Spring Semester 2017
Iowa State University

LING 410X - Language as Data

Course Handbook

Instructor: Sowmya Vajjala

- *Office:* 331 Ross Hall
- *Email:* sowmya@iastate.edu

Course Objectives: This course aims to introduce students to methods of discovering language patterns in text documents and applying them to solve practical text analysis problems in their disciplines. Data of any form (text, numbers, images etc.) is available in large amounts now like never before. Text is one of the major forms of big data and hence text analysis is in huge demand in the information technology industry now. Apart from the technological applications, it is also useful in various disciplines like business intelligence, sociology, psychology and literature to name a few. For example, key word extraction and sentiment analysis are very useful in Business analytics. Authorship detection and stylometric analyses are examples applications for literature. Studying mental disorders through patient written samples is gaining prominence in clinical psychology. In this background, this course introduces some commonly used methods to work with textual data. After a brief primer in the fundamentals of linguistics and its role in text analysis, the course will introduce the students to writing R scripts (as it is easier to do exploratory analysis and visualization in R without learning a lot of programming principles) to perform text analysis and visualize textual data.

Learning Outcomes After finishing this course, students will know:

1. some common methods for performing automatic text analysis
2. some real-life applications of text analysis
3. how to apply these methods to solve text analysis problems in their domain areas
4. how to visualize textual data using various tools and methods

Pre-requisites: Junior Standing. LING 120 is a preferred but not a mandatory pre-requisite.

Course Details:

- on Tuesdays and Thursdays from 9:30 – 10:50 am
- Meets in Ross 0406 on Tuesdays, and Ross 0137 (Lab) on Thursdays. Note that the Thursday classroom is different from what is put up on class scheduler.

- *Office hours:* Tuesdays and Thursdays, 11-12 noon (please email beforehand if there are specific issues to discuss. If anyone cannot make it during these hours, send me an email to fix an appointment.)

Credits:

- Credit Points: 3
  (Expect to spend 5-6 hours per week outside the class to work on the problems and assignments.)

Nature of the course and expectations: Primary mode of instruction is by lectures and hands-on lab sessions. The course will have regular assignments that deal with various methods of corpora creation and text analysis using software tools, and a final project. The corpora and resources used in this course will address the methods to solve various text analysis related to the student's discipline.

Students enrolled in the course are expected to

1. regularly and actively participate in class, and submit the assignments on time (80% of the grade)
2. work hard, and prepare well for the classes

Grading Policy There are 6 Assignments in this course covering 75% of your final grade, and a final project carrying 25 marks (which involves visualizing textual data). Plus/minus grading will be used (93% = A, 90% = A-, 87% = B+, 83% = B, 80% = B-, etc.).

The following are the scheduled deadlines for this class

- Assignment 1: 24 January (10 Marks)
- Assignment 2: 7 February (10 Marks)
- Assignment 3: 21 February (15 marks)
- Assignment 4: 10 March (15 marks)
- Assignment 5: 2 April (15 marks)
- Assignment 6: 15 April (10 marks)
- Group project: (25 Marks total)
  - Initial report due: 21 March (5 marks)
  - Project presentation: 2-5 May (5 marks)
  - Project report, and code submission: Finals week (15 marks)

Attendance Policy Attendance is not mandatory, but recommended.

Class etiquette: Please do not read or work on materials for other classes in this class. Come to class on time and do not pack up early. Electronic devices like mobile phones, tablets etc should not be used in the class. Laptops should not be open in class unless there is a concrete, assigned activity. If for some reason, you must leave early or you have an important call coming in, or you have to miss class for an important reason, please let me know (via email) and get it approved *before* the class. Being absent from the class does not allow you to skip submitting any assignments that were assigned in that class.

Academic Conduct: Generally, you are encouraged to work in groups, discuss, and exchange ideas. At the same time, you are expected to do your assignments by yourself and with honesty. For the text you write, you always have to provide explicit references for any ideas or passages you reuse from somewhere else. Note that this includes text taken from the web. You should cite the url of the web site in case no more official publication is available. Specifically, the class will follow the University policy on academic dishonesty. Anyone suspected of academic dishonesty will be reported to the Dean of Students Office: http://www.dso.iastate.edu/ja/academic/misconduct.html

Disability Accommodation Iowa State University complies with the Americans with Disabilities Act and Sect 504 of the Rehabilitation Act. If you have a disability and anticipate needing accommodations in this course, please contact (instructor name) to set up a meeting within the first two weeks of the semester or as soon as you become aware of your need. Before meeting with (instructor name), you will need to obtain a SAAR form with recommendations for accommodations from the Student Disability Resources, located in Room 1076 on the main floor of the Student Services Building. Their telephone number is 515-294-7220 or email disabilityresources@iastate.edu . Retroactive requests for accommodations will not be honored.

Harassment and Discrimination Iowa State University strives to maintain our campus as a place of work and study for faculty, staff, and students that is free of all forms of prohibited discrimination and harassment based upon race, ethnicity, sex (including sexual assault), pregnancy, color, religion, national origin, physical or mental disability, age, marital status, sexual orientation, gender identity, genetic information, or status as a U.S. veteran. Any student who has concerns about such behavior should contact his/her instructor, Student Assistance at 515-294-1020 or email dso-sas@iastate.edu, or the Office of Equal Opportunity and Compliance at 515-294-7612.

Dead Week Policy This class follows the Iowa State University Dead Week policy as noted in section 10.6.4 of the Faculty Handbook: http://www.provost.iastate.edu/resources/faculty-handbook

Textbooks The primary textbook is: "Text analysis with R for students of literature" by M.J.Jockers and you are not obligated to buy it. The course will also rely on a wide range of freely accessible online tutorials and videos related to various methods of text analysis. (example: https://github.com/kbenoit/ITAUR-Short).

Syllabus - topics covered

1. Introduction
   Text analysis, real world applications, usefulness for various disciplines
2. Introduction to Linguistics and the role of linguistic knowledge in solving text analysis problems, with examples
3. Installing R and working with it.
4. Corpus preparation: methods to select, process and clean corpora
5. Keyword and Key-phrase extraction methods
6. text classification methods and their application for sentiment detection
7. topic modeling and its applications
8. methods of visualizing textual information

Scheduling and Deadlines (tentative) Note that the following session plan is subject to change; it only constitutes the current state of our planning as the semester unfolds.

1. Tuesday, January 10: Introduction to the course, expectations etc.
2. Thursday, January 12: R set up, basics practice (lab)
   A1 assigned. Due on 24th January
3. Tuesday, January 17: Overview of Linguistics and its relevance to discipline specific problems related to text processing
4. Thursday, January 19: introducing text processing in R
5. Tuesday, January 24: Corpus preprocessing and cleaning - Introduction and issues involved
   A2 assigned on pre-processing text. Due on 7th February
6. Thursday, January 26: Corpus cleaning continued (Reading from HTML, PDF, XML, JSON etc.) + Practice.
7. Tuesday, January 31: Scraping data from Twitter, NYT etc.
8. Thursday, February 2: corpus cleaning: conclusion + learning to use R Markdown
9. Tuesday, February 7: Introduction to vocabulary analysis. Keywords and Phrases extraction - overview, and applications
   A3 assigned on vocabulary and phrase analysis. Due on 21st February
10. Thursday, February 9: KWIC and other such tools: usage, analysis and lab
11. Tuesday, February 14: Words to Phrases (ngrams etc)
12. Thursday, February 16: Conclusion of the topic and exercises.
13. Tuesday, February 21: Text classification overview
    A4 assigned. Due on 10th March. 15 marks.
14. Thursday, February 23: Text classification and R

15. Tuesday, February 28: Text classification continued
16. Thursday, March 2: Text classification conclusion
17. Tuesday, March 7: Revision of concepts so far. Description of final project ideas
18. Thursday, March 9: Revision, Final project ideas discussion and decisions made.
19. Tuesday, March 14: Spring break
20. Thursday, March 16: Spring break
21. Tuesday, March 21: Topic Modeling
    A5 assigned. Due on 2 April. 10 marks.
22. Thursday, March 23: Topic Modeling
23. Tuesday, March 28: Topic Modeling
24. Thursday, March 30: Topic Modeling
25. Tuesday, April 4: Visualizing textual data
    A6 assigned. Due on 15th April. 10 marks.
26. Thursday, April 6: Visualizing textual data
27. Tuesday, April 11: Visualizing textual data
28. Thursday, April 13: Visualizing textual data
29. Tuesday, April 18: Group exercises on exploring domain specific problems and solving them.
30. Thursday, April 20: Group exercises on exploring domain specific problems and solving them.
31. Tuesday, April 25: Conclusion and Revision
32. Thursday, April 27: Conclusion and Revision
33. Tuesday, May 2: Project presentations.
34. Thursday, May 4: Project presentations.