November 30 Class Exercise

Compare at least three text documents (you may use the 5 chapters of Luke in the folder on Luke_Mining folder on LearningHub, or any other documents of your choice)

1. Remove all English stopwords as well as any other common words that you do not want in your evaluation
2. Remove numbers, extra spaces, punctuations.
3. Your evaluation should ignore case
4. Prepare a chart that shows the frequency of the words across all documents (top 20 words only). Save the image
5. Prepare a word cloud with no more than 150 words. Only words that occur more than twice should be included. Save the image
6. Prepare a chart that shows the frequency of the words by document. (same words you used for # 4). Save the image

Upload your files to LearningHub

Note: to create a Corpus from the text files, put the files in a folder in your working directory and use

corpusName <- Corpus(dirSource("your folder") where corpusName is the name for the object you will create and dirname is the directory where your text files are stored.

The Luke data came from here: https://www.gutenberg.org/cache/epub/6529/pg6529.txt

other resources here: https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf