Fall Semester 2016
Iowa State University

LING 520 - Computational Analysis of English

Course Handbook

Instructor: Sowmya Vajjala

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

Course Objectives: Use of language processing tools in Applied Linguistics is increasing day by day especially in Computer Assisted Language Learning (CALL) and Intelligent-CALL (ICALL). Apart from adapting to the technology, learning to tweak software tools you use at work will enable you to customize existing technologies to your needs. This will also let you enhance the technologies with new text processing tools that are derived from your domain expertise. In this context, the objective of this course is to teaching natural language processing methods and tools for applied linguists. In this course, you will learn basic text processing techniques such as - pattern extraction from text, parts of speech tagging, spelling check-correction, text classification, and syntactic parsing.

Pre-requisites: Knowledge of programming (in any language) is mandatory.

Course Details:

- meets in Ross 312, on Tuesdays and Thursdays from 11:00-12:20 pm
- *Office hours:* Tuesdays and thursdays, 10-11 am (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 more time outside the class to work on the problems and assignments. That is not because I don't know how to teach or you are not smart enough. It is the nature of the course and programming can be frustrating and exciting at the same time even with 20 years of experience.)

Nature of the course and expectations: This is a 3 credit, graduate level course. Primary mode of instruction is by lectures followed by some discussion. We will have regular assignments, a final project and an oral exam for the project. Readings for each topic are specified in the syllabus and it is expected that the students read them before coming to the class. There may be a few additional (mostly optional) readings or videos from other sources for some of the topics.

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. finish a programming project as a final exam for the course and attend an oral examination about it (20% of the grade)
3. work hard, and prepare well for the classes

Grading Policy There are 4 assignments with 15 marks each, one assignment for 20 marks and one final project for 20 marks. The assignments will come about once in 2 weeks you usually will have 2 weeks of time for submission. For the final exam, you have to implement a small programming project, which is for 20 marks. The project should be decided by the end of October from a list of given projects. You are allowed to come up with your own idea, but should get my approval before starting to work on that. The final exam grade will also have a short oral defense where you have to explain what you did and answer questions. Plus/minus grading will be used (93% = A, 90% = A-, 87% = B+, 83% = B, 80% = B-, etc.).

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 be used only for activities related to the classwork. 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. Do not ask questions that can easily be answered by looking at the textbook or in online discussion forums. Value my time, your time, and everyone else's time. Asking such questions will not result in any answers from me **in the class** even if you write negative feedback about it.

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. It is common to search in websites such as stackoverflow.com for solutions to some problems or to fix the bugs in your program. However, copying full code samples from somewhere else (including your colleague's program) is considered academic dishonesty. Generally speaking, the class will follow Iowa State University's policy on academic dishonesty. Anyone suspected of academic dishonesty will be reported to the Dean of Students Office.

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

1. Speech and Language Processing - Jurafsky and Martin (I don't insist on buying as it is very expensive, but it is a useful book, and language agnostic). I will be using material from 2nd and 3rd editions. Default is 2nd edition (print book). Lot of draft chapters from this book are freely available online.
2. NLTK Book by Bird, Klein, and Loper (http://www.nltk.org/book/ - free ebook)
3. Language and Computers (Dickinson, Brew and Meurers) - optional. Good to get a general overview, without getting too technical. Slides for this book are available online.

(If you can only buy one textbook, I would suggest "Language and Computers" for non-expert programmers and "Speech and Language Processing" for those who want to continue working in NLP.)

Syllabus - topics covered

1. Introduction
   - Course overview
   - Introduction to Natural Language Processing (NLP)
   - Programming concepts review and practice

   Readings: Chapter 1 in Jurafsky and Martin, Chapter 1 in NLTK Book - Perl users can also read the chapter to get a general picture.
2. Text processing

- Basic text processing - getting word/ngram frequencies, regular expressions, tokenizing and sentence splitting.
- Calculating distance between words - application for spelling check-correction.
- Installing NLP tools for Python and Perl:
  (a) Python: NLTK http://www.nltk.org/install.html (Install NLTK and NLTK-Data)
  (b) Perl: Clairlib, Stanford CoreNLP toolset Perl interface http://search.cpan.org/~kal/Lingua-StanfordCoreNLP-0.02/lib/Lingua/StanfordCoreNLP.pm

Readings: Chapter 2 and 3 in J&M (2nd Edition). Chapter 3 in the NLTK Book (Perl users - I still suggest you to just browse through the chapter)
(Assignment 1 on the first two topics. Two ungraded problem sets with 10 questions each are given for programming practice.)

3. Morphological analysis

   - Regular expressions
   - Inflectional and Derivational Morphology

   Readings: Chapters 2,3 in J&M (2nd Edition). Link to Draft Chapter 2 from J&M 3rd edition: https://web.stanford.edu/~jurafsky/slp3/2.pdf
   (Assignment 2 on these topics. Ungraded problem set with 10 questions will be given for programming practice.)

4. Basics of probability, Language Modeling, Part of speech tagging
   Readings: Chapters 4 and 5 in J&M, Chapter 5 in NLTK book.
   Draft Chapters 4 (https://web.stanford.edu/~jurafsky/slp3/4.pdf) and 9 (https://web.stanford.edu/~jurafsky/slp3/9.pdf) of J&M 3rd Edition has most of this covered.
   Lecture Videos: Week 6 and 7 in Radev's coursera course
   (Assignment 3 on these topics)

5. Text Classification (Naive bayes, K-nearest neighbours algorithm, logistic regression)
   Readings: Chapter 7 in J&M (Edition 3, url: https://web.stanford.edu/~jurafsky/slp3/7.pdf), Chapter 6 in NLTK Book. Chapter 5 in Dickinson et.al. book.
   (Assignment 4 on this topic)

6. Parsing: constituency and dependency parsing
   Optional readings: Chapters 12, 13, 14 in J&M (Watching relevant video lectures will be sufficient).
   Videos: Week 4 and 5 in Radev's coursera course
   (Assignment 5 on parsing)

7. Overview of other topics in NLP: semantic analysis, discourse analysis; overview of NLP applications (Machine Translation, Information Extraction etc.,)
   Optional readings: Chapters 18, 20, 21 in J&M.
   Optional videos: Week 8–12 in Radev's coursera course.

8. NLP Applications in CALL: writers aids, language tutoring systems
   Optional Readings: Chapters 2 and 3 in Dickinson et.al. textbook (going through the slides for these chapters on the book's website will also be fine)

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, August 23: Course orientation; What is NLP?
2. Thursday, August 25: Why is NLP hard? What are some common text processing problems? Linguistics overview.
   (Assignment 1 given on basic text processing and programming, for 15 marks. Deadline - 10th Sep.)
3. Tuesday, August 30: Programming review + Practice
   (Problem Set 1 given for practice in and out of class)
4. Thursday, September 1: Programming Review continued. Installation of NLTK for Python users. Installing Perl libraries (Stanford CoreNLP package) for PERL users. Practice exercises.
5. Tuesday, September 6: Regular expressions review and practice.
6. Thursday, September 8: NLP - Preprocessing tasks (tokenizing, sentence splitting)
7. Tuesday, September 13: Preprocessing tasks continued: spelling correction and normalization; Assignment 1 discussion. (Assignment 2: Text processing and Regular expression programs. 15 marks. Deadline - 27th Sep. Problem Set 2 given for practice)
8. Thursday, September 15: Morphological Analysis (Book: Chapter 2 and 3 in J&M)
9. Tuesday, September 20: Morphological Analysis (Chapter 3 in J&M + Chap 4 in NLTK book)
10. Thursday, September 22: Ngrams, Language models, POS tagging (Chapter 4 and 5 in J &M, Chapter 5 in NLTK book)
    (Problem set 3 given for practice)
11. Tuesday, September 27: continued.
    (Assignment 3: on Tagging. 20 Marks, Deadline: 15 October)
12. Thursday, September 29: continued; Assignment 2 discussion.
13. Tuesday, October 4: continued.
    problem set 4 given for practice.
14. Thursday, October 6: Release of example projects for final projects (Deadline to decide: 5 November) and Revision of morphology, and tagging.
15. Tuesday, October 11: Text Classification (Chapter 6 in NLTK)
16. Thursday, October 13: Text classification - continued
17. Tuesday, October 18: Text classification - continued. Assignment 3 discussion.
    (Assignment 4: on text classification and others. 15 marks. Deadline: 29th October)
    problem set 5 given for practice

18. Thursday, October 20: Text classification - conclusion. Hui-Hsien's talk on her thesis. General Revision, Midterm feedback
19. Friday, October 21: Tutorial session @312.
20. Tuesday, October 25: Parsing, CFGs, PST.
21. Thursday, October 27: CFGs/PST continued.
22. Tuesday, November 1: Dependency parsing. (Assignment 5: on parsing and others. 15 marks. Deadline: 15th November; Discussion on Assignment 4; problem set 6 released)
23. Thursday, November 3: Parsing - conclusion + Discussion about decisions on Final projects. First report due on 5th November.
24. Tuesday, November 8: Other topics in NLP: Discourse analysis, Semantics etc.
25. Thursday, November 10: NLP applications: Machine Translation, Question Answering etc.
    problem set 7 for practice.
26. Tuesday, November 15: NLP and its application to Language learning and assessment, feature extraction from text, text classification etc
27. Thursday, November 17: status report on projects in the class. Discussion on Assignment 5. Problem set 8 released.
    problem set 8 for practice.
28. Tuesday, November 22: Thanksgiving break, no classes
29. Thursday, November 24: Thanksgiving break, no classes
30. Tuesday, November 29: NLP for CALL continued
31. Thursday, December 1:
32. Tuesday, December 6: Revision (Last week of classes)
33. Thursday, December 8: Revision (Last week of classes) Project submission (20 Marks) - Deadline Friday, December 9 Midnight.
34. December 12-16: Exams week. Oral exams scheduling.