Spring Semester 2018
Iowa State University

LING 410X - Language as Data

Assignment 4
Submission Deadline: 10 March 2018, end of the day

Instructions: This assignment is for 15 marks. The first question has 5 marks and second one has 10 marks. Upload your submission as a zip file containing R markdown file and its doc/pdf version, and any result files you want to submit. Follow the your_Lastname_A4.zip format. Late submissions are allowed, but will not be awarded full credit.

Question 1

In this assignment, you will be working with a publicly available dataset called sentiment_sentences which contains sentences that are manually annotated as having positive or negative sentiment (this is downloaded from the example datasets in LightSide textmining workbench: `http://ankara.lti.cs.cmu.edu/side/download.html`). I split the data into training and test sets and they are uploaded with the assignment as .csv files which can be read in R. Each line has two columns, separated by a comma character, and the sentence is enclosed in quotes. Using what you know in R so far, find out what are the most frequent 5 words in the training and test data, along with their frequencies. Write a 1 page report giving the answer and how you got it, what kind of pre-processing you did, etc. Explore using read.csv(), to read columns of strings in a file.

Question 2

There is a helpfile along with this assignment which shows you step-by-step procedure to train and test a sentiment classification approach. I am just using bag of words as features there. Your task is the following:

- Using the given training-test data, train a bag of words classification model and analyse the results you obtain (in the final csv file you generate). How is your algorithm doing? Is it correct? How many times is it correct? What do you think of the results.
- Did the model do stemming? Did it remove stopwords? If it did not, how do you add those into the process? Try to repeat the steps adding one of these missing pre-processing stuff and write about how it changed the accuracy on test data.

- If you change SVM in train_model step to some other algorithm, did you see any change in your results?
- I am not showing the prediction accuracy in my help file. Is there a way to show it?

-your report can be upto 2 pages long.