Ling 571 Computational Corpus Linguistics

Schedule # 22058
Fall 2016
TuTh 9:30am–10:45am
Room NE-085

Advances in technology have revolutionized the way linguists approach their data. Using computers, extremely large bodies of text ("corpora") can be collected and analyzed at a level of detail that only a decade ago would have been unthinkable. Sources like the British National Corpus, the Corpus of Historical American English, and Google Books collection allow us access to language use across an unprecedented range of time and space. For anyone studying human communication or culture, the accelerating growth of the World Wide Web and other natural language resources have made techniques for dealing with very large texts more important than ever.

Through a combination of lectures, demonstrations, and hands-on exercises, this course will give students an introduction to the skills necessary for computer-aided text manipulation. Students will learn to search text databases using on-line tools, to write Python programs to manipulate large natural language corpora, to apply quantitative linguistic measures to existing texts, and to formulate, carry out, and describe their own corpus-based linguistic research projects.

This class has no pre-requisites.

Instructor

Rob Malouf

Office:	SHW 244
Office hours:	Tu 1:00–2:00, Th 8:30–9:30, or by appt
Email/GTalk:	`rmalouf@mail.sdsu.edu`
Web:	`malouf.sdsu.edu`

Requirements

The final grade will be based on homework assignments (30%), a midterm project (30%), and a final project (40%). Through the term, there will be several hands-on homework assignments in which

students apply the techniques learned in class to actual corpus materials. Since it's important to not get behind on assignments, late assignments will be accepted for partial credit **for one week only** after the due date unless prior arrangements are made. Working in groups is encouraged, but please include the names of all coworkers on the assignment.

The midterm will be a take-home programming assignment, for which students will be required to replicate a published corpus analysis using Python. The final project should be a program (with documentation) to perform some substantial corpus processing task chosen by the student. Alternatively, the final project can be the collection and annotation of a new corpus, or a research project that makes crucial use of novel corpus data. More details about both projects will be given later in the term.

No form of academic dishonesty, including cheating or plagiarism, will be tolerated in the class. Following Executive Order 1006, all instances of academic dishonesty will be reported to the Center for Student Rights and Responsibilities for investigation. For more information about the judicial process, see `http://csrr.sdsu.edu`. For more information about what plagiarism is and how to avoid it, see `http://its.sdsu.edu/tech/plagiarism.html`.

If you are a student with a disability and believe you will need accommodations for this class, it is your responsibility to contact Student Disability Services at (619) 594-6473. To avoid any delay in the receipt of your accommodations, you should contact Student Disability Services as soon as possible. Please note that accommodations are not retroactive, and that accommodations based upon disability cannot be provided until you have presented your instructor with an accommodation letter from Student Disability Services. Your cooperation is appreciated.

Readings

There are four required textbooks for this course:

- Martin Weisser. 2016. *Practical Corpus Linguistics: An Introduction To Corpus-Based Language Analysis.* Wiley.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Press.
- Allen B. Downey. 2012. *Think Python: How to Think Like a Computer Scientist.* Green Tea Press. `http://greenteapress.com/wp/think-python/`

- Martin Wynne (ed.). 2005. *Developing Linguistic Corpora: A Guide to Good Practice.* Oxbow Books. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

The first two books are available from the campus bookstore (or the other usual places) and you should do you best to get them. The last two are available free online and there's no reason to buy a printed copy.

Software

For the programming assignments in this class, we will be using a software package called Enthought Canopy (https://www.enthought.com/products/canopy/). You'll want to get the free Academic version: go to the Enthought website, click on 'Create account', and register using your SDSU email address. Once you've done that, you should be able to download the Academic version. Try this as soon as you can, and if it doesn't work please let me know immediately!

Proposed schedule

Week 1	Introduction
Week 2	Very large corpora
Week 3	Vocabulary
Week 4	Corpus construction
Week 5	Encoding
Week 6, 7	Annotation
Week 8	Python
Week 9, 10	N-grams
Week 11	Readability
Week 12	Collostructional analysis
Week 13	Syntax
Week 14	Web scraping
Week 15	Review