LIS 490IDS
Fall 2016
Homework #7

Due Tuesday Oct 25, 2016 in moodle and in class

Prof. V. Stodden

Q. 1) (3 points)

Write down a *general* regular expression to match the following:

(a) Words with @ symbols in them, e.g., h@te or v|c0din

(b) An IP address (Four sets of 1 to 3 digits separated by periods, e.g., 100.12.162.0)

(c) An email address that ends with .com, .edu, .net, .org, or .gov

Q. 2) (19 points) Carry out the following exercises on the State of the Union speeches database (available in moodle).

(a) Use readLines() to read in the speeches (available as a text file in moodle) where the return value is: character vector with one element/character string per line in the file

(b) Use regular expressions to find ***

(c) Use *** to identify the date of the speech.

(d) Use regular expressions to extract the year.

(e) Use regular expressions to extract the month.

(f) Use *** to extract the name of the president State of the union speeches.

(g) Use regular expressions and R to return the number of speeches in the dataset, and the number of presidents that gave speeches.

(h) Chop the speeches up into a list there is one element for each speech. Each element is a character vector. Check: does your number of list elements match your answer above?

(i) Eliminate apostrophes, numbers, and the phrase: (Applause.)

(j) Make all the characters lower case.

(k) Split the sentences up where there are blanks and punctuation to create “words”.

(l) Drop any empty words that resulted from this split.

(m) Create a word vector for each speech.

(n) Normalize the word vectors to get term frequencies.

(o) (5 points) Carry out some exploratory analysis of the data and term frequencies. For example, find the number of sentences, extract the long words, and the political party. Plot and interpret the term frequencies. What are your observations?