Spring Semester 2018
Iowa State University

LING 410X - Language as Data

Assignment 3 - words, phrases and ngrams
Submission Deadline: 24 Feb 2018, end of the day

Instructions: This assignment is for 10 marks. First question carries 4 marks and second question carries 6 marks. Upload your submission as a zip file in your_Lastname_A3.zip format. If any of the programs does not run and throws errors, you cannot get credit for that. Late submissions are allowed, but will not be awarded full credit.

Question 1

Read the following URL's content into R and calculate the following: `http://www.gutenberg.org/cache/epub/2446/pg2446.txt`

- lexical variety in that article (unique words/total words). Consider only words - not punctuation markers and other such non-word characters.
- 10 most frequent words in the article

Now, write a short 1 page report on the article you picked, and the results you got, and how you got them (i.e., R commands you used)
Here is a useful tutorial on the second part of the question: `http://johnvictoranderson.org/?p=115`.

Question 2

Download the content of any 2 books by one author from gutenberg.org (in .txt format!). In R, read the files, remove punctuations, lowercase all words, split the text into words based on white space seperation, and make a list of unique unigrams, 2 and 3 grams in the book. Compare the lists for both books thus created, and create a list of unigrams, bi and trigrams that appear in both texts. Each time, print the following to console: Number of n (n=1,2,3) grams in Book1, in Book 2, and their intersection. Submit your assignment as a R-markdown submission. R packages you may find useful: readr, ngram.