Experimental Course Proposal
Iowa State University

ENGL XXX
Discovering patterns in texts: An introduction to text analysis (working title)

Written by: Sowmya Vajjala on February 8, 2016

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

Reason for Proposal

(description for this on the catalog page: Need for the course, intended use, programmatic justification - are the fields under this term in the proposal form)

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 of "data science" and "big data". Text is one of the major forms of big data, and finding patterns from large text corpora is one of the major applications of big data analytics. Text analysis and pattern mining play a major role in several industrial applications like opinion mining, summarization, information extraction for specific domains (such as legal texts, financial documents etc.). With the availability of such large datasets, various subfields within linguistics and literature studies too started working with large datasets and software (http://www.culturomics.org/ is a popular example). This resulted in the increase in the demand for linguists who know how to work with textual data on computers, both in academia and in the industry. This experimental course is being proposed in this background.

Course Description:

This course aims to introduce undergraduate linguistics students (and any others interested in text processing) to tools and techniques that can assist them in discovering language patterns in text documents and study language use. Starting with simple unix tools for matching handcrafted patterns in texts, we will move to querying large text databases for information extraction. The course will conclude with the gentle introduction to some of the state-of-the-art visualization tools for large text corpora. The focus in this course in on practical applications of text processing rather than the development of new methods to process text or the development of new software that does more sophisticated analysis. Thus, the aim of this course is not to make anyone a software developer or a computational linguist. The aim of this course is to introduce linguistics majors to some methods of processing text corpora, and enable them to write small text processing programs themselves for doing text analysis and interpretation.

Pre-requisites: Familiarity with using computers and an interest in working with text corpora. LING 120 is a preferred but not a mandatory pre-requisite. Enthusiasm to learn how to write computer programs is a must.

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 writing

code for automatic extraction of textual information from various corpora, and an oral exam. The corpora used in the course will be freely downloadble versions, and will address various data problems related to language (e.g. opinion analysis, politeness evaluation, information extraction, searching texts, estimating language change across times etc).

Learning Outcomes After finishing this course, students are expected to be comfortable with various methods of extracting information from text documents. They would have gained the handson knowledge required to perform text analysis in their undergrad studies, future graduate studies, or while working at some company.

Textbooks and Other Resources The course does not have a specific textbook, and will rely on a range of freely available ebooks, online course videos and free and open-source software. The primary resources are:

- Unix for Poets by K.W.Church (available as a free ebook - http://web.stanford.edu/class/cs124/kwc-unix-for-poets.pdf)
- Selected chapters from: "Introduction to Data Technologies" by Paul Murrell (available as a free ebook - http://stat.auckland.ac.nz/~paul/ItDT/itdt-2013-03-26.pdf)
- Selected chapters from: Python for Informatics: Exploring Information by Charles Severence (available as a free ebook: http://pythonlearn.com/book.php)

Syllabus - topics covered

1. Introduction to text processing, applications in real world, applications for linguistic areas.
   planned duration: 2 weeks
   primary material: web articles, videos and some hands on experience with tools like google n-gram viewer, COCA corpus GUI etc.
2. Simple hands-on text processing tools for small scale analyses
   - Counting and sorting words in a text
   - Extracting useful information from a dictionary
   - Computing n-gram statistics for a text
   - Making concordances
   - Searching a directory of text files for patterns (introducing information retrieval concepts)

   Duration: 4 weeks, primary material: "Unix for Poets". Other tutorials/videos will be there.
3. Querying and processing large text databases
   - How to query a database (SQL)
   - How to process information obtained from a database (Python crash course)

- Using the above two to do sentiment analysis on texts (classification), to extract information from texts (information extraction) and to group texts based on their central theme (clustering).
  Note: most courses of this kind are using R. It is definitely much easier for me to teach with R because of these course descriptions that people even shared with me via email, but I still need to think on the R vs Python part as I believe Python has more industry relevance than R while at the same time is used widely in academic settings too)

Duration: 5-6 weeks, primary material: Selected chapters from Murrell's and Severence's books.

4. a brief introduction to interpreting visualization tools for text data
   Duration: 1-2 weeks of handson experience with visualization tools (creating tag clouds, hierarchical clusters)