CS3907-80/CS444-10 Big Data and Analytics

Class project #3

Text Analytics in R

1. Data Set: acq

The problem is to process a large set of documents (50) to understand how text analytics works.

The function acq in the R package tm references a corpus of 50 documents.

You can start by following the slides in Lecture 7.

You should do at least the following:

a. For the complete set of documents, try the functions in lecture 7. What happens? Does it yield anything understandable about the documents.

b. Find the 15 longest documents (in number of words).

c. For each document work through the examples given in Lecture 7 to display the dendrogram and the WordCloud.

For the following you will need to write R functions to help you compute the results.

Use the packages textreuse, wordnet, zipfR

d. Prior to removing the punctuation, find the longest word and longest sentence in each document from the 15 largest documents.

e. Print a table of the length of each sentence in each of the 10 documents.

f. For each sentence of each document, remove the punctuation. Display the sentences.

g. For each word print its part of speech using the Wordnet package.

h. Analyze word frequency using functions from package zipfR.

2. Deliverables: You will deliver your results by putting a zipfile in your group’s Blackboard file, with the following naming convention: Group-N-Project-3.zip, where N is your group number. Your deliverable should encompass the following items:

A listing of all R functions that you have written

A document giving your results which should include your assessment of applying the different techniques to the data provided.

Remember to save your workspace! In your Group area would be a good place so all members can get to it.

Include in your Word document the results required

(use a CTRL-ALT-PrintScreen) to grab the screen

You may use Irfanview 4.40, [irfanview@gmx.net](mailto:irfanview@gmx.net). Paste in the screen image, and copy the image as JPEG to drop into your Word document.

3. Due Date: May 4, 2017 COB

4. Project #3 Value: 25 points

a. Document R functions: 3 points

b. Presentation and discussion of results from the experiments that you run using the different functions from Lecture 7: parts (a) through (h) 2 points each. Include plots where applicable.

c. Write an R function to search through the documents to find a specific word or phrase. Print the document number, line number, and word index in the sentence. Demonstrate with three examples. Use words of 6 characters or more as your test cases. 3 points.

d. Analysis of what this project helped you learn about data science, e.g., the exploration of data which is what you have been doing: 3 points