CS4907/CS6444 Big Data and Analytics
Class Project #3
Due Saturday COB before Final Exam

Text Analytics in R

Data Set: Tarzan of the Apes by Edgar Rice Burroughs, use chapters I – XV

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

1. Create a VCorpus after you separate the chapters into individual files.

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. **Follow the Rubric!! You will be graded according to the Rubric!!**

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:

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 zip file 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:

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.

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

3. Due Date: COB (Saturday before the Final Exam)
4. Project #3 Value: 35 points
a. Items a above – 10 points
b. Follow the Rubric – do all methods – 23 pts
Discussion of what your learned – 2 points
Total 35 points