A Network Analysis of the Definition of Love.

The Issues Encountered and how I Solved Them

Sally A.M. Hogenboom
11377909
word count: 2110 exlc. references & figurecaptions

12-12-2017

Contents

Introduction

Humans - whether researcher, lay-person, or writer - tend not to agree on definitions of concepts. Similarly, we seem to be unable to reach a consensus on the definition of love. Even dictionaries do not seem to agree. For example, the Cambridge Dictionary (2017) describes love as: "To like another adult very much and be romantically and sexually attracted to them, or to have strong feelings of liking a friend or person in your family." In contrast, the Merriam Webster Dictionary (2017) describes love as: "Strong affection for another arising out of kinship or personal ties.". Even though, in sentiment, these definitions may be similar, there are already obvious differences; Cambridge dictionary describes sexual attraction, but the Merriam Webster does not. Achieving consensus on a definition of a concept is thus very difficult, possibly even impossible. The current study sought to explore the use of Network Analysis as an attempt to syntesize and identify commonalities between definitions. The prior goal being the identification of clusters of words that - together - make up the concept of love. However, to the best of my knowledge such a technique has not been used before, and I thus will discuss the considerations made, the pitfalls of existing functions and protocols, and my solutions to overcome them.

The Datasets

The first decision that I made concerns the combination of available data. I have opted to divide the data between definitions / descriptions made by lay people (Personal Definitions) and writers (Literary Quotes), because writers tend to talk about concepts in a more descriptive sense. I have thus ran the analyses on two distinct datasets. All materials are available in the online appendix [1].

Personal Definitions

I conducted a 1-question survey under my peers (N = 27) asking them to report their personal definitions of love. I did not specify the type of love (e.g., companionate, romantic) as these appear constructs that are only separated by researchers, and not by lay-people in real life. The second source of data came from UrbanDictionary (N = 33). UrbanDictionary allows users to add their own definitions for words. I opted to combine these two sources into the dataset `pd_data` as both sources are from lay people (N = 60).

Descriptives [2]

- Shortest Description: 1 word
- Longest Description: 48 words
- Mean: 8.78 words
- SD: 9.5 words

Literary Quotes

A different source of definitions of love comes from the literature. Writers have been known to describe love in unique ways. I have thus searched the internet for "Quotes of Love". Such searches result in a long list of website quoting famous authors. A selection was made and resulted in the following two sets of data: QuoteDB (N = 100; 2017), and a list of wise, witty, and cynical quotes by Seltzer (N = 96; 2011).

Descriptives[3]

- Shortest Description: 1 word
- Longest Description: 52 words

[1] https://github.com/SHogenboom/NetworkAnalysisDefitionOfLove
[2] Supporting Custom Function: SentenceLenghts
[3] Supporting Custom Function: SentenceLenghts

- Mean: 5.91 words
- SD: 4.71 words

Overview Plots

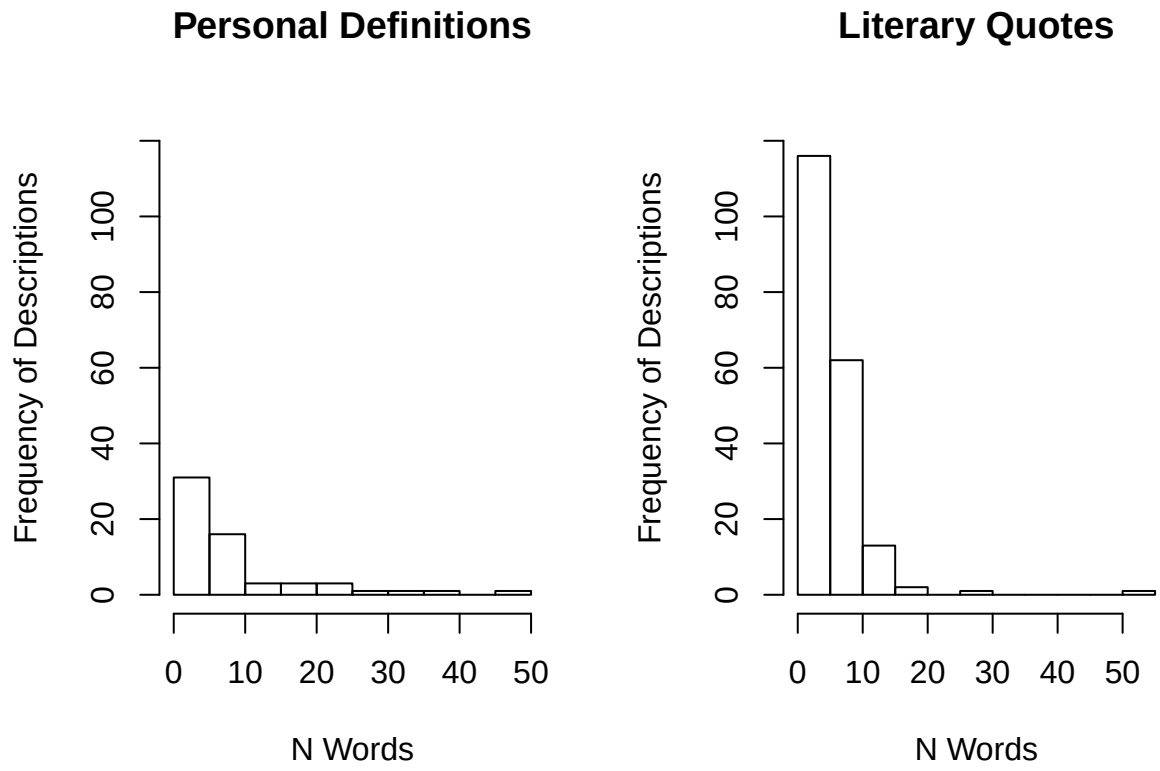

Figure 1. An overview of the total amount of words in the descriptions after removal of stop-words.

Figure 2. [**Left**] Personal Definitions, [**Right**] Literary Quotes. An overview of the Top 50 most frequently used words.[4] Frequency differences are captures by relative differences in the size of the words portrayed.

[4]Supporting Custom Function: WordFrequency

Network 1: Bigrams

Example Network

One of the ways in which the data may be visualized is by plotting bigrams. Bigrams are combinations of two words that directly follow each other. For example consider the statement: "Love is caring deeply for another". Without stop-words (these are removed at the initial processing stage) that sentence looks like this: "love caring deeply". The example sentence contains two bigrams: "love-caring" and "caring-deeply". Plotting these in a directed acylic graphical (DAG) network allows you to follow the nodes, and thereby actually read the definitions. Consider another example statement: "love is feeling deeply emotional.". Combined, these two statements in a DAG would look like this:

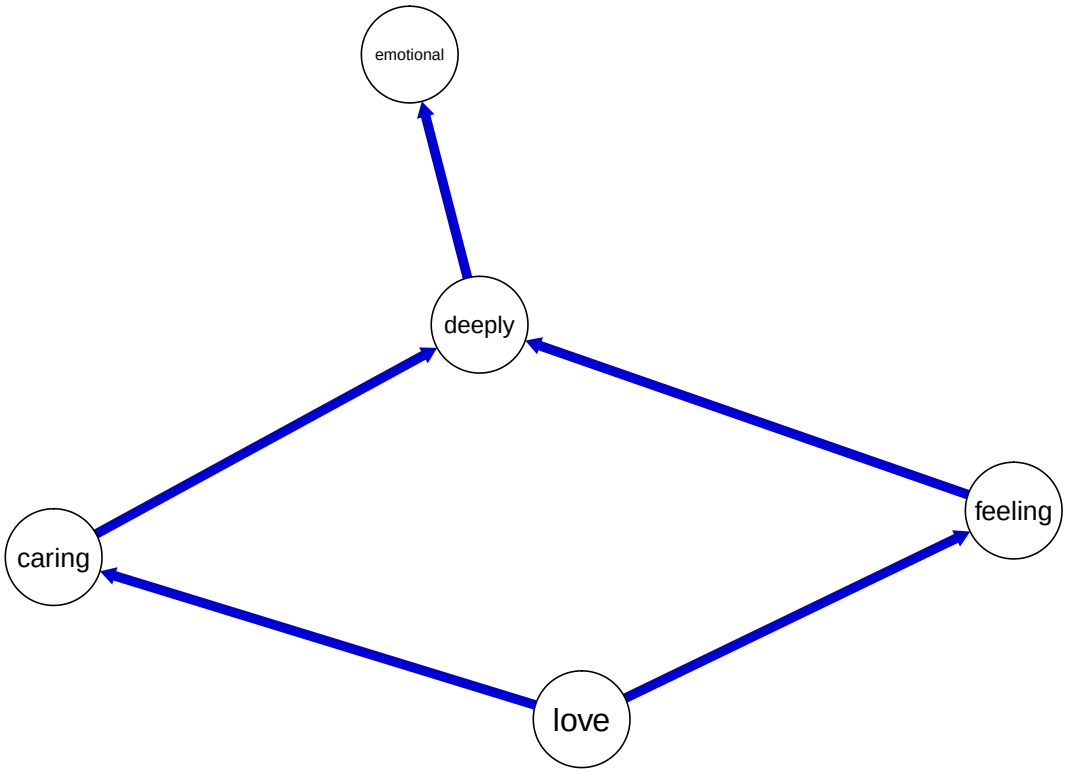

Figure 3. An example of a Directed Acyclic Graphical Network of two example definitions of Love. Following the directed edges (i.e., the arrows) allows you to track back how definitions may have been expressed.

Raw Bigram Networks

The raw bigram networks (similar to the example) of the two datasets are included below. Please note, the network of Personal Definitions contains **685 bigrams**, and the network of Literary Quotes contains **1330 bigrams** - the networks are thus far from legible even if they were printed on a larger plot area. I will address how to deal with such issues in more detail below, however, I first want to consider what happens when minimal adjustments are made to the data:

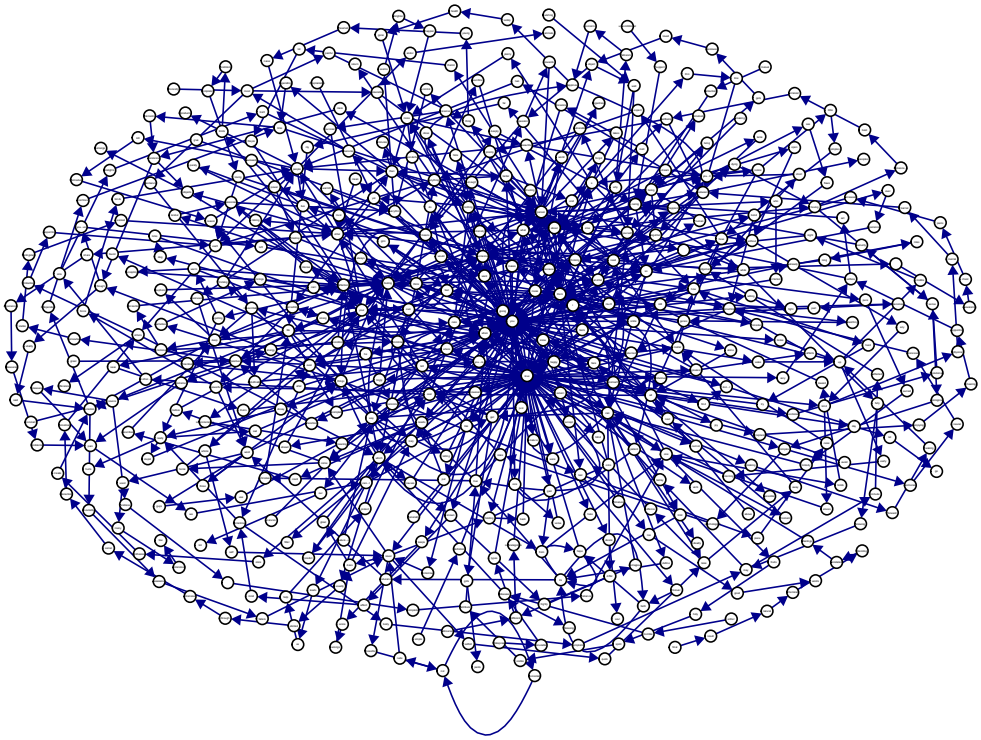

Figure 4. An unadjusted network of Personal Definition Bigrams (sets of two consecutive words). The total number of bigrams included is 685.

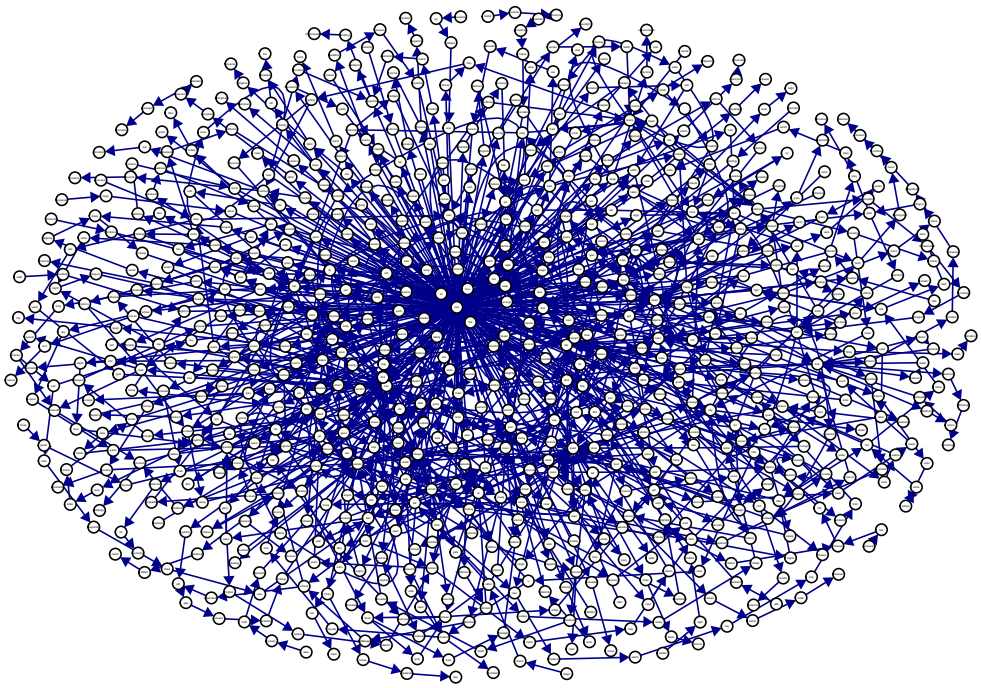

Figure 5. An unadjusted network of Literary Quote Bigrams (sets of two consecutive words). The total number of bigrams included is 1330.

What lessons can be learned?

Removal of stop-words

There exist a great number of manuals on Natural Language Processing, one of which is the TidyText book (2017). This book was initially used as a guideline on how to extract meaningful sections of text (i.e., tokens), however I also noticed a problematic step in their bigram procedure; stop-words are removed after bigram extraction. Let me illustrate why this is problematic. Bigrams are extracted from the original sentences, after removal of punctuation and transformation to lower case. At this stage a sentence may look like: "the feeling of deep care for another human being". Extraction of bigrams (i.e., two consecutive words) will result in a list of "the-feeling", "feeling-of", "of-deep", "deep-care", "care-for", "for-another", "another-human", "human-being". It was then advised to search through the list of bigrams and remove any that contained a stop-word (a library was provided of 1149 words). In the case of the example that would leave: "deep-care", "human-being". Although we may intuitively know that we cannot experience deep care without also experiencing a feeling, this example illustrates how removing stop-words after extraction of the bigrams may remove insightful words/concepts. Consequently, I opted not to remove the stop-words at this stage, but rather before extracting bigrams (through use of the `NLP` package). A second insight was provided by the fact that different packages utilize different stop-word libraries and thus will produce greatly different outcomes. For example, the initially used `tidytext` package included a stop-word library containing 1149 words, in contrast the later used `NLP` package includes a library that contains only 174 English stopwords. One should thus not only consider when to remove stopwords, but also which libraries are utilized as one may in fact remove too many or too little words. For the second set of networks I have opted to remove both the stop wordsfrom the `tidytext` library, as well as those included in the `NLP` library.

Linear relationship N Bigrams & N Descriptions

The first thing I noted was that the larger the dataset (the Literary Quotes data contains more entries than the Personal Definitions), the more bigrams seemed to be included in the network. Although some increase in the amount of bigrams should be expected (illustrating differences between statements), I did not expect the increase to be as strong as it currently is:

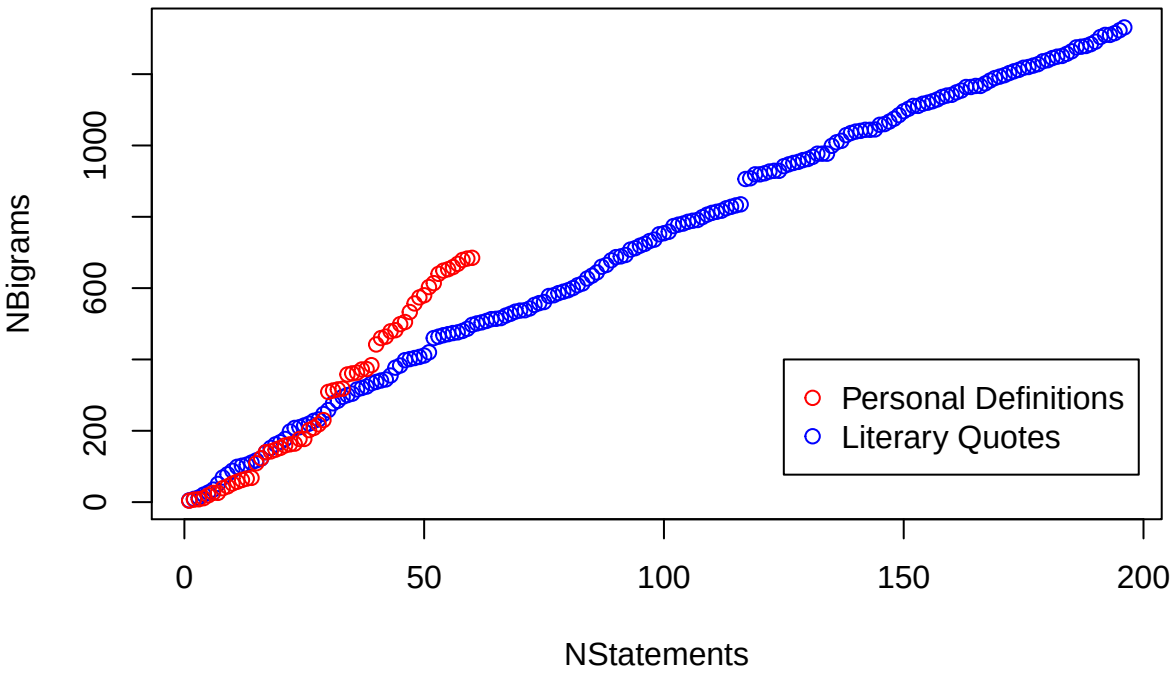

Figure 6. Increase of Number of unique bigrams with the increase of amount of statements in the dataset.

Red circles represent the Personal Definitions, where blue represent the Literary Quotes.

Uniqueness

The increase in number of bigrams with the amount of statements included in the data sample illustrates that there is likely to be relative little overlap between and within statements; a high uniqueness of the data. This assumption is supported by the overview of edge weights (i.e., amount of times a bigram occurs within the dataset):

Table 1. *The frequency of unique bigrams in the sample of Personal Definitions. For example, there were 665 bigrams that occured only once in the entire dataset.*

```
##
##   1   2   3
## 665  17   3
```

Table 2. *The frequency of unique bigrams in the sample of Literary Quotes. For example, there were 1214 bigrams that occured only once in the entire dataset.*

```
##
##    1    2    3    4    5
## 1214  103    8    4    1
```

This yields an unsurprising pattern of centrality measures. I have included a sample from the Literary Quotes dataset to illustrate that, in general, the nodes have low centrality.

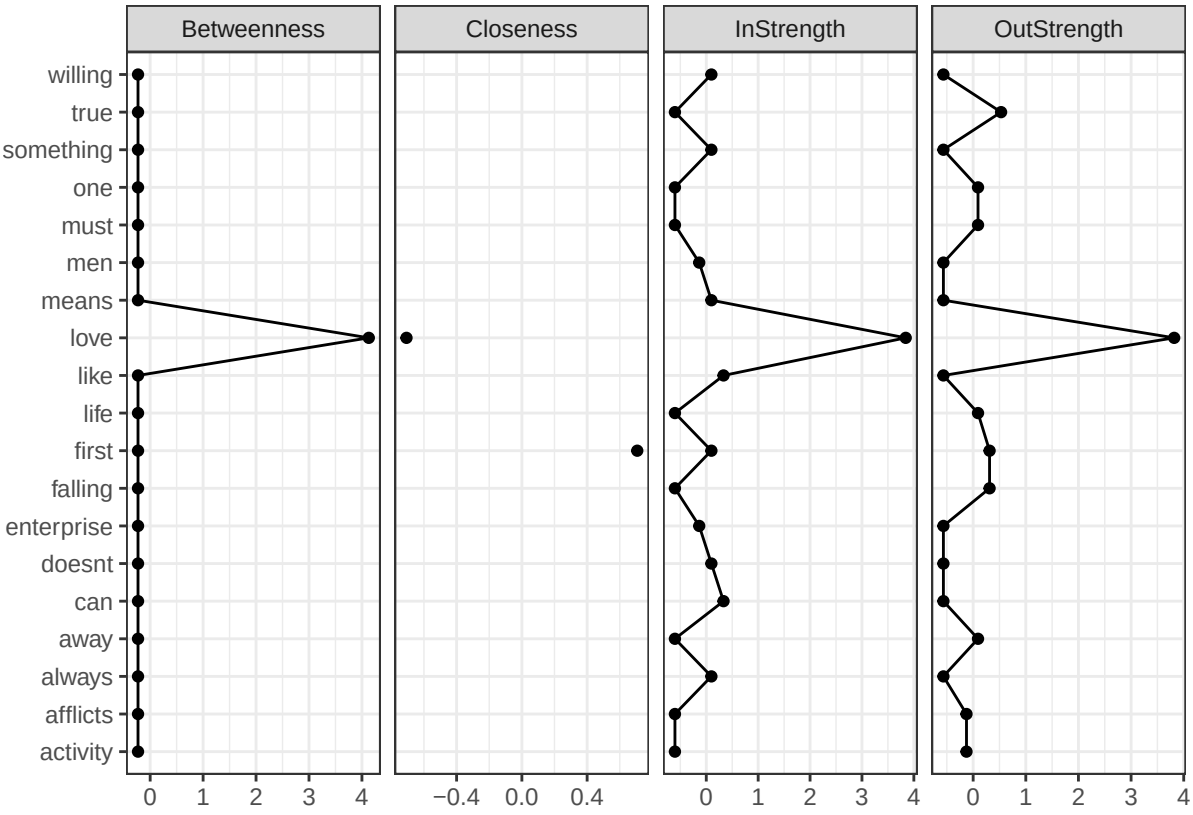

Figure 7. A subset of data from the Literary Quotes Bigram Network showing low centrality for the different nodes (i.e., words).

Some nodes, of course, are more central than others, which is already indicated by the high betweenness and in going edges for the node 'love'. I have calculated the centrality measures for the entire networks to

determine whether, other than 'love', there are more influential nodes.

```
centralTop5(pd_centrality)
```

```
## $`Top5 Betweenness`
##      love   someone    person    giving       can
## 107750.79  41010.97  31819.33  16840.00  15335.94
##
## $`Top5 Closeness`
##      love     youll     youre affection   another
##         0         0         0         0         0
##
## $`Top5 OutDegree`
##    love feeling  person someone     can
##      50      14      14      13       9
```

```
centralTop5(writers_centrality)
```

```
## $`Top5 Betweenness`
##      love       one       can     loved     never
## 428665.11  64234.58  52747.64  39919.21  27859.48
##
## $`Top5 Closeness`
##    true falling   first    love    away
##       0       0       0       0       0
##
## $`Top5 OutDegree`
##   love    one    can people  never
##    175     29     21     15     14
```

These centrality measures show that there are some very influential nodes (e.g., 'love'), however, also that the that number is limited to 1 (steep decline in centrality from node 1 to 2). I would expect that the nodes that are currently the most central, will also be central in the second Network approach. The results also show that tokenization of the nodes has created non-words such as "ll". This is the direct consequence of ineffective tokenization packages, which cause words such as "you'll" to split into "you" and "ll". This particular issue has been resolved in the procedure for network 2 where this type of punctuation "'" is removed before the text is further tokenized.

Conclusion

In an attempt to find common ground between definitions of love I started by extracting and plotting bigrams. Bigrams are unique combinations of consecutive words and thus allow for the visualization of directed networks. However, it was demonstrated that the amount of bigrams increases directly with the amount of descriptions included in the analysis (see Figure 6), and that only the node 'love' has a high centrality in the entire network structure. I have learned that 1) a decision is required as to when stop wordsshould be removed, 2) bigrams of large datasets are little revealing, and 3) the order of text processing must be taken into account to ensure valuable verbal information is not lost.

Network 2: Co-occurrence of Words in Sentences

The first network approach demonstrated that extracting and plotting a directed network of bigrams may result in too much noise and consequently does really allow for inferences. The second approach therefore considers larger tokens: entire statements. Edge weights will now be determined by how often a given word

(i.e., node) occurs in a statement with another word. In addition to including larger tokens, additional attempts are included to aim and decrease the amount of nodes included in the network. Firstly, I will remove both the `tidytext` and the `NLP` stop-words database from the statements. Secondly, a threshold is set to exclude any words / nodes that only occur a relatively low amount of times (N=8) in one single sentence. Thirdly, the data was converted to a binary format as co-occurrence between two words did not occur more than three times in one sentence. Converting the data to a binary format thus allowed the plotting of an Ising Network with eGlasso estimation (gamma = 0; to allow finding of as many edges as possible, and 'AND' == FALSE)[5].

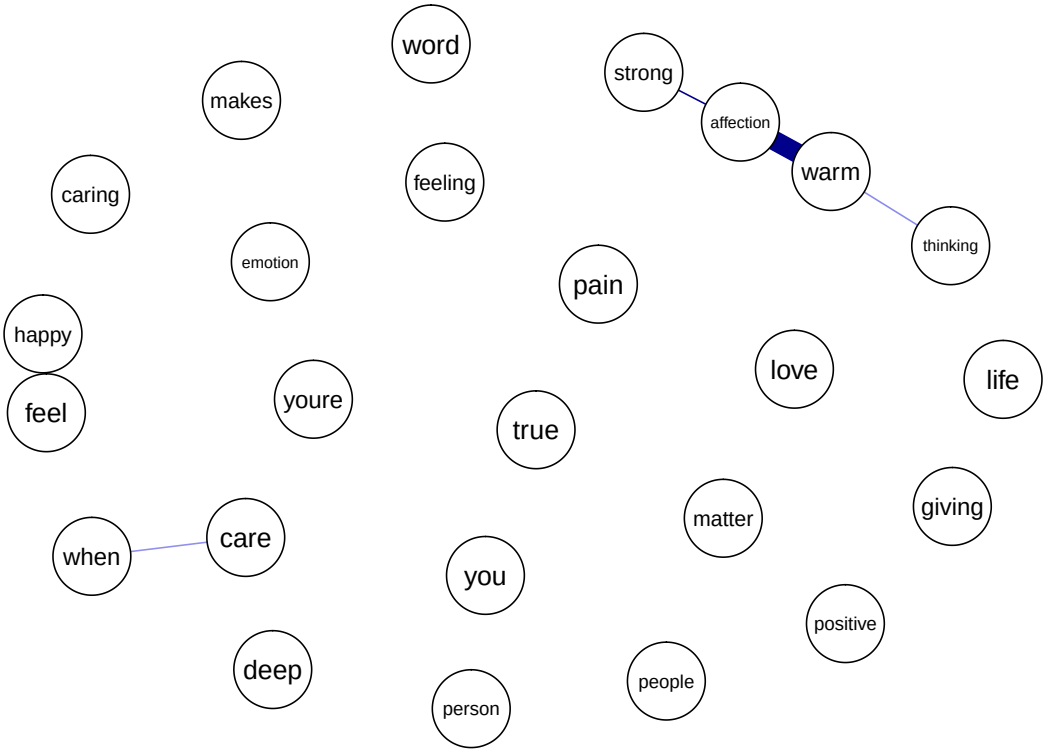

Figure 8. A Binomial Ising Network of the co-occurence of words in the Personal Definitions of Love.

[5]For a detailed discussion of the settings and the implications see Borkulo (2017)

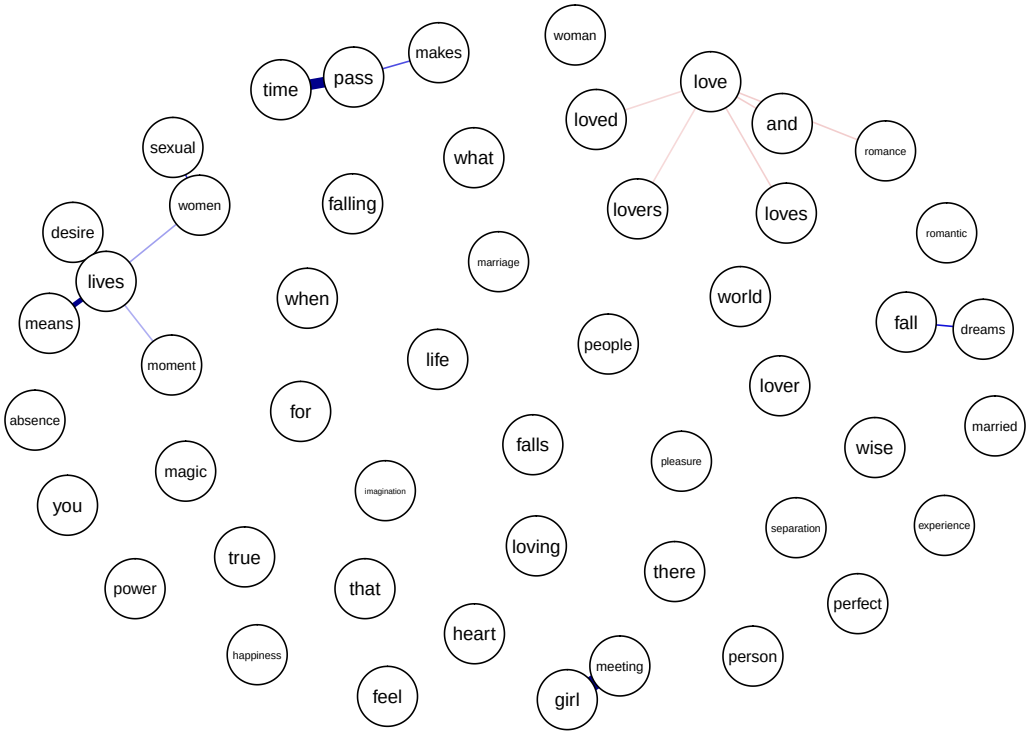

Figure 9. A Binomial Ising Network of the co-occurence of words in the Literary Quotes of Love.

What lessons can be learned?

Uniqueness of Definitions

We saw earlier that the uniqueness of definitions is very high. This could be seen in the direct increase of bigrams with the amount of included definitions, as well as in the low centrality indices. It is therefore perhaps not unsurprising to see that the frequency of co-occuring words, even within an entire sentence, is still very low. Setting a threshold for a single word (i.e., node) to occur in at least three different sentences (thereby increasing the chances of co-occurance with other words), drastically reduces the network from 337 to 25 nodes in the Personal Definitions and from 661 to 50 in the Literary Quotes. Even then, the networks clearly show a lack of overlap as a relatively large number of nodes remain unconnected.

Synonyms

As can been seen from the networks, a lot of words are very similar to one another. For example, in the Literary Quotes network words such as 'love', 'loved', 'loving' etc. are all included. In fact, these nodes are negatively correlated. Indicating that use of the one synonym prevents authors from using the other synonym. Therefore, if one wants to capture any form of text input as a network structure, one should take into account that synonyms may bias the data. As of yet I have been unable to find a Natural Language Processing package that deals with synonyms in an adequate sense. Grouping all synonyms together might, however, reduce the variability between definitions.

Combination of Data

Although I previously argued for a separation of the data, I did want to explore what would happen if the two datasets were combined. The result is plotted below:

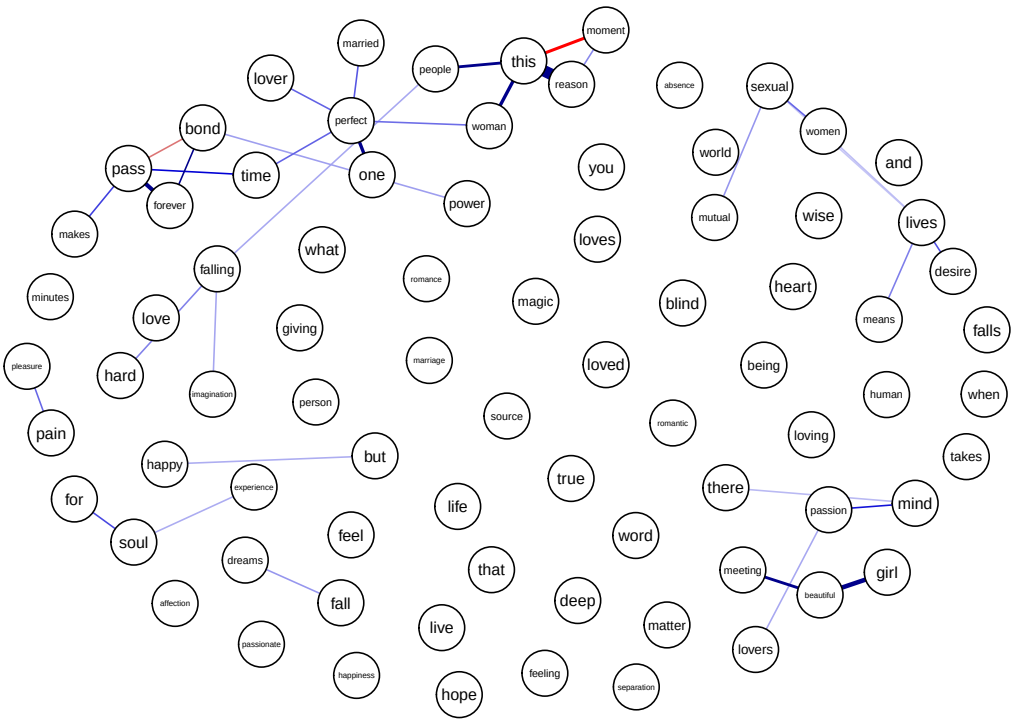

Figure 10. A combined Ising Network of data from the Personal Definitions and Literary Quotes

The combination of data has an interesting and unexpected effect: different edges arise than occurred in the separate networks. This is an interesting effect because it may indicate that there might be more overlap between the two datasets than I expected. After all, only an increase in co-occurring words could have created more edges and nodes to appear.

Conclusion

In an attempt to find common ground between definitions of love I explored the co-occurrence of words in sentences. As becomes evident from the Figures (8,9,10); there is little common ground to be found. Interestingly, even with a larger tokenization, the amount of information visualized by the network is rather little. Or one might argue that is demonstrates exactly the opposite; the lack of visible edges indicates that there is large variation in the definitions of love.

General Conclusion

Two network analyses of the definitions of love have revealed that there is little overlap between lay-persons and writer's definitions of love. Accurate visualization of a verbal structure is further complicated by the lack of clear procedures such as which stop wordsshould be removed, and during what time of the process. Furthermore, analyses of any form of text would greatly benefit from the ability to automatically remove synonyms from the text.

References

1. Borkulo, C.V. (2017). A Tutorial on R Package IsingFit. https://cvborkulo.files.wordpress.com/2017/06/tutisingfit.pdf. Published on: 01/07/2017, Accessed on: 12/12/2017
2. Cambride Dictionary (2017). Meaning of "love" in the English Dictionary. https://dictionary.cambridge.org/dictionary/english/love. Published on: unknown. Accessed on: 05/12/2017
3. Merriam Webster Dictionary (2017). Definition of love. https://www.merriam-webster.com/dictionary/love. Published on: 11/12/2017, Accessed on 12/12/2017
4. Seltzer, L.F. (2011). Love Quotes: The Wisest, Wittiest, and Most Cynical. Published on: 12/02/2011, Accessed on 05/05/2017
5. Silge, J., Robinson, D. (2017). Text Mining With R - A Tidy Approach. http://tidytextmining.com. Published on: 07/05/2017, Accessed on: 01/12/2017
6. QuoteDB (2017). Love. https://www.quotedb.com/categories/love. Accessed on: 05/05/2017

Appendix

An Online Appendix is available at https://github.com/SHogenboom/NetworkAnalysisDefitionOfLove. Items included in the appendix are:

- Personal Definition
- Quote DB - database of Literary Quotes
- Psychology Today - database of Literary Quotes
- FinalAssignment.Rmd - RMarkdown File including all code required to run the analyses
- FinalAssignment.pdf - Final Assingment Report