output: bookdown::pdf_document2: template: templates/template.tex bookdown::html_document2: default bookdown::word_document2: default documentclass: book #bibliography: [bibliography/references.bib, bibliography/additional-references.bib] editor_options: markdown: wrap: 72 # Text Mining, Natural Language Processing, and Sentiment Analysis —

0.1 and here is a bit of text mining i am doing, basic frequency

```
knitr::opts_chunk$set(
    echo = FALSE,
    message = TRUE,
    warning = TRUE
)
library(tidyverse) #for various data manipulation tasks
library(tidytext) #for text mining specifically, main package in
↪  book
library(stringr) #for various text operations
library(gutenbergr) #to access full-text books that are in the
↪  public domain
library(scales) # for visualising percentages
library(readtext) # for reading in txt files
library(wordcloud) # for creating wordclouds

library(syuzhet)
```

The focus group discussions provided a lot of feedback for how the participants experienced their interactions with the chatbots, and how the CEPEH team can improve them, improve the design and development processes, and improve uptake and sharing.

One method of analysing this data is with use of text mining and data manipulation, creating word clouds, sentiment analysis, and using a model which can distinguish the unique themes in text, and highlights for us what text is used to create these themes.

Therefore, we have created a model to allow efficient and intelligent analysis of this open/free focus group data.

0.1.1 Tokenising

Firstly, we tokenised the words from the FGDs. A Token is “a meaningful unit of text, most often a word, that we are interested in using for further analysis”. For each word we give it a property that we can call upon later.

The data manipulation for this included removing punctuation, converting to lower-case, and setting word type to word (and not such types as “characters”, “ngrams”, “sentences”, “lines” etc)

Stop words

The model then removed words with meaningless function. These are called stop words. Words like “the”, “of” and “to” are the most frequent words found, technically, but are of little interest to us.

We also created a custom list of stop words for CEPEH. We know participants may mention other objects, and the list was as followed: found; chatbot; chatbots; presentation.

The data was ready for analysis by the model. We ordered it to find the most frequent words. Below is a table with the 6 frequently occurring words, showing how Stop words have now been filtered.

word	n
information	11
helpful	8
understand	8
idea	7
ideas	7
lot	7

This word list can then be used for sentiment analysis, (see *Sentiment Analysis* section), in addition to frequency of words.

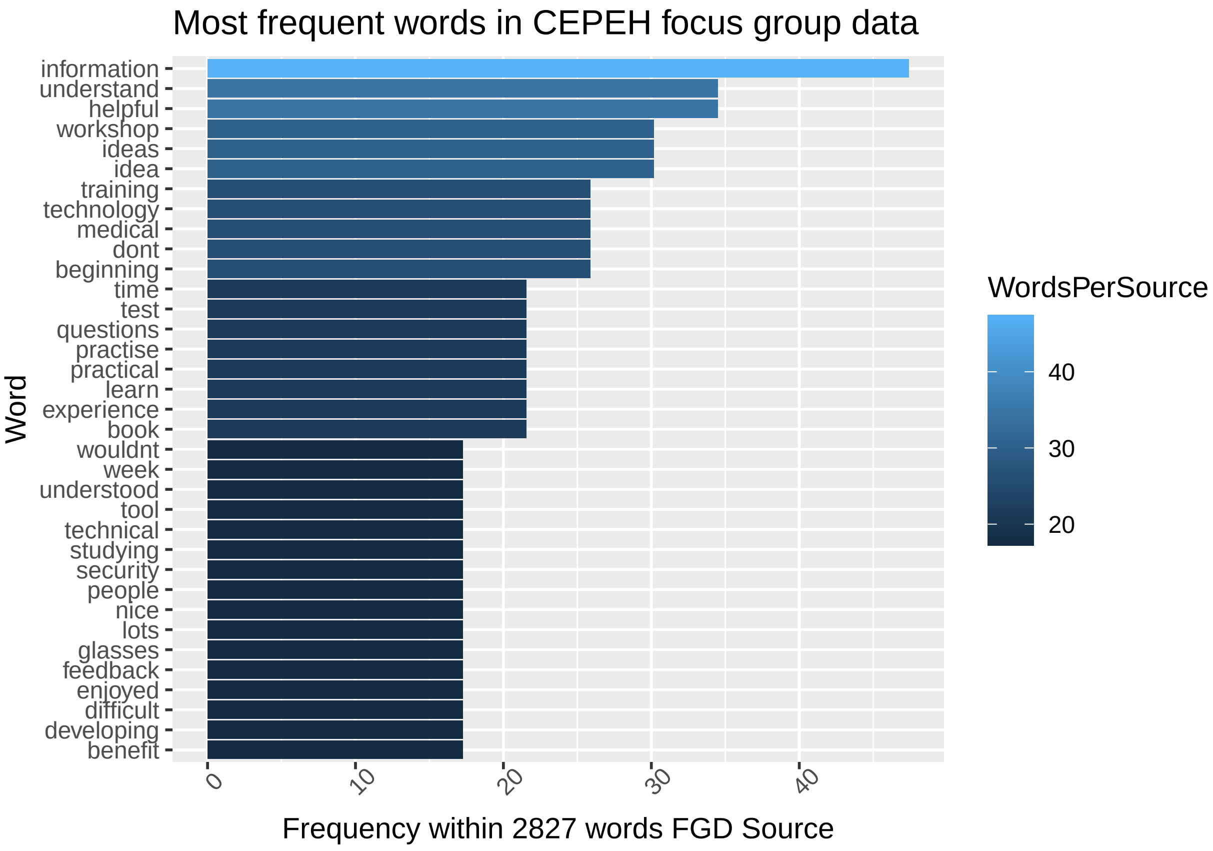
0.1.2 Plotting word frequencies - bar graphs

Normalised frequency

With this information a list of top words from the participants in the FGD can be rendered and after some modifications, a graph of the top 20 words is produced, with better aesthetics. This is a better way to understand this data, and the axis can be normalised for the frequency of occurrences in accordance with the source text. The raw text had 2827 words in total. Therefore we can mutate the ratios to reflect this.

Plotting normalised frequency

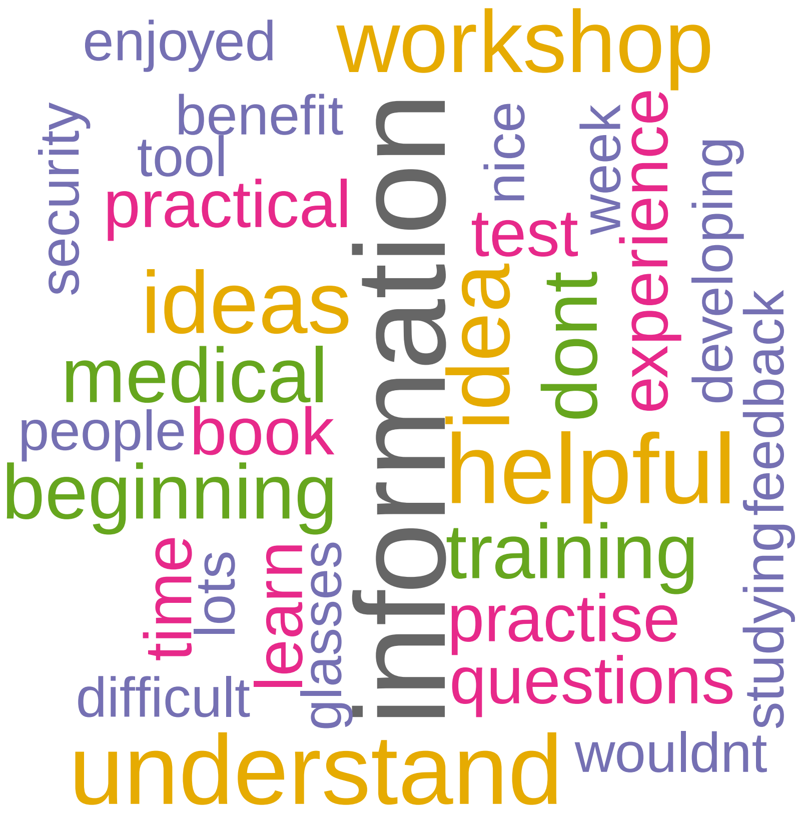
Now we can plot, for example, the 20 most frequent words when normalised by the source text.



In summary, this understanding of frequent words can help to understand common concurrences and extrapolate to a larger audience. If scope and impact of CEPEH chatbots increased we can understand the type of themes and trends may occur, based on such FGD analysis.

0.1.3 Word clouds

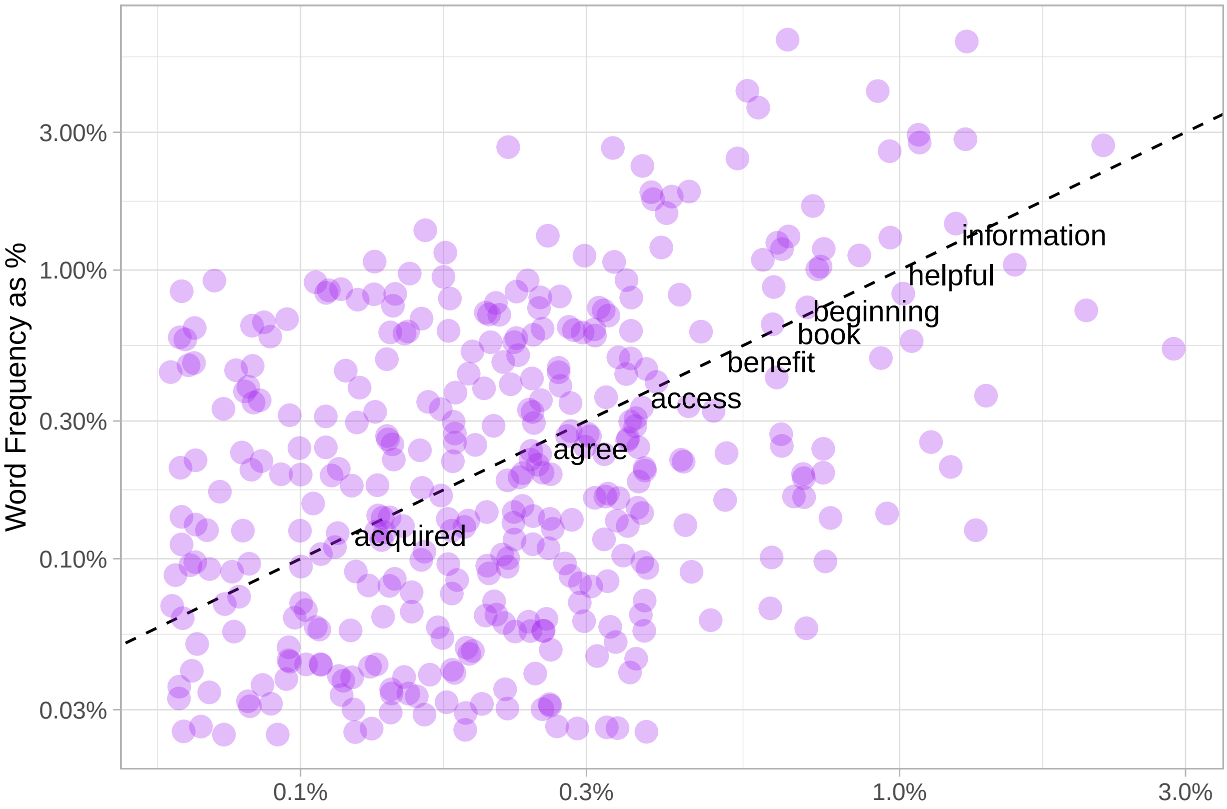
To visualise the most frequent words in another format, below is a word cloud which presents the word size to indicate the frequency- words that occur more often being displayed in a larger font size. This has a normalised data frequency in accordance to the FGD source document analysed.



We understand the context has been reduced for each word. However, in general there can be categorised positive/negative words from the word cloud: Positive words are- benefit, practical, nice, helpful, learn, ideas, and enjoyed Negative words are- difficult, test (who likes a test?), don't, and 'lot' may be negative if there is a 'lot' of information.

The vocabulary of Texts

Here is a graph that has plotted the words in places depending on the word frequencies. Additionally, colour hotspots shows how different the frequencies are - darker items are more similar in terms of their frequencies, lighter-coloured ones more frequent in one text compared to the other.



0.1.4 Sentiment analysis

What is the sentiment of all participants? What is types of emotional words are being used? The preparation of these words has some use in understanding the frequencies, but their emotional valence are not compared. The table above has the word *'helpful'* which has a positive connotation, however there are 386 words, with many having several occurrences.

As the table below shows. the FGD data has been analysed for sentiment of each word, and has been calculated to have 62 positive emotional valence of words, with 24 negative valence of words. These are from a **Bing sentiment lexicon** which is the most popular English language dictionary.

negative	positive	total_score
24	62	38

Unfortunately, there is little research using sentiment analysis for chatbot related focus group results that can help to understand the scoring found. However, on a basic interpretation the higher the score the better the chatbots were discussed in the FGD's. A score of 72% (62/(24+62))) would be in 3/4th quartile in distribution of sentiment distribution. Alternatively, 62/24 = 2.58 would state the ratio that for every 1 negative word recorded, there were 2.58 positive words recorded.