Experimental Course Proposal
Iowa State University

LING XXX
Language as Data

Written by: Sowmya Vajjala on February 22, 2016

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

Reason for Proposal

Data of any form (text, numbers, images etc.) is available in large amounts now like never before. This resulted in a wave of new technologies and jobs that fall under the umbrella terms "data science" and "big data". 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. Despite the growing need for people who can do automatic text analysis, no such course exists for non-technical audience in the university, to name a few areas. During my conversations with faculty members from the school of business and department of sociology, I learnt that such a course will be largely useful for the advanced undergraduate and graduate students from their respective disciplines as well. This experimental course is being proposed in this background.

Course Description:

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. 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.

Pre-requisites: Juniors and preferably non-linguistics majors. LING 120 is a preferred but not a mandatory pre-requisite.

Nature of the course and expectations: Primary mode of instruction is by lectures and handson 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 exam. The corpora and resources used in this course will address the methods to solve various text analysis related to the student's discipline.

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

Textbooks and Other Resources The primary textbook is: "Text analysis with R for students of literature" by M.J.Jockers. The course will also rely on a wide range of online tutorials and videos related to various methods of text analysis. (example: https://github.com/kbenoit/ITAUR-Short).

Syllabus - topics covered

1. Introduction to text analysis, applications in real world, and some hands on experience with some text analysis tools like google n-gram viewer (1 week)
2. Installing R and getting some basic text statistics like word frequencies (1 week)
3. Introduction to Linguistics and the role of linguistic knowledge in solving text analysis problems, with examples (2 weeks)
4. Corpus preparation: methods to select, process and clean corpora (2 weeks)
5. Methods of studying word distribution and their application for extracting key words and phrases (2 weeks)
6. Topic modeling and its applications (2 weeks)
7. text classification methods and their application for sentiment detection (2 weeks)
8. methods of visualizing textual information (2 weeks)