FIT3152 Data analytics. Tutorial 11: Text analytics

Contents:

- Text analytics, text processing and clustering in R

1. Work through the examples in the lecture slides.

2. Text pre-processing

   Create a Term-Document Matrix of the following data following the process outlined in the lecture notes: (Tokenise, Convert case, Filtering – including removing stop words, Stemming)

ID	Document
Doc1	Jazz music has a swing rhythm
Doc2	Swing is hard to explain
Doc3	Swing rhythm is a natural rhythm

3. Create 3 text documents following the example in the lecture notes and create a Term-Document matrix using R.

   Calculate the Cosine Distance between each pair of documents.

4. A number of reports relating to UFO sightings have been stored in the file (UFOsample.csv). Read the data into R, process and cluster the text. Adapt the following code fragment below to read each row of a csv file as a separate document into a corpus:

```
UFO = read.csv("UFOsample.csv", header = FALSE)
UFO = data.frame(doc_id = row.names(UFO), text =UFO[1])
colnames(UFO) = c('doc_id', 'text')
docs = Corpus(DataframeSource(UFO))
```

5. A selection of Usenet articles taken from 20 newsgroups are in the zipped folder (mini_newsgroups_mixedup.zip). These documents were obtained from the UCI KDD Archive of data sources for machine learning. Ref: http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.data.html. The topics are: alt.atheism; comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc.

   Note: these articles have been vetted for expletives but some may contain offensive content. Therefore you are not encouraged to read these in detail. (Newsgroup topic appears in header)

(a) Process and cluster the data. Inspect the Term-Document Matrix or word frequency counts. This should identify words that appear commonly in each article, regardless of topic.

(b) Using the information gained in Part (a) re-process the articles to eliminate these specific words or phrases and then re-cluster the data. If you are using hierarchical clustering, partition the data into 20 clusters.

(c) Inspect the clusters obtained in Part (b), do the clusters contain articles on the same or similar topics? Hint: look at the document IDs. Newsgroup topics are coded (approximately) by topic as TTTxxx where TTT is the topic code.

6. The text for the novel *David Copperfield*, by Charles Dickens was obtained from Project Gutenberg, ref: https://www.gutenberg.org. An edited version of this text – with material not in the original book removed is in text file (David Copperfield PG Edit.txt).

(a) Process the data for text analysis and create a data frame of the top 100 or 200 most frequently appearing words. Plot a column graph of these word frequencies. What do you observe?

(b) Using the data in Part (a) inspect the data frame and using Wikipedia or any other source identify the main characters listed. Reprocess your data to selectively remove these characters and re-plot the word frequencies.