CS3907/CS6444 Big Data and Analytics

Class project #2

Text Analytics in R

1. Data Set: Dreams.txt by Henri Bergson

The problem is to process a large document to understand how text analytics works.

You can start by following the slides in Lecture 4.

You should do at least the following:

a. Try the functions in lecture 4.

- What happens?

- Do they yield anything understandable about the document.

b. Find the 10 longest sentences (in number of words) in the text.

c. For each paragraph, 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 of the paragraphs. Print a table of the length of the shortest and longest sentences in each chapter.

e. Use WordNet to mark the parts of speech for the first five paragraphs for nouns and verbs having a length of 5 or greater.

f. Analyze word frequency using functions from package zipfR.

g. Generate bigrams and trigrams for all words in the first three paragraphs.

Now, there are several other packages that you can use. By now, you know to load packages, read their PDF files, and apply their methods to do analysis.

Read text\_analysis\_in\_R.pdf.

h. Process the text from *text* using corpusTools, stringi, corpustools, and quanteda. Describe the methods you use, the results, you get, and what you understand about the theme of the book.

By now, you should see that Data Science is an empirical science. So, these packages provide tools that can give greater insight into the text. At a minimum, choose three (3) functions from each package and apply them to Chapter 1.

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-2.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: July 29, 2019 COB

4. Project #2 Value: 40 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) 3 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. 8 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: 5 points