Syllabus

1 Overview

Prerequisites: Linguistics 50, 53 and 111/112.

This course is a practical tour of computational methods that will be useful for practicing linguists. The basics of how text files are encoded on computational devices will be discussed, as well as common methods for manipulating these files, extracting their contents, and converting between formats. In addition, modern tools and algorithms for indexation and search, including regular expression grammars and databases, will be covered. Students will learn how to train some common tools (e.g., concordances, indices, part of speech taggers, chunkers, and parsers) in a variety of languages. The tools, methods, and metrics of modern annotation in computational linguistics will be discussed in considerable depth, including the mechanisms for crowdsourced annotation. After completing this course, students will be able to construct simple programs in Python and Javascript. No background in programming or UNIX will be assumed.

1.1 Course Schedule

The course will be divided into 6 subject areas, with the following order of presentation.

1. Introductions (2 week): Introductions to programming, basic concepts and data structures. Overview of file types and encoding; conversion systems. Basic web scraping.
2. Regular Expressions and Chunking (1 week): Introduction to Regular Expressions; Building chunkers and extractors.
3. Sorting and Searching (2 weeks): Tools for gathering data; methods for compiling data and doing simple queries; concordance building; indexing a corpus
4. Linguistic Annotation Guidelines (2 weeks): Best practices in annotation; metrics for interannotator agreement; using crowdsourcing to gather annotations
5. Advanced Querying (1 week): Introduction to structured query languages, including SQL and tgrep.
6. Interacting with the Web (1 week): CGI scripting; javascript visualization

2 Requirements

The course has six formal requirements: lecture, section, weekly problem sets, participation in two online experiments, and a final project.

2.1 Lecture

It is imperative that you attend class, especially given that there is **no textbook** for the class. This class will also require diligent and accurate note-taking – please do not rely on class slides/handouts to summarize information for you. Speaking up in class will affect your participation grade.

2.2 Section

Section attendance is **mandatory**, and will affect your participation grade. In section we will discuss the issues raised in class, go over technical details that passed by too quickly, and help with the homework problems. If you cannot attend one of the two sections, you may arrange to visit a TA's office hours. If you cannot do either, you should not take this class.

2.3 Weekly Problem Sets

There will be **7 homework assignments** in this course. While many of the problems will be routine applications of what was discussed in class, several will require imagination, cleverness, and lots of thinking-time.

Homework will be given out on Wednesdays and be due **at the start of class** on the following Wednesday, unless indicated otherwise. **You fail the course if you miss more than 2 assignments.** You should work on homeworks in pairs and submit one joint assignment. Please indicate on your submission who you collaborated with. If you collaborated with no one, please mark it.

We will deduct points for the following:

- not listing your collaborators **I will deduct 30% for this!**

LATE HOMEWORK POLICY: Each student is allowed two late assignments, which must be turned in within one week of the original due date. Late submissions will be given half credit.

2.4 Participation in Experiments

You will be required to participate in two experiments during this quarter. The experiments are each worth **1%** of your grade. The experiments can be of your choosing from the list the department has. I will make this list available to you for perusal.
If you cannot participate in an experiment, you may write a 4 page summary of one of the supplementary readings on the website.

In addition, you may enroll in seven additional experiments for extra credit. Extra credit will be computed as an additional 5% on one homework assignment.

2.5 Final Project

There will be a final project due in class on the last day class. This project will be an extension of one of your homework assignments according to the "further refinements" section of the assignment, and will represent a significant effort towards developing a practical tool and field testing it. The guidelines for the project will be discussed during the fourth week of class. The final is worth **24%** of your grade.

2.6 Participation

A word about participation. Your participation grade is 11% and is determined across how actively you are involved in the intellectual life of the class. Attendance in section will affect your participation grade, as will how you talk in class and section. However: coming to class (and sitting silent) will not count as a good participation grade – at best, it's a C. I don't take attendance in class, and if you aren't raising your hand and talking, you aren't participating (in my book).

2.7 Grade Computation

We will use the following system to compute your grades:

- ▷ Problem sets: 63% (each problem set is worth 9% of your grade)
- ▷ Final Project: 24%
- ▷ Participation: 11%
- ▷ Experiments: 2%

2.8 Matters of Etiquette

In this class, we treat each other with respect and compassion. That means we do not show annoyance at someone else's confusion, nor laugh at people's questions (unless intended).

We also don't interrupt class discussion by yelling something out out of turn; please raise your hands if you want to bring something to the class's attention. Because of the size of this class, you won't always be able to contribute something to a discussion. I'm sorry. If the issue is really pressing, mention it to your TA, come up after class, or come to office hours. I am always happy to relay an interesting remark to the rest of the class.

If you feel like you're being systematically marginalized, please let one of us know. This room is not conducive to larger lectures; I often can't see in the back. Regardless where you are sitting, it is never my intention to keep students from discussion!

Finally, a brief word about email: I try to respond to email as quickly as possible, but I have a life outside this classroom. Here are my promised response latencies: a) during the week, 24 hours; b) during the weekend, by Monday morning. That means, in particular, that I will guiltlessly file away hurried questions to me on Thursday evening or Friday morning. I'm sorry to be so harsh, but I found out the hard way that otherwise I'm besieged with messages while trying to eat breakfast.