Spring Semester 2018
Iowa State University

LING 410X - Language as Data

Assignment 2 - Working with different data formats
Submission Deadline: 10 Feb 2018, end of the day

Instructions: This assignment consists of two questions. Each question carries 5 marks. Upload your submission as a zip file in your_Lastname_A2.zip format. Late submissions are allowed, but will not be awarded full credit.

Question 1

Go to the following link: `http://www.gutenberg.org/ebooks/2446` where the same book exists in different formats. Your task is to read the HTML document and the plain-text (.txt) file onto your Desktop and read the files in R using readLines() function. Compare the numbers you got in both steps. If they are different, try to guess why they are different looking at the text of both versions. Which one of these formats is easy to process for R in your opinion? Why? Write this up in a 1-2page document and submit as pdf.

Optional challenge: You can also try to figure out and use XML library in R to parse HTML files. If you decide to go the XML library route, this will require you to install the package "XML", which can parse HTML. You should read its documentation to figure out how to work this!

Question 2

"GuardianR" is a R package provide us with functionalities to search and access news articles from "The Guardian" newspaper. Your task in this question is to setup this package to work, learn to work with it (seeing the GuardianR manual : `https://cran.r-project.org/web/packages/GuardianR/GuardianR.pdf`) and do the following:

1. Search for articles about Justin Trudeau between 1st-15 January 2018 and on 1st-15th January 2014 (Do not worry about the actual content of what you get. My purpose is to make you search for a string with two words - in this case - Justin Trudeau).

2. How many results did you get for 2018, and how many did you get for 2014?

3. Prepare a spreadsheet with 2014 results and 2016 results (2 columns for each year - id and wordcount.)
4. Submit this spreadsheet as a part of your assignment.

Here are a few r-functions that will be useful in this process:

- if results is a dataframe:
  1. names(results) gives you the column names of the data-frame results.
  2. nrow(results) gives you the number of rows in results
  3. ncol(results) gives you the number of columns in results
  4. results["id"] prints you all text in column "id" in results

Note: There were several R libraries for several content rich websites such as : Wikipedia, Gutenberg, NYTimes, Twitter. However, all of them need not necessarily form the exact same structure of output as this. The best approach is to check the documentation for that package, and follow a trial and error process to understand the format.