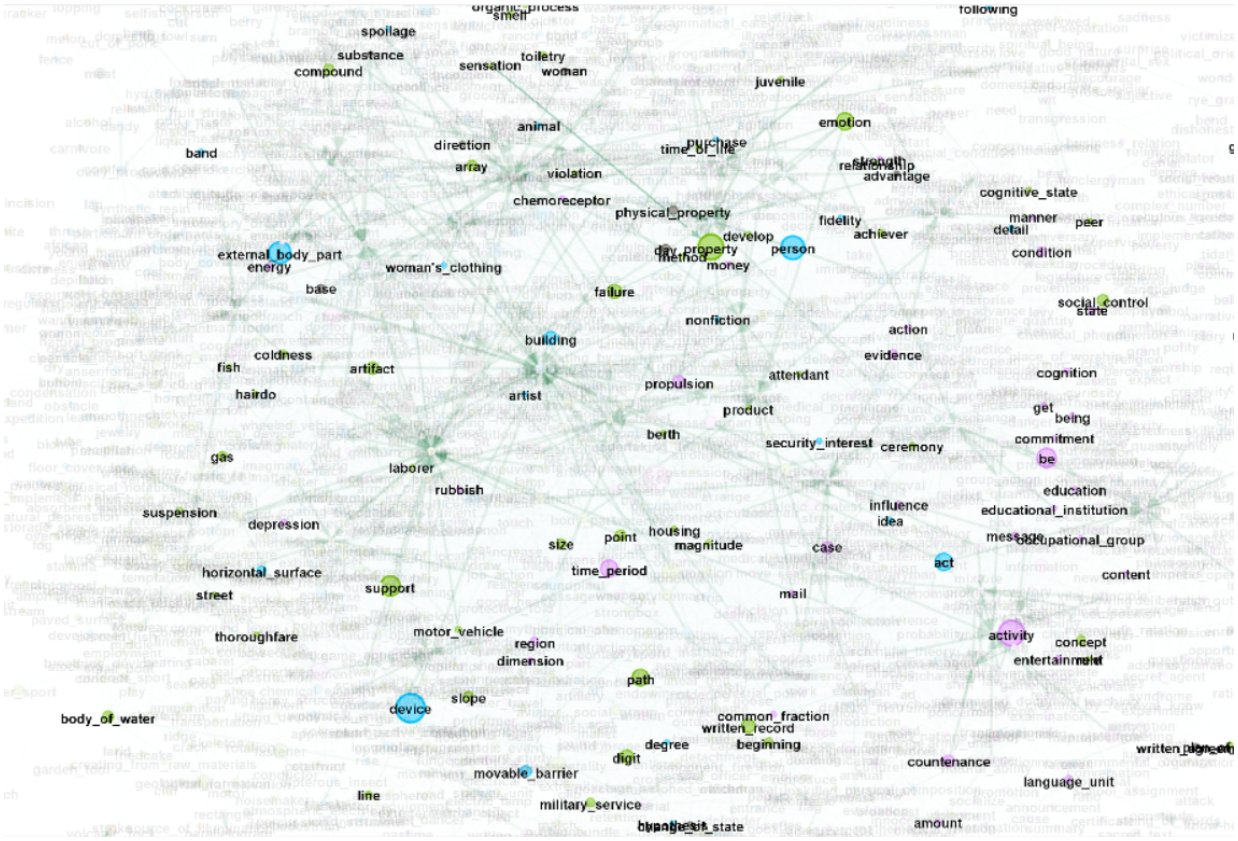


In the above network, we can observe the following characteristics:

- **Nodes:** 2154; **Edges:** 17933; **Avg Degree:** 8.325; **Avg Weighted Degree:** 0.651
- **Avg Clustering Coeff:** 0.082; **Avg Path Length:** 3.505; **Communities:** 12 (Res=2.0)

We can see that the network is very dense and has a lot of nodes with strong associations with their neighbors. The large nodes indicate the betweenness centers and an interesting thing that can be noted is that the strongly associated nodes are not the betweenness centers. We will explore some examples to find further insights in this network.

Sub-Topic: Property

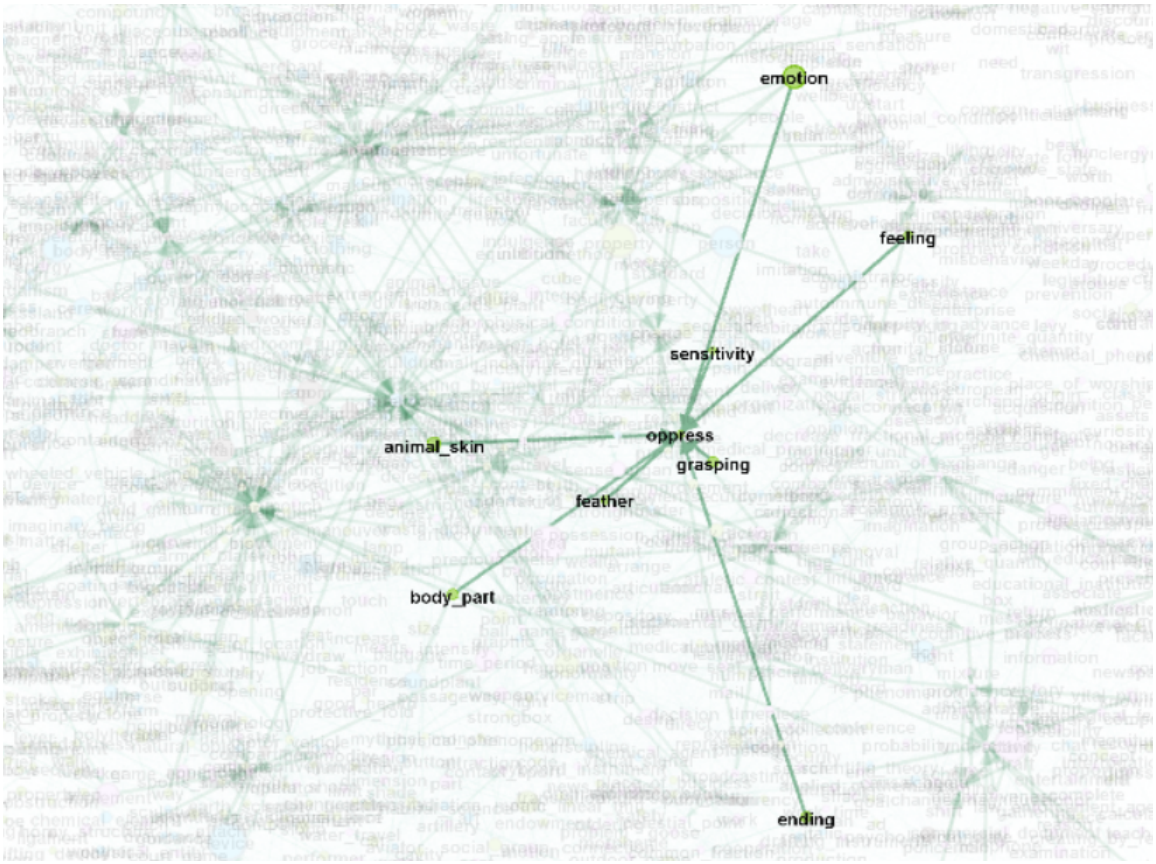


In the network, the node with the label Property has been taken into consideration.

Observations:

- It can be seen that it is one of the betweenness centers and is well connected across different communities.
- The number of neighbors is also quite high.
- An insight that can be observed is that although it has a high neighbor count, the associations with them are weak. This could indicate that there are some hidden properties that could give more insight upon further exploration.

Sub-Topic: Oppress



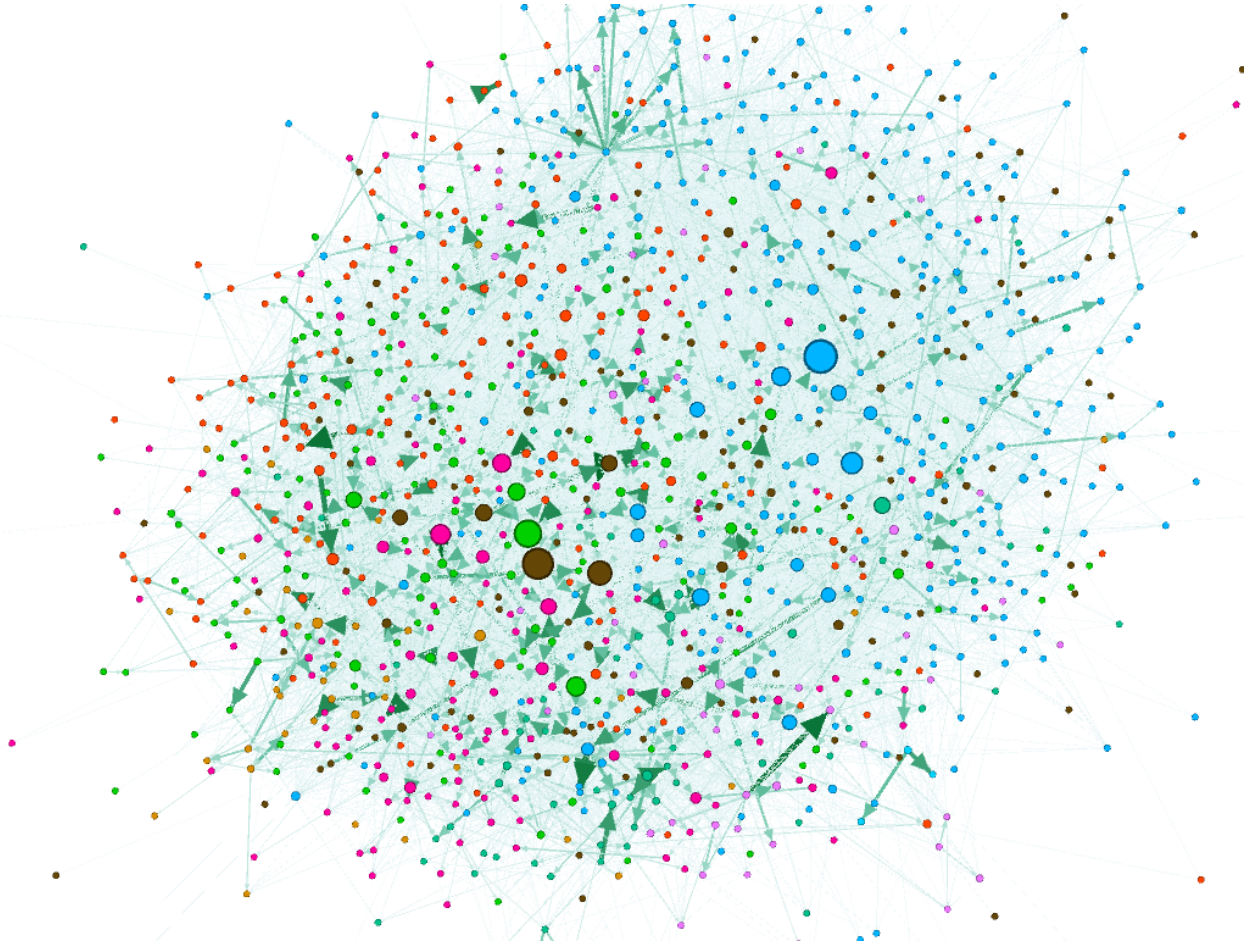
In the network, the node with the label Oppress has been taken into consideration.

Observations:

- It can be seen that it is strongly associated with all its neighbors in the network.
- The number of neighbors is quite low.
- An insight that can be observed is that although it has a low neighbor count, the associations with them are strong. This could indicate that there are some hidden properties that suggest more association with certain properties that can be found with further exploration.

Topics Association:

The network visualized for the Topics Association is shown below:

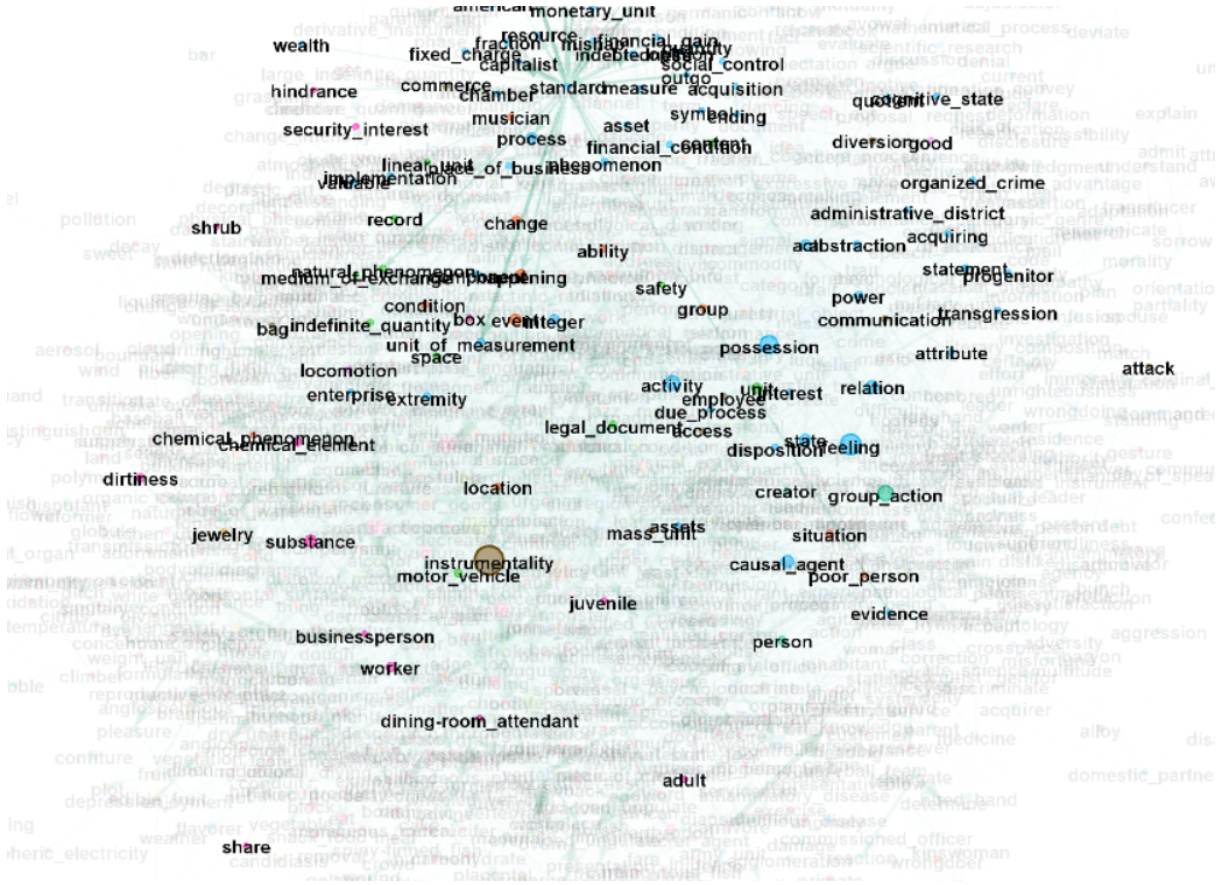


In the above network, we can observe the following characteristics:

- **Nodes:** 1185; **Edges:** 14951; **Avg Degree:** 12.617; **Avg Weighted Degree:** 0.795
- **Avg Clustering Coeff:** 0.155; **Avg Path Length:** 2.76; **Communities:** 10 (Res=1.25)

We can see that the network is very dense and has nodes with a mix of strong and weak associations with their neighbors. The large nodes indicate the betweenness centers and it spans across the different communities formed. It could also indicate that this follows a scale-free graph, that is, small-world or even ultra small-world. We will explore through some examples to find further insights in this network.

Topic: Standard

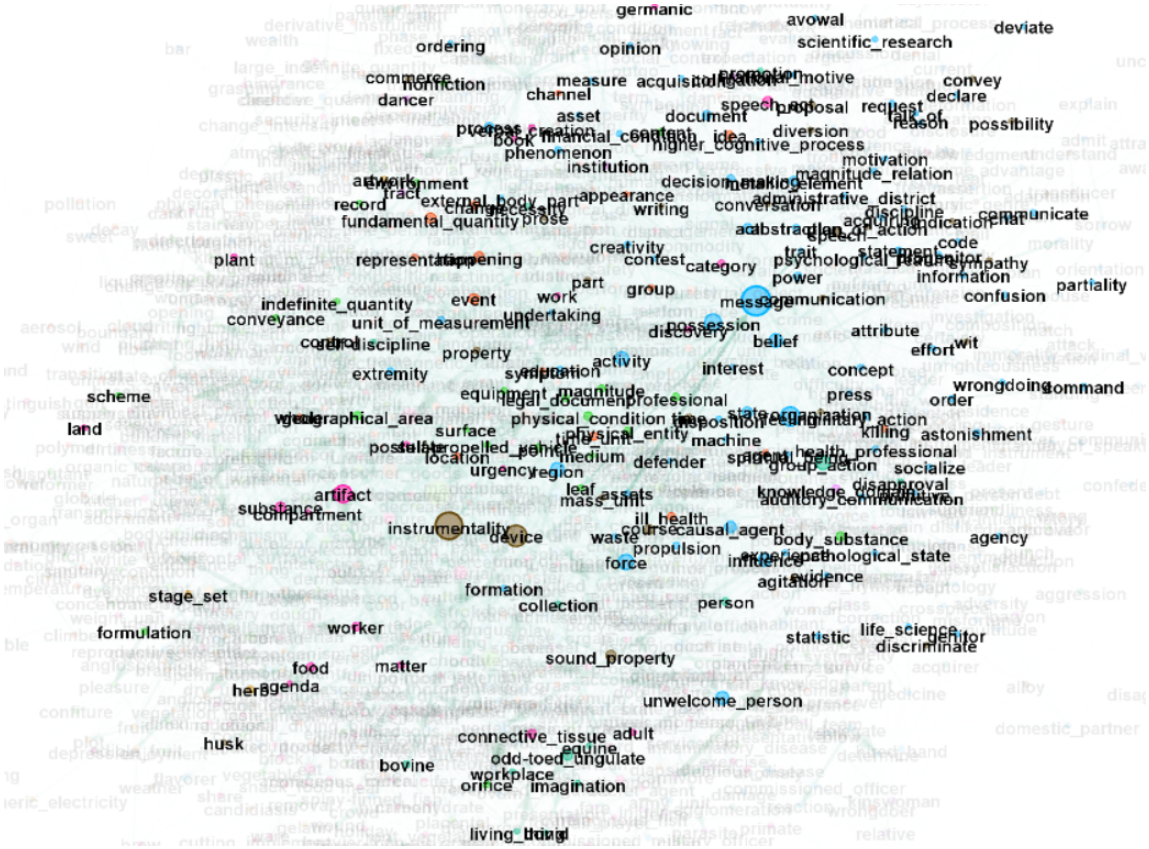


In the network, the node with the label Standard has been taken into consideration.

Observations:

- It can be seen that although it is not a betweenness center, the node is well connected across different communities.
- The number of neighbors is also quite high.
- There is also a good mix of strong and weak associations with the neighbors.
- This highly indicates that the network follows a small-world or even an ultra small-world network. This can be further verified with more exploration.

Topic: Message



In the network, the node with the label Message has been taken into consideration.

Observations:

- It can be seen that it is one of the betweenness centers and is well connected across different communities.
- The number of neighbors is also quite high.
- There is also a good mix of strong and weak associations with the neighbors.
- This also suggests that the network is a small-world or even an ultra small-world network.

Discussion and Conclusion:

- From the Part-of-Speech tagging, we understand that there is a strong relationship between Nouns, Verbs, and Adjectives but could be biased as there is not enough data about other parts of speech.
- The vowels association network follows a heterogeneous network.
- In the Subtopics association, it can be seen that there are some subtopics that have the highest association with others and these are not as well connected as others.
- In the Topics association, the network simulates a real-world network where it acts like a scale-free graph (small world).

Limitations:

Despite all the methodologies and techniques implemented, we found some limitations based on the results we got. The following show the limitations of the project:

- Visualizing and analyzing a high number of nodes and edges in a tool like Gephi made computation a challenge and thus we had to select a subset of dataset to derive results.
- Another limitation was the data acquisition of highly relevant data sources to map its features to the current dataset.

Future Scope:

- The study can entail an application of Machine learning models to predict the target words on the CUE words.
- Including sub-features like the sound of the word, topic modeling, and applications of concepts like community detection, etc on a larger dataset with more features can make this study interesting.

Citations:

- Adaptive Factorization Network: Learning Adaptive-Order Feature Interactions Weiyu Cheng, Yanyan Shen,*Linpeng Huang, Shanghai Jiao Tong University{weiyu cheng, shenyy, lphuang}@sjtu.edu.cn
- Nelson, McEvoy & Schreiber: http://w3.usf.edu/FreeAssociation/AppendixA/index.html
- https://unsplash.com/photos/C5SUkYZT7nU
- https://pixabay.com/photos/abc-alphabet-alphabet-letter-blank-3523454/