Lab 8 Exercises - Text Visualisation.

Look at the examples folder to find Five Main Steps.
This is a pdf of the IHaveADream Jupyter Notebook, which is an implementation of a post on http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know. The R version works, but the Markdown comments aren't shown, so look at the .PDF for those.

The visualisations shown here are based on Martin Luther King's "I have a dream" speech. Download and run either the .ipynb or the .R.

Task 1 (1 mark)

1. Find transcripts of a different speech on the Internet, noting and recording where the data was taken from. (Can be done in OS)
2. Convert the file to a utf-8 encoded txt file and store it in your ./data directory
3. Read the data into the program and load it into a corpus.
4. Inspect the corpus, to see if there are any obvious problems (e.g. non-English punctuation marks, etc.)
5. Using tm_map, convert the content to lower case, remove numbers, stopwords, punctuation and whitespace.
6. Inspect it again, to make sure step 5 worked properly.
7. Build a term-document matrix and generate a word-cloud.
8. Plot the most frequently used words in the speech.

Task 2 (1 mark)

1. Write a function that will take in a text file and produce a word cloud and word frequency plot.
2. Find song lyrics on the Internet, noting what they are and the source.
3. Using your function, produce a word cloud and a frequency plot for the lyrics.
4. Write a critical analysis of the word clouds produced by both the speech and the song lyrics.

Task 3 – In markdown or commented block (1 mark):

1. Read back over the notes for week 6 - visualising text and state Zipf's Law in your own words.
2. How does your function address Zipf's Law? Discuss this in terms of the parameters you are not changing.
3. What are the challenges to this visualisation of text?