LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

09 January 2018

Class outline

- Introductions
- Motivation for the course
- Course objectives and Pre-requisites
- Course Logistics
- Syllabus
- Group activity
- Pre-course questionnaire

Introductions

About me

1. In ISU as Asst. Professor since January 2016.
2. PhD in Computational Linguistics, 2015.
3. Teaching experience:
   - Language as Data (this course, last year)
   - Language and Computers (LING 120), Statistical Natural Language Processing (LING 515)
   - Introductory courses on programming and NLP for applied linguistics students
   - Technical communication for undergrad engineering students
   - Graduate level topics seminars for computational linguistics students (2012-13)

About you?

1. Name
2. What do you do in ISU?
3. What are your interests related to computational analysis of language?
4. Why did you enroll in this course?

Motivation for the course

Why this course?

1. There is a lot of data available everywhere now. Text is one form of such data.
2. We write comments on amazon.com, read news, write blog posts, use twitter - all these forms of internet usage create lots and lots of textual data everyday.
3. Knowing how to work with text and extract some kind of information from it is a valuable and industry relevant skill.
4. This is the main reason for the creation of this course.

Some context from the news

- Literature
  1. "data mining reveals the six basic emotional arcs of story telling" (https://goo.gl/i1xWTu)
  2. authors "claim they created an algorithm that identifies the literary elements that guarantee a book a spot on the besteller lists." (https://goo.gl/Hsjfmjl)
- Political science:
  1. on linguistic analysis of debate transcripts from recent elections (https://goo.gl/BNxTaa, https://goo.gl/7KIyWj)
  2. 9/11 anniversary speeches: what next analysis tells us (https://goo.gl/dCj477)
- Clinical psychology: "can you detect a manic episode on Twitter?" (https://goo.gl/i5V7ST)

... and so on. Not all these are super successful or anything. These are just a few examples to show the relevance of what you will learn in areas other than technology and computer science.

Why did I choose R?

1. R is fastly becoming a popular language for data science and statistical analysis
2. R has a lot of support for creating visualization tools
3. Finally, you don't have to immerse yourself into programming to be able to write R code for your work. (my personal opinion)

Why did I choose R?

1. R is fastly becoming a popular language for data science and statistical analysis
2. R has a lot of support for creating visualization tools
3. Finally, you don't have to immerse yourself into programming to be able to write R code for your work. (my personal opinion)
4. So, I believe it is a suitable language to teach about doing text analysis to non-CS students.
5. What can CS students benefit: knowledge about text processing and R, programs about analysing textual data.

Course Objectives and Pre-requisites

Goals for the course

1. Teach you basic methods and techniques of text processing
2. Teach you how to use R to analyse your own data
3. Teach how to create visualizations of text data
4. Make you a comfortable R user who can search for and utilize existing R libraries to find solutions to your text processing problems
5. Make you work on a practical project that is relevant to the outside world

What are not the goals for this course

1. Make you an expert programmer
2. Make you an expert R programmer
3. Make you a statistical analysis expert

Pre-requisites

None. General curiosity about language, and a willingness to work with computer programs and tools.

Course Logistics

Meeting and Location

- Curtiss 0225 on Tuesdays, and Ross 0137 (Lab) on Thursdays, 9:30-10:50 am
  (Note that the Thursday classroom is different from what is put up on class scheduler.)
- *Office hours:* Tuesdays and thursdays, 11 am-12 noon (please email beforehand if there are specific issues to discuss. If this time does not work for you, send an email, and we can meet at a convenient time)
- course website: on Canvas.
- Credits: 3

Format and Grading

Course Format

- weekly lectures and practical sessions
- 6 assignments (70%) + 1 final project (25%)
- 5% for classroom participation/discussion participation

Assignment and project deadlines

- Assignment 1: 27 Jan 2018 - 10 marks
- Assignment 2: 10 Feb 2018 - 10 marks
- Assignment 3: 24 Feb 2018 - 10 marks
- Assignment 4: 10 Mar 2018 - 15 marks
- Assignment 5: 31 March 2018 - 15 marks
- Assignment 6: 14 April 2018 - 10 marks
- Group project: (25 Marks total)
  - Initial report due: 7 April 2018 (5 marks)
  - Project presentation: 24-26 April (5 marks)
  - Project report, and code submission: Finals week, 3rd May (15 marks)

(3 assginments are already uploaded. Rest will be up in 2–3 weeks)

Some general rules:

- attendance: 80% attendance requirement. Attendance is counted through per-class questions asked in the class, which can be answered in the discussion forum.
- missing a deadline is okay, but you will not get full credit.
- long absence due to illness etc: please inform and follow university procedures.
- cheating and plagiarism: see the course handbook, and university policy against plagiarism.
- classroom behavior: please be punctual and do not do personal work in the class.
- Disability accomodation: Please speak to Disability Resources Office (DRO) to officially request an accomodation.

Other Issues

- validating enrollment: who is enrolled? who is just here?
- feedback about the course:
  1. Talk to me directly, or leave anonymous feedback at: https://goo.gl/forms/9o4AmL9bpOfsHlRF2 or leave a paper feedback in my mailbox.
  2. Be confident enough to confront me and talk to me if there is a concern.

Syllabus

Topics

- Introduction to the course, R, and linguistic analysis
- Corpus preparation: methods to select, process and clean textual data
- Keyword and Key-phrase extraction methods
- text classification methods and their applications
- topic modeling and its applications
- methods of visualizing textual information

Text Book

1. Primary textbook: "Text analysis for students of literature" by Matthew Jockers
   - It is freely available as pdf from university network (from the publisher)
   - I will be using several other free online tutorials and stuff - urls will be given in appropriate locations
   - Software: R, RStudio (a graphical interface for R), and several text processing libraries in R (will talk about them as needed).

Any questions so far?

Next Class ..

- To do before next class:
  1. Read the syllabus handbook carefully
  2. If you have your own laptop, get that for thursday's class to install required stuff
  3. Please note: Thursday's class is in ROSS 0137
- Next class:
  1. Installing R, Rstudio
  2. Working with R - tutorial
  3. Assignment 1 description

Group Activity -1

Form into groups of 3 (know your classmates!) and figure out the answer for the given word sentiment classification problem.

Answers? -1

Here is how I grouped the words: molistic, slatty, blitty, weasy, sloshful - perhaps belong to one group. strungy, struffy, danty, cloovy, cluvious, brastic, frumsy - perhaps belong to one group. If so, then, answer to first question will be C and second question will be D.
source: NACLO 2007 puzzles. (http://nacloweb.org/resources/problems/2007/N2007-AS.pdf)

Group Activity - 2

Form into groups of 3 (know your classmates!) and figure out the answer for the given search relevance problem.

Answers? -2

solution: http://nacloweb.org/resources/problems/2007/N2007-BS.pdf