Spring Semester 2018
Iowa State University

LING 410X - Language as Data

Assignment 1
Submission Deadline: 27 Jan 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_A1.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

Pick one of the following disciplines that is the closest to your major or the one that interests you the most:

- Psychology
- Sociology
- Literature
- Political Science
- Linguistics
- Business, Management and Finance
- Life sciences
- Agronomy and Agriculture
- Information technology

Now, do a little bit of research and list 5 problems where text processing is useful in these areas. Give a list of stuff you read (news articles, wikipedia etc) as references at the end of your report. Your report should be a pdf file, can be upto 3 pages long, and without any double spacing.

Question 2

Read the tutorial pdf I am providing with this document (A1-HelpFile.pdf), and answer the following questions.

Consider this paragraph from ISU website on Cyclone Athletics:
For only the second time in its history, the Iowa State Athletics Department finished in the Top 40 of the nation's most-successful athletics departments from a competitive standpoint. The 2014 Learfield Sports Directors' Cup final tally has the Cyclones in 38th place with 585.75 points. The school's all-time best finish is 34th in 2010. Iowa State had the fifth-best finish among Big 12 schools and tops in the state of Iowa. The school earned 160 points in the spring season, which equaled its all-time record. The womens track & field (60), womens golf (51) and mens golf (49) accounted for the spring points. ISU had averaged only 46 points in the spring season previously. Iowa State also earned points from wrestling (64.5), men's basketball (64), women's cross country (63), gymnastics (59.25), women's indoor track & field (51.5), men's indoor track & field (46.5), men's cross country (27), women's basketball (25) and volleyball (25).

Based on what you learnt in the tutorial, answer the following questions using R. Submit the lines you typed in R along with the output you got as a pdf file.

1. Read this paragraph into a string variable called cyclones.
2. What are the number of characters in cyclones?
3. Uppercase cyclones and store it in another variable cyclones_upper
4. Substitute all occurrences of numbers of the form XY.Z (i.e., numbers with only 1 digit after decimal) with NUM and show the output.
5. Split cyclones text wherever a period (.) occurs and print the output. What is the type of output you get (string, list, numbers etc)?
6. Convert cyclones text into title case and print the output. What does the output look like?