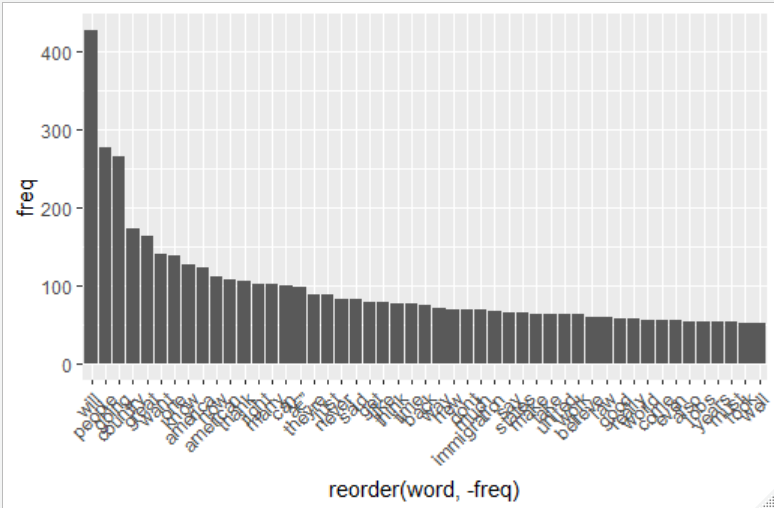
Text mining exercises

<https://rstudio-pubs-static.s3.amazonaws.com/265713_cbef910aee7642dc8b62996e38d2825d.html>

I did this example instead, similar to the one on GDrive but has the files included so I used the same files as they did in the example

The example has a good summary of the basics for the text mining techniques at the end

Text mined 11 Trump Speeches saved as text files, cleaned up the text to leave just useful words, and then checked the frequency of the 20 lowest and 20 highest frequency words. I firstly used the tm library, and then the ggplot2 library to plot the frequencies in a bar chart. After that I checked the correlation between the words country and American as they tend to appear frequently in the texts.



$country

nothing cities countries jobs

0.95 0.94 0.94 0.92

come biggest donors second

0.91 0.90 0.90 0.90

begin border plan crimes

0.88 0.88 0.88 0.87

globe meant thousands means

0.87 0.87 0.87 0.86

workers also despite take

0.86 0.85 0.85 0.85

$american

restore task fair budget

0.97 0.93 0.92 0.91

cycle new promises dollars

0.89 0.89 0.89 0.88

finally millions national tens

0.88 0.88 0.88 0.88

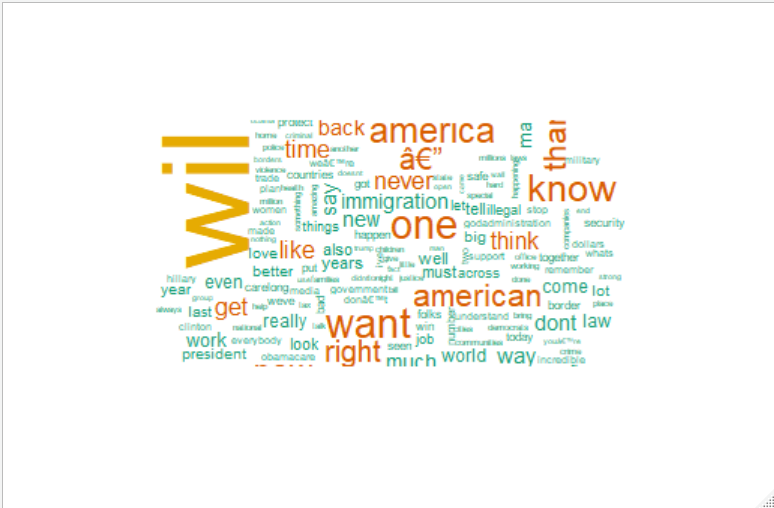
foreign middle justice program

0.87 0.87 0.86 0.86

break joining united

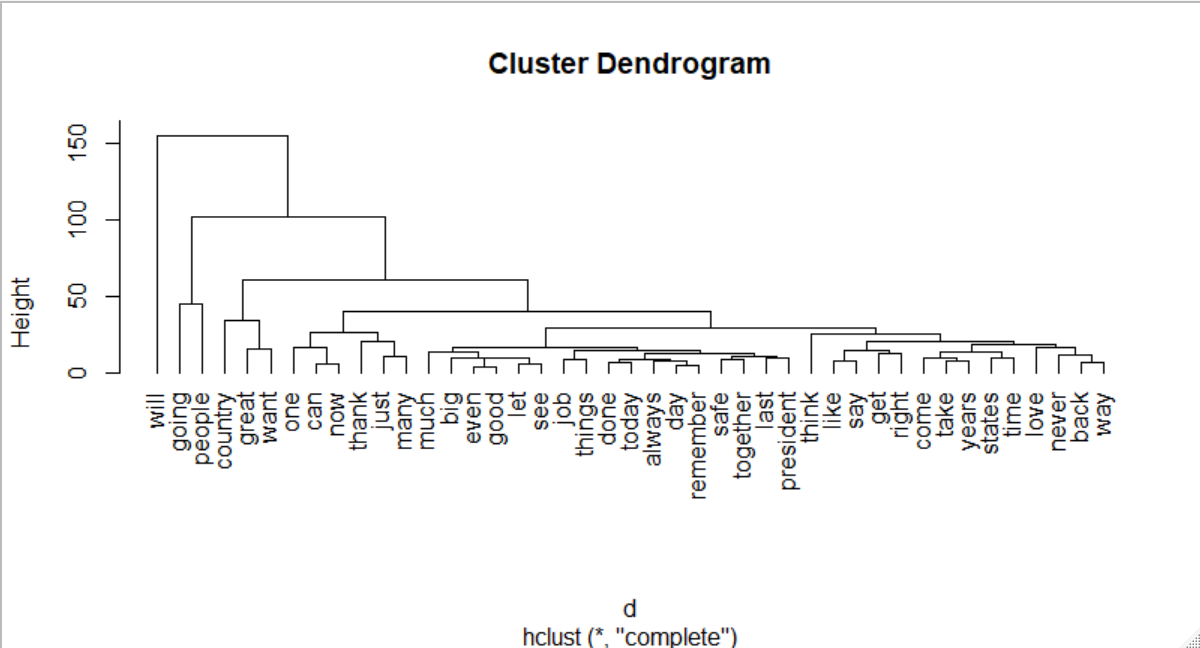
0.85 0.85 0.85

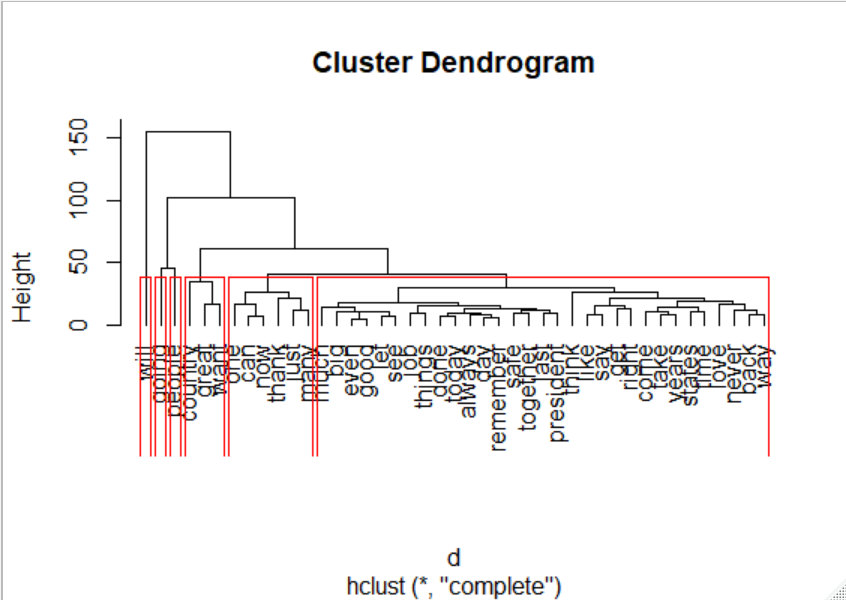
Next up are words clouds to visualise the most frequent terms, using the packages wrdcloud and RColorBrewer



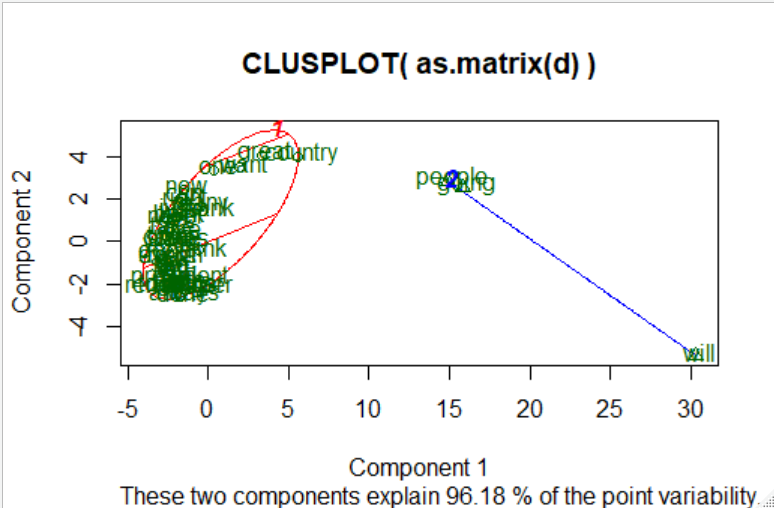
Clustering:

Removing infrequent words from the text before clustering.





K-means clustering



Got a different result to the example, got inverse numbers and a mirrored plot, not sure what went wrong