Table of Contents

Textual Data Analysis - Diary from a Trauma Surgeon: 10 Weeks of Covid-19

1. Project Overview

Text data is a long-established form of evident for argumentation within many humanities disciplines (Arnold et.al 2019). Diary entries for example, helps organise out thoughts and make them apprehensible. We can record daily events, though and feels about certain experiences or opinions. It can be used as a tool to reflect on daily performance and take steps for improvement also in professional life.

In this report, option 4 was selected for textual analysis. It is the diary of Los Angeles-based trauma surgeon Annie Onishi, she documents 10 weeks of her life working inside a hospital during the COVID-19 pandemic. This project will use textual data analysis to analyse, extract, interpret and present insights for this report. Techniques that were used includes sentimental analysis, n-grams, unigrams, bigrams and text clustering. These combined will examine the general mood and trends as well as contributing factors in Annie's diary.

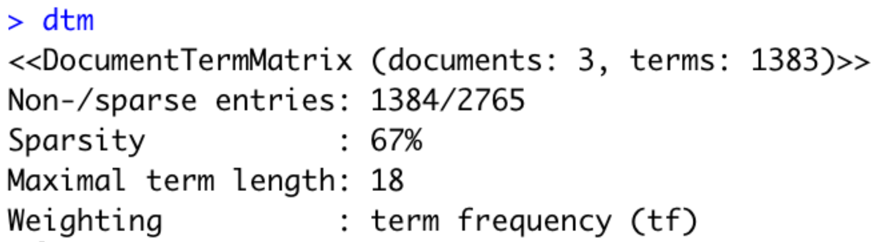
2. Data Processing

Unnecessary words that do not contribute much to text analysis were removed. To further avoid unnecessary computations the text was:

- Converted to lower case
- Stop words such as "I", "like, "and". This step is beneficial in finding aspects from a sentence that are generally described by nouns and emotions are conveyed by adjectives.
- Punctuation was removed and replaced with " "
- Numbers removed
- Strip White Space
- Remove sparse terms
- It was also concluded in this report that removing numbers enhanced accuracy and sparse terms increased accuracy in sentiment analysis and hierarchical clustering and included as part of data processing

3. Exploratory Data Analysis (EDA)

Figure 1 shows the corpus consisting of 3 'documents'. The document term matrix (dtm) contains matrix of 1383 terms that occurred in the corpus. The term frequency rating (tf) is the raw count of the row term found in the document columns. The term has 2785 zeros and 1384 non-zero values. "Sparsity" is 67% indicating much of the matrix contains zeros.



```
> dtm
<<DocumentTermMatrix (documents: 3, terms: 1383)>>
Non-/sparse entries: 1384/2765
Sparsity           : 67%
Maximal term length: 18
Weighting          : term frequency (tf)
```

Figure 1 Sample of Document Term Matrix

Figure 2 shows the most common terms which occurred in the document and plotted on Figure 3.

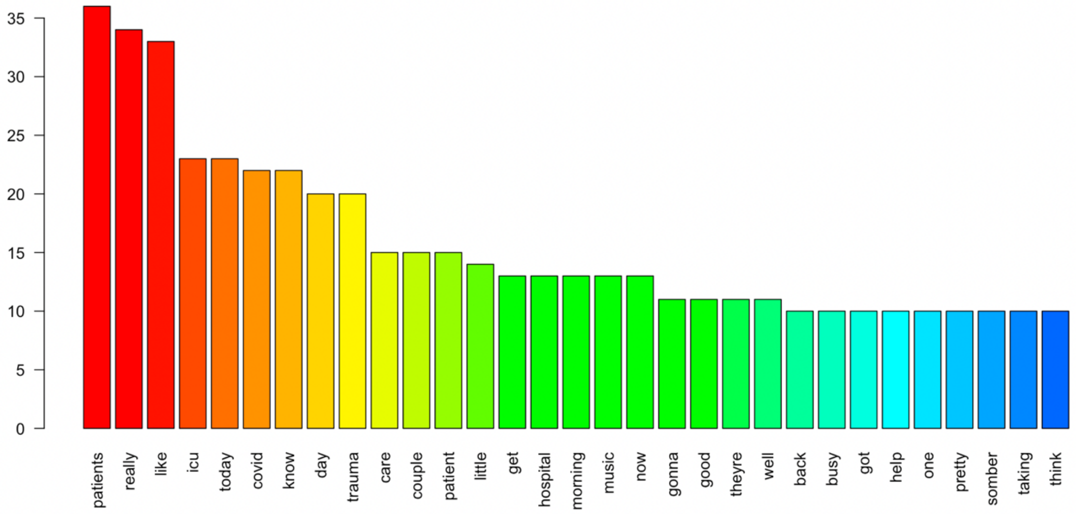


Figure 2 Barplot of Word Frequency

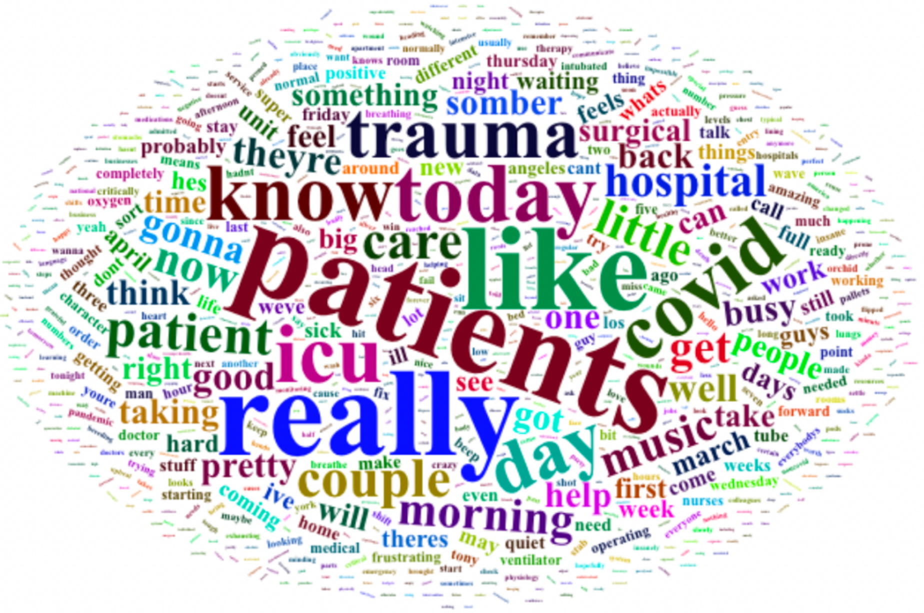


Figure 3 Word Cloud

The most frequent words in the world cloud appear larger in size. The natural language process (NLP) of stemming reduced words to their root or base terms, for example 'patient' and 'patients' and so on.

Upon examination, the top five words included:

Patients, really, like, icu, today, covid.

Word clouds emphasize the frequency of words but not their importance. They also do not provide context. Overall, it does not give much insight to the text so further text analysis methods were used in other R packages.

4. Results and Recommendations

4.1 Unigrams, Bigrams and n-grams

Word tokenisation was performed in this process. Below is figure of top 10 unigrams for the text corpus. These terms give insights Annie's working environment.

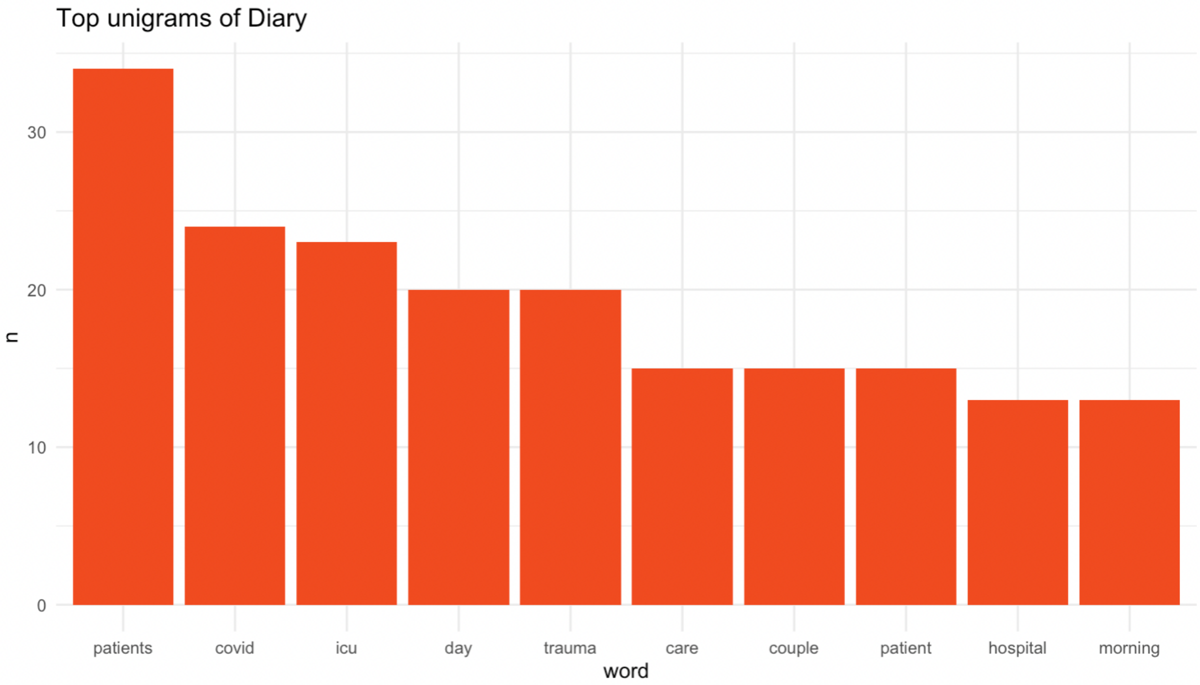


Figure 4 Top 10 Unigrams of Annie's Diary Entry

Some words occur together more frequently. Word pairs are generated from the existing sentences and maintain their current sequences. The other figure below are top 10 bigrams for text corpus.

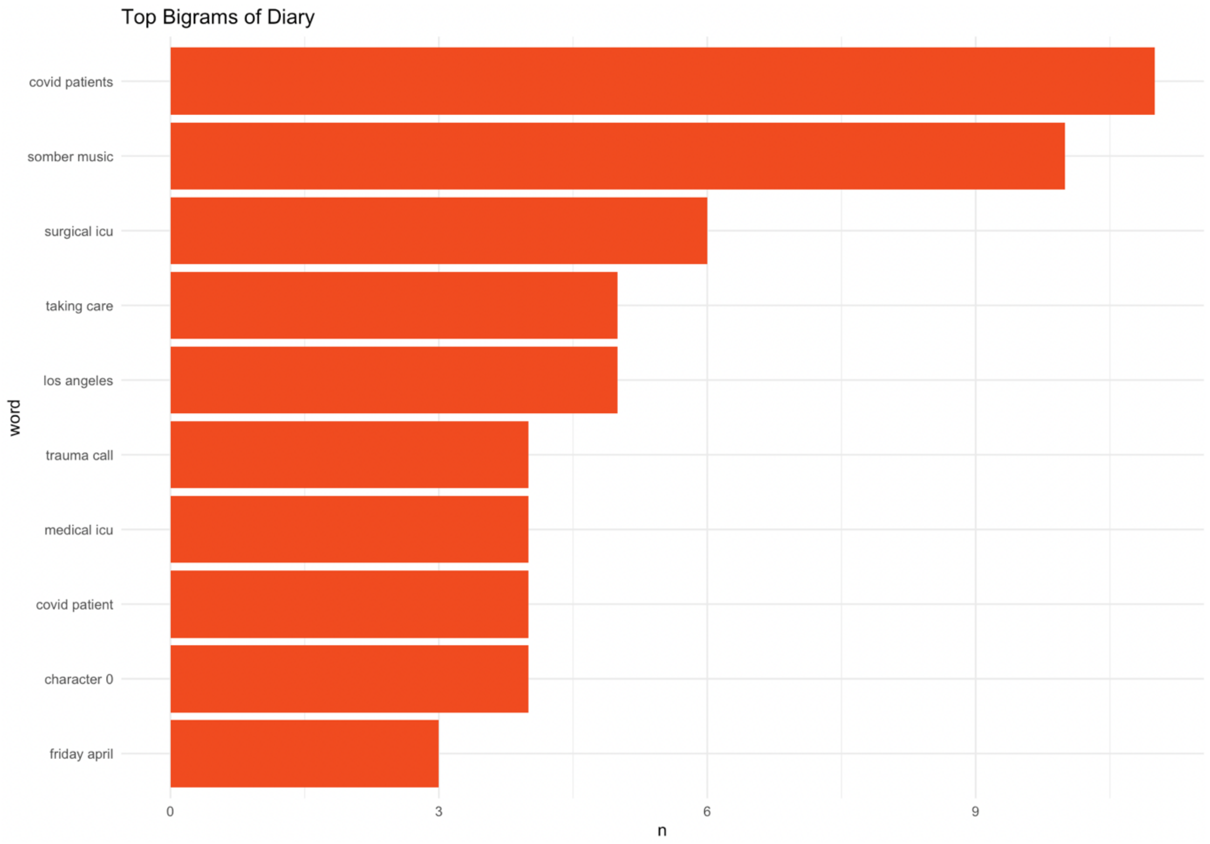


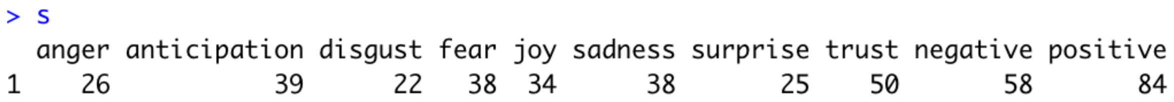
Figure 5 Top 10 Bigrams of Annie's Diary

This result can be used to corelate to the general sentiment of the descriptions in the text.

4.2 Sentiment Analysis:

Sentiment analysis also known as opinion mining is a method of detecting a whether an authors viewpoint on a subject is positive or negative. It is defined as "the process of obtaining meaningful information and semantics from text using natural processing techniques and determining the writer's attitude, which might be positive, negative, or neutral" (Onyenwe et al. 2020).

Levels of sentiment analysis is possible at a sentence level, document level and aspect level. For this case, at a document sentiment is detected from the entire document. In this report, sentiment analysis was quantified in 10 levels throughout the text file:



```
> s
  anger anticipation disgust fear joy sadness surprise trust negative positive
1    26           39      22   38  34      38       25    50       58       84
```

Figure 6 Scores Levels of Sentiment Scores

We can see Annie tried to maintain a positive mood throughout the 10 week period. Sentiments analysis revealed the overall sentiment of 10 weeks of her life working inside a hospital during the Covid-19 pandemic. We can see a ratio of 29:42 of negative vs positive sentiment. 42% of the time it is positive so Annie tries to be optimistic about the experience.

Other notable sentiments: anticipation and trust could also explain the overall positive mood during her 10 weeks.

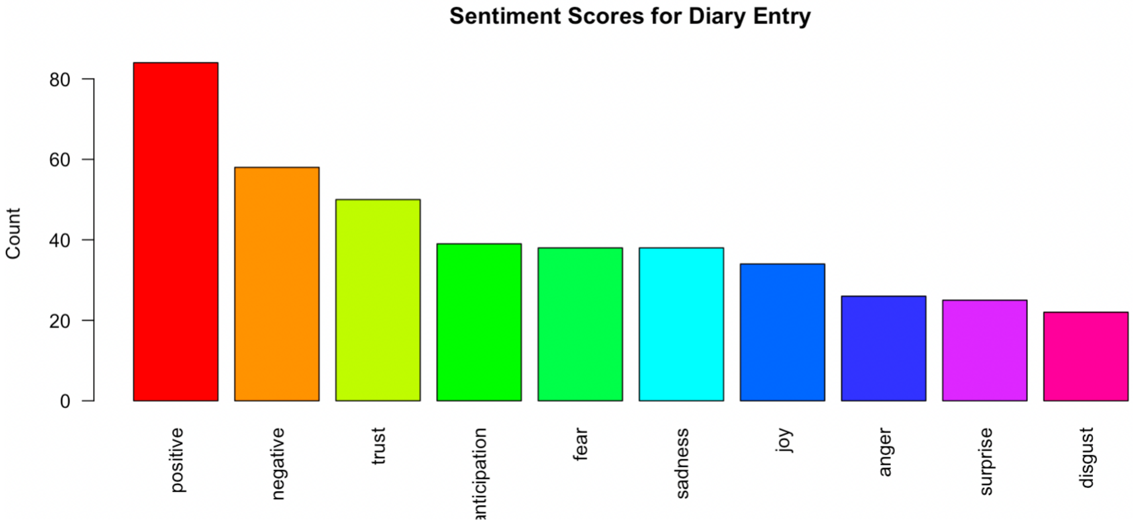
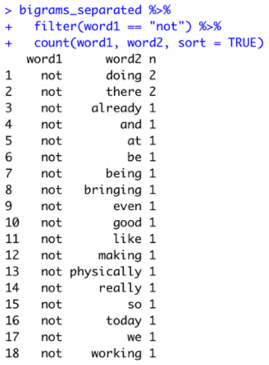


Figure 7 Visualisation of Sentiment Scores

4.3 Resolution bigrams for sentiment analysis

One challenging aspect of document level sentiment classification is considering the link between words and phrases and the full context of semantic information to reflect document composition. It simply counts appearance of positive or negative words. The word’s context does not have much presence. It requires a deeper understanding of the internal structure of sentiments and dependant words (Hassani et al. 2020). For example, the words ‘happy’ and ‘like’ will be counted as positive even in a sentence such as ‘I am not happy’ and ‘I do not like it’.

We can use bigrams to provide some more context in sentiment analysis by can examining how often sentiment-associated words are preceded by ‘not or other negative words. This also be used to ignore or reverse their contribution to the sentiment score.

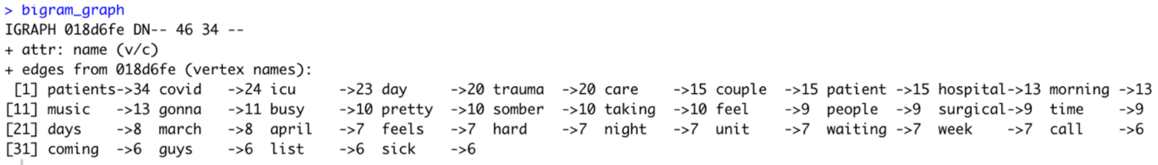


```
> bigrams_separated %>%
+   filter(word1 == "not") %>%
+   count(word1, word2, sort = TRUE)
   word1      word2 n
1    not      doing 2
2    not      there 2
3    not    already 1
4    not        and 1
5    not         at 1
6    not         be 1
7    not      being 1
8    not   bringing 1
9    not       even 1
10   not       good 1
11   not       like 1
12   not     making 1
13   not physically 1
14   not     really 1
15   not         so 1
16   not      today 1
17   not         we 1
18   not    working 1
```

Figure 8 Resolution bigram for sentiment analysis

4.4 Visualising Network of Bigrams

The bigrams network is useful in showing the general idea of the content in the information gathered in the diary.



```
> bigram_graph
IGRAPH 018d6fe DN-- 46 34 --
+ attr: name (v/c)
+ edges from 018d6fe (vertex names):
 [1] patients->34 covid    ->24 icu      ->23 day      ->20 trauma   ->20 care     ->15 couple   ->15 patient  ->15 hospital->13 morning  ->13
[11] music    ->13 gonna    ->11 busy     ->10 pretty   ->10 somber   ->10 taking   ->10 feel     ->9  people   ->9  surgical->9  time     ->9
[21] days     ->8  march    ->8  april    ->7  feels    ->7  hard     ->7  night    ->7  unit     ->7  waiting  ->7  week     ->7  call     ->6
[31] coming   ->6  guys     ->6  list     ->6  sick     ->6
```

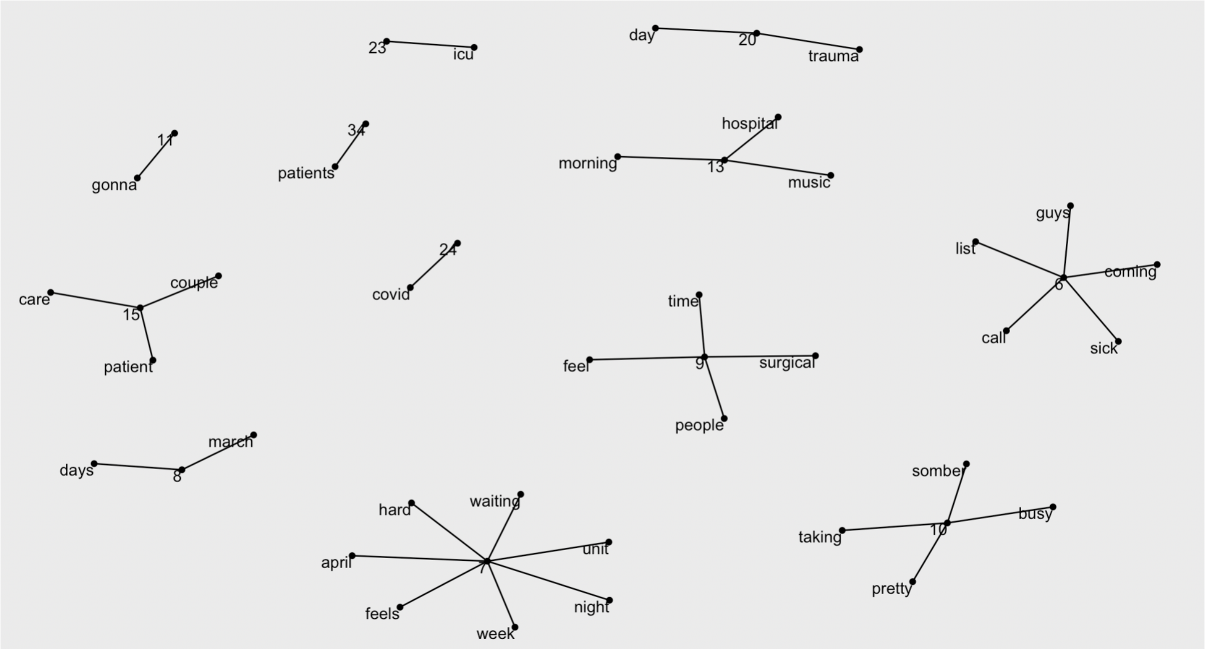


Figure 9 Visualising a Network of Bigrams

5. Text Clustering

5.1 Hierarchical words clustering using dendrogram

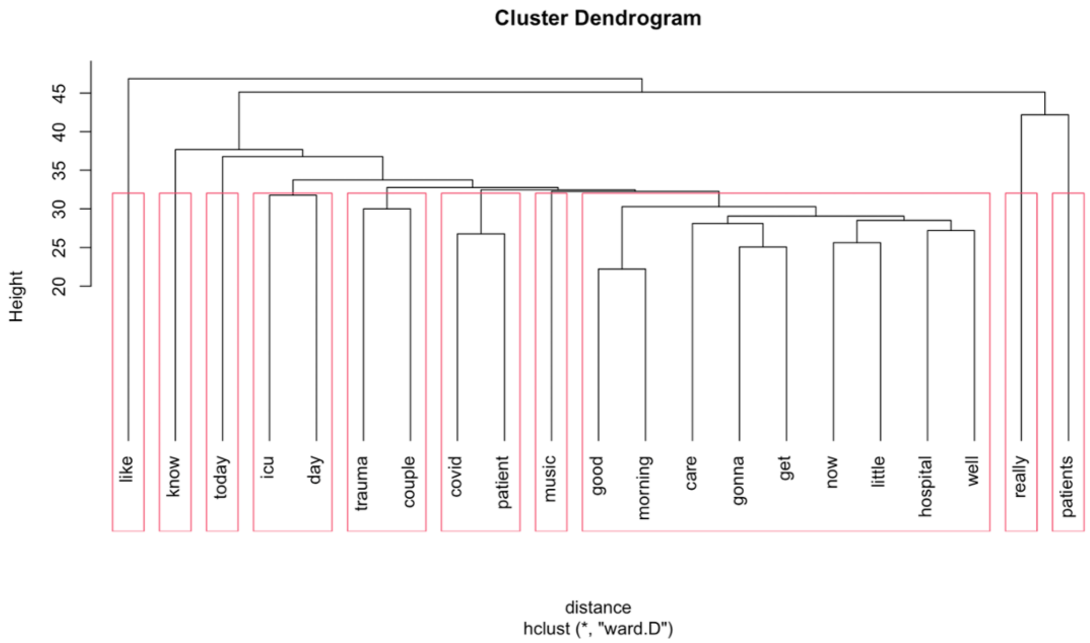


Figure 10 Hierarchal Word Clustering Dendrogram

Words are clustered into word groups through hierarchies; cluster analysis based on dissimilarity of words. One way is to use the Euclidean distance as used in this report. The larger the distance, indicates more dissimilarity between two words. With the distance matric, we can then cluster the words. This is visualised in the dendrogram above with each box around each cluster. 10 clusters were used. It is seen that terms ‘trauma’ and ‘care’ are kept in their own clusters and ‘music’, ‘really’ are kept in their own.

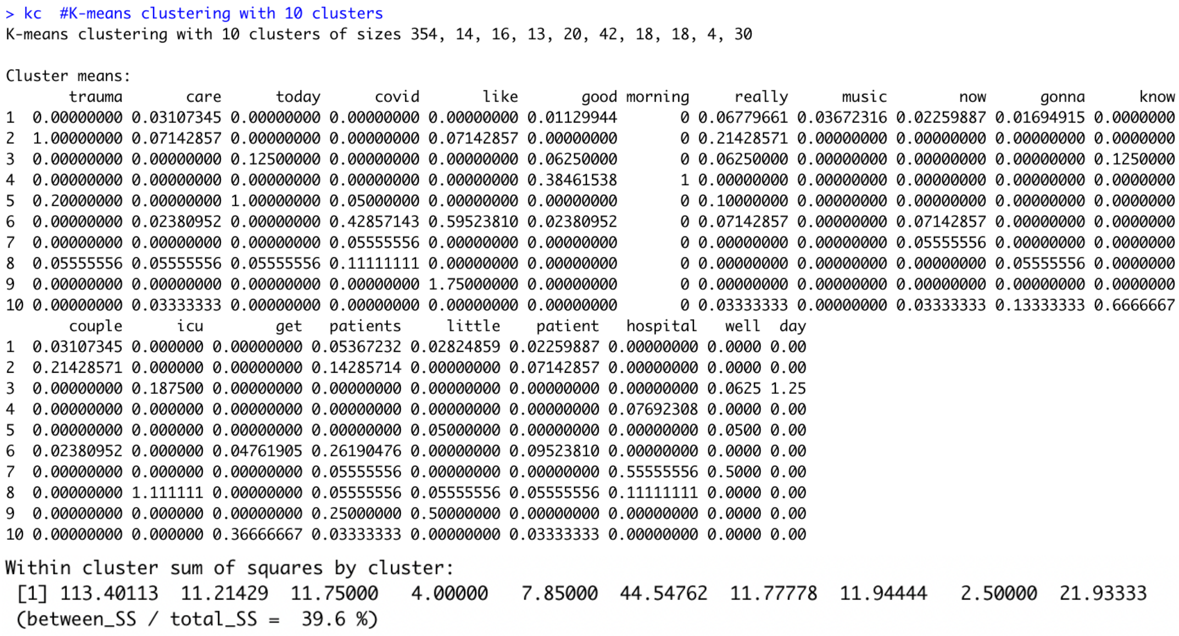
The following was done to improve the model:

- Sparsity percentage was cut down to 0.98 by removing sparse terms

By linking together segments in the diary by the same topic, we can compare how Annie felt about different aspects of her work.

5.2 Non-Hierarchical k-means clustering of words

k-means clustering aims to partition the observations into k clusters in which each observation belong to the cluster with the nearest mean.



```
> kc  #K-means clustering with 10 clusters
K-means clustering with 10 clusters of sizes 354, 14, 16, 13, 20, 42, 18, 18, 4, 30

Cluster means:
       trauma       care      today      covid       like       good morning      really      music        now      gonna      know
1  0.00000000 0.03107345 0.00000000 0.00000000 0.00000000 0.01129944       0 0.06779661 0.03672316 0.02259887 0.01694915 0.0000000
2  1.00000000 0.07142857 0.00000000 0.00000000 0.07142857 0.00000000       0 0.21428571 0.00000000 0.00000000 0.00000000 0.0000000
3  0.00000000 0.00000000 0.12500000 0.00000000 0.00000000 0.06250000       0 0.06250000 0.00000000 0.00000000 0.00000000 0.1250000
4  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.38461538       1 0.00000000 0.00000000 0.00000000 0.00000000 0.0000000
5  0.20000000 0.00000000 1.00000000 0.05000000 0.00000000 0.00000000       0 0.10000000 0.00000000 0.00000000 0.00000000 0.0000000
6  0.00000000 0.02380952 0.00000000 0.42857143 0.59523810 0.02380952       0 0.07142857 0.00000000 0.07142857 0.00000000 0.0000000
7  0.00000000 0.00000000 0.00000000 0.05555556 0.00000000 0.00000000       0 0.00000000 0.00000000 0.05555556 0.00000000 0.0000000
8  0.05555556 0.05555556 0.05555556 0.11111111 0.00000000 0.00000000       0 0.00000000 0.00000000 0.00000000 0.05555556 0.0000000
9  0.00000000 0.00000000 0.00000000 0.00000000 1.75000000 0.00000000       0 0.00000000 0.00000000 0.00000000 0.00000000 0.0000000
10 0.00000000 0.03333333 0.00000000 0.00000000 0.00000000 0.00000000       0 0.03333333 0.00000000 0.03333333 0.13333333 0.6666667
       couple      icu        get   patients     little    patient   hospital   well  day
1  0.03107345 0.000000 0.00000000 0.05367232 0.02824859 0.02259887 0.00000000 0.0000 0.00
2  0.21428571 0.000000 0.00000000 0.14285714 0.00000000 0.07142857 0.00000000 0.0000 0.00
3  0.00000000 0.187500 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.0625 1.25
4  0.00000000 0.000000 0.00000000 0.00000000 0.00000000 0.00000000 0.07692308 0.0000 0.00
5  0.00000000 0.000000 0.00000000 0.00000000 0.05000000 0.00000000 0.00000000 0.0500 0.00
6  0.02380952 0.000000 0.04761905 0.26190476 0.00000000 0.09523810 0.00000000 0.0000 0.00
7  0.00000000 0.000000 0.00000000 0.05555556 0.00000000 0.00000000 0.55555556 0.5000 0.00
8  0.00000000 1.111111 0.00000000 0.05555556 0.05555556 0.05555556 0.11111111 0.0000 0.00
9  0.00000000 0.000000 0.00000000 0.25000000 0.50000000 0.00000000 0.00000000 0.0000 0.00
10 0.00000000 0.000000 0.36666667 0.03333333 0.00000000 0.03333333 0.00000000 0.0000 0.00

Within cluster sum of squares by cluster:
 [1] 113.40113  11.21429  11.75000   4.00000   7.85000  44.54762  11.77778  11.94444   2.50000  21.93333
 (between_SS / total_SS =  39.6 %)
```

We want cluster variability to be low and higher cluster distance percentage to increase.

- Hierarchical clustering using ward D method
- Group outcome into 10 clusters gave the best result
- Cluster-to-cluster distance/variability of 39.6% in the data can be explained by the 10 groups
- Elbow method was used to determine the best number of clusters

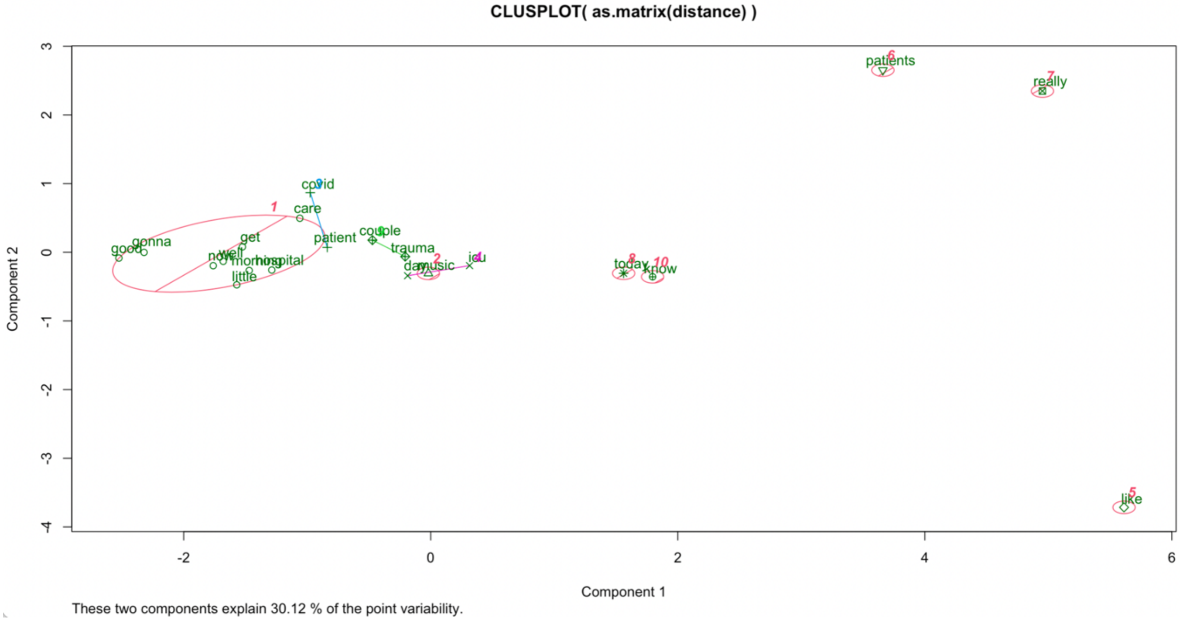


Figure 11 Text Clustering using k-means

6. Potential applications of text analysis in other contexts

Text mining is one of the most critical ways of analysis and processing unstructured data that forms 80% of the world data. It incorporates and integrates the tools of information retrieval, data mining, machine learning and statistics which can be applied in multidisciplinary fields. Other potential applications of text analysis in other contexts include cybercrime, healthcare, and finance.

Text mining and cyber crime:

In cybercrime, text mining techniques are an effective way to detect and predict criminal activities. It primarily focuses on capturing data from chat rooms and social networking sites to apply text mining algorithms.

There have been different approaches to capturing data from 'bag-of-words' approach to detect cyber-predator communications to leveraging communications theory to develop more sophisticated features for input of the classifier.

Dewes et al. 2003, uses multi-layered approach for capturing web chat from various sources including Internet Replay Chat (IRC) and web-based (HTTP and Java) chat systems is used. This is by casting a wide net and capturing all network traffic that passes though a particular router. Several filters are applied to the chat and non-chat traffic. Some other researchers focus on technical difficulties encountered when trying to parse chat log data. (Tuulos & Tirri 2004).

Text mining in healthcare

Recently, text mining tools have been utilised in healthcare research. It is to find patterns in doctors reports identifying patterns in patient data. Emerging concepts precision health and learning healthcare system which uses artificial intelligence and machine learning methods to develop patient models and personalised predictions of diagnosis and care.

Several research studies focus on the processing of textual information available in healthcare datasets. An example of a study by Cerrito and Cerrito 2006 analysed electronic medical records from emergency department of a hospital over a six month period using text mining. They found that similar complaints were treated different depending on the physician on call. This affect quality of healthcare and costs. Therefore, text-mining prior to expert treatment can provide physicians on call will help in providing patients optimised treatment. This can lead to the development of concepts protocols to deliver personalised treatment.

Text mining applications in finance

Finance is a major sector that is continues to see growth in every sector. Text mining is an emerging field of research in the domain of finance (Gupta et al. 2020). The analysis of large volumes of financial data is both a need and an advantage for corporate government and general public and can benefit from using text-mining. Some widely uses techniques in the analysis of textual data are:

- Sentiment Analysis (SA)

This technique extracts the underlying opinions within textual data. In finance stock market prediction is one of the applications in which SA has been used to predict stock market trends and prices from financial news articles (Al-Natour & Turetken 2020). The main objective of SA are broadly divided in two categories: emotion recognition and polarity detection. Where emotion detection is focuses on a set of emotion labels and polarity detection is bases on a classifier approach with discrete outputs (Cambria 2016).

- Ontologies

Ontologies-based text mining serve domain knowledge and the form of relations between entities (perceptions in their level order). An example is WorldNet that uses Ontology which contains 110,000 unique connections and 150,000 words. Feldman and Hirsh 1996, developed a system called FACT (Financial Associations in Collections of Text) where it discovers the co-occurrence and associations of the terms in text corpus.

Developments in big data analysis can be exploited successfully in text-mining applications in many contexts. Obvious extensions to this report include text analysis of additional document corpora, and more sophisticated use of text analytics in handling and predicting a significant amount of data. There is also room for future improvement of text analysis such as combining numerical data with textual data to yield better predictions.

Appendix

R code

References

- Al-Natour S & Turetken O, 2020, 'A comparative assessment of sentiment analysis and star ratings for consumer reviews.' International Journal of Information Management, vol 54 https://doi.org/10.1016/j.ijinfomgt.2020.102132

- Cambria E, 2016. 'Affective computing and sentiment analysis.' IEEE Intelligence Systems vol 31, no.2, pp 102-107. https://doi.org/10.1109/MIS.2016.31

- Cerrito, P.B., & Cerrito, J.C. 2006. 'Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs', Data Mining and Predictive Modelling.

- Dewes C, Wichmann A & Feldmann, A., 2003, 'An analysis of Internet chat systems.', IMC'03: Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement , pp. 51– 64.

- Feldman, R. & Hirsh, H., 1996, 'Mining Associations in Text in the Prescence of Background Knowledge', KDD.

- Gupta, A., Dengre, V., Kheruwala, H.A. & Shah, M., 2020, Comprehensive review of text-mining applications in finance. Financial Innovation, vol **6, pp** 39. https://doi.org/10.1186/s40854-020-00205-1

- Hassani, Hossein & Beneki, Christina & Unger, Stephan & Mazinani, Maedeh & Yeganegi, Mohammad. (2020). Text Mining in Big Data Analytics. Big Data and Cognitive Computing. Vol 4. No 1

- Onyenwe, I., Nwagbo, S. % Mbeledogu, N. *(2020).* The impact of political party candidate on the election results from a sentiment analysis perspective using #AnambraDecides2017 tweets. Social Network Analysis and Mining, vol 10, pp 55. https://doi.org/10.1007/s13278-020-00667-2

- Taylor, A., Ballier, N., Lissón, P., & Tilton, L. 2019. Beyond lexical frequencies: Using R for text analysis in the digital humanities. *Language Resources and Evaluation, vol 53*, no 4, pp 707-733. doi:https://doi.org/10.1007/s10579-019-09456-6

- Tuulos, V.H., & Tirri, H. 2004. Combining Topic Models and Social Networks for Chat Data Mining. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, pp 206-213.