**Big Data**

**Text Analytics in R**

Data Set: A Scandal in Bohemia by Arthur Conan Doyle.

The problem is to process a large document and analyze it.

**1**. Create A VCorpus after you separate the chapters. You should have three documents in your corpus.

a. Prior to removing the punctuation, find the 10 longest words and 10 longest sentences in each chapter. Prepare a table of this data as well as showing these items.

b. Read Introduction to Text Analytics.docx. Apply the methods in that document to the documents in your VCorpus. Also, Read text\_analysis\_in\_R.pdf.

**Show all you work with explanation of what you are doing. Tables are a good way to show your work in a compact manner.**

**After removing the stop words and punctuation:**

c. If the dendrogram is unreadable, remove words of length 2, 3, 4, etc. to see how the dendrogram is affected. Show the dendrogram each time.

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

Use the packages textreuse, wordnet, zipfR

d. Use openNLP to mark the parts of speech for the 10 longest sentences found in part b for nouns and verbs having a length of 5 or greater. Prepare a table with this data.

e. Analyze word frequency using functions from package zipfR. Prepare a table with this data.

f. Generate bigrams and trigrams for all words whose length is greater than 6 characters in the 10 longest sentences of each chapter. Prepare a table with this data.

g. Perform the procedures in Introduction to Text Analytics re Word Cloud and Sentiment Analysis.

h. Install the packages stringi and quanteda and select three additional methods from each – not the ones used in (g) – apply them and show the results.

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 the document.

**2**. Deliverables: You will deliver your results by putting a zipfile.

Your deliverable should encompass the following items:

You should prepare a report discussing in detail what you did and showing the results.

You should present the function calls and results from running the functions in Intro to Text Analytics. Discuss what it tells you.

Discussion what this project helped you learn about text analytics, e.g., the exploration of data which is what you have been doing. This should be at least two robust paragraphs.

Include in your Word document the results required

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

Paste in the screen image, and copy the image as JPEG to drop into your Word document.

Show ALL of your work meaning use screen shots to show the function calls and the results.

a. 1 point

b. 5 points

c. 2 points

d. 2 points

e. 2 points

f. 2 points

g. 4 points

h. 1 point

Discussion of what your learned – 1 point

Total 20 points