LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

16 January 2018

Class outline

- Week 1 Quick recap: Swirl lessons
- Linguistic Knowledge and Text Processing
- Counting words in a text - R example

Week 1 Review

Week 1 review

- Course overview
- R installation and swirl lessons.
- Is there anyone who did not start working with R?

Week 1 review

- Course overview
- R installation and swirl lessons.
- Is there anyone who did not start working with R?
- How many Swirl lessons did you do?

Week 1 review

- Course overview
- R installation and swirl lessons.
- Is there anyone who did not start working with R?
- How many Swirl lessons did you do?
- Any questions or concerns so far on anything from last week's classes?

Week 1 review

- Course overview
- R installation and swirl lessons.
- Is there anyone who did not start working with R?
- How many Swirl lessons did you do?
- Any questions or concerns so far on anything from last week's classes?
- Did anyone take a look at the Assignment 1 description?

Few questions about Thursday's lessons

- What is a vector in R? How are they created?

Few questions about Thursday's lessons

- What is a vector in R? How are they created?
- What does ?c do on R console?

Few questions about Thursday's lessons

- What is a vector in R? How are they created?
- What does ?c do on R console?
- What is "recycling"?
- What does a paste() function do with vectors?

Recap of Lesson 2

Note: Directory is what we would call a "Folder"

- getwd()
- setwd("some folder path")
- list.files(), list.files("somefolder") (Quotes are important!)
- file.create("somename.txt")
- file.exists("somename.txt")
- file.info("somename.txt")
- file.rename("oldname.txt","newname.txt")

What is a folder or file path?

file/folder "path"

- What is a path?

file/folder "path"

- What is a path?
- How do we find it?

file/folder "path"

- What is a path?
- How do we find it?
- Relative path vs Absolute path

file/folder "path"

- What is a path?
- How do we find it?
- Relative path vs Absolute path
- file.path function in R

file/folder "path"

- What is a path?
- How do we find it?
- Relative path vs Absolute path
- file.path function in R

Lesson 3: Sequences

```
seq(1:20)
seq(0,10,by=0.5)
seq(0,10,length=20)
seq(20:1)
rep(c(1,2),times=10)
rep(c(1,2),each=10)
```

Question: What does seq_along() do?

Linguistic Knowledge and Text Processing

Text Processing

- Different tasks require different kinds of language awareness and text preprocessing
- Example tasks: getting word collocations, word frequencies, getting POS tags, getting "meaning" of a sentence
- Example pre-processing: splitting text into words or sentences, spelling normalization, lower casing, removing punctuation and so on.
- In this course, we will primarily discuss text mining at the level of words, without going into deeper linguistic analysis.

Time for some exercises :-) Form into groups of 2-3 people.

Scenario -1

If you are asked to get the most frequent 10 words used in "Alice in Wonderland", what are the pre-processing steps you will do and how will you get the list of words and their frequencies?

Scenario -2

If I give you the text transcripts of all speeches made by all presidential candidates since 2000 and ask you who uses the most number of adjectives, how will you get the answer? What kind of pre-processing should you do and not do? Let us assume there is a way to get the part-of-speech tags for all words in any sentence.

Scenario -3

If I ask you to find out what are the 10 most frequent words tweeted from Des Moines area in the past 10 days, how will you get the answer? Assume there is a way to get tweets specific to a location and in a given time frame. What are the pre-processing issues you may encounter in this case?

Scenario -4

Continuing on the twitter problem, will it be easier if I ask you to find out how many people shared a URL in their tweets instead?

Scenario -5

If I give you a pdf file and ask you to count the number of words in it through R, do you think it is easy or difficult? Why?

Counting the frequencies of words in a text - R example and discussion

Task: Let us see if we can count the 10 most frequent words in the play "A Doll's House" by Henrik Ibsen

What do we need for doing this?

- Some way to read the file into R (from Project Gutenberg)
- Some way to discard all the Gutenberg metadata, and keep only the actual text we need
- Lowercase all text (we don't need case info), some way to split it into individual words
- Someway to get counts for all unique words in the text
- Optional: someway to plot/visualize whatever we extracted.

Procedure and R-functions

- scan() function - lets you to read a file on your computer into R environment.
- which() function - gives you the line number of the string you search for.
- tolower() - converts everything into lower case
- strsplit() - splits a string into words
- table() - makes a table of words and frequencies
- sort() - arranges a table in some order - ascending or descending
- plot() - is a function to make simple plots

R code for counting word frequencies

```
setwd("~/Dropbox/ClassroomSlides-BothCourses/LING410X/")
english <- scan("DollsHouse-Eng.txt", what = "character", sep = "\n")
english.start <- which(english == "DRAMATIS PERSONAE")
english.end <- which(english ==
      "(The sound of a door shutting is heard from below.)")
actual_english <- english[english.start:english.end]
actual_english_string <- paste(actual_english, collapse = " ")
english_lower <- tolower(actual_english_string)
english_words <- strsplit(english_lower, "\\W+")
sorted_freqs_english <- sort(table(english_words), decreasing = TRUE)
plot(sorted_freqs_english[2:11], type="b")
```

Next Class

- Regular Expressions and looking for patterns in strings
- Practice exercise about counting frequencies, and looking for patterns
- To Do before the class: Read Chapters 1–3 in the textbook.