

COMM 620 Data Retrieval and Processing Techniques

Time: Wed 12:30-3:20p

Classroom: ANN 309

Instructor: Professor Lian Jian

Office: ANN 414A

eMail: ljian at usc

Office hours: Wed 10:30-11:30a or by appt

Introduction

This course is a research methods course that covers a new and increasingly popular method of conducting social science research: large scale data analysis. The advent of the Internet has enabled social scientists to have access to extremely large datasets about the behavior of millions of people, e.g., edits on Wikipedia, transactions on eBay, ratings on Yelp. However, collecting and analyzing this data isn't straightforward and requires some specific skills. The goal of this course is to expose PhD students to the skills required for this type of research, e.g., writing scripts for crawling and parsing data from websites or their API, collecting data into MySQL databases, and simple data visualization skills.

Pre-requisites

No prior programming background is required.

I will assume that you are somewhat familiar with a stats package, e.g., SPSS, R, or Stata, to do simple data plotting and tests.

Logistics

You need to bring your laptop to class.

Grading Scheme

60%	Eight short homework tasks (to be turned in on Blackboard by the following class)
30%	Presearch project
10%	Participation in class, chat, project group, and being nice to buddy classmates, etc.

Readings

We will use UDACITY online course "[Intro to Computer Science](#)" (create an account and then click "Start free course") as a resource for this course. Assigned segments from this course will be announced weekly.

As a recommended resource, this course [Introduction to Computer Science and Programming Using Python](#) is also good if you want additional material to learn Python

And here is one more cool book about Python for you [Automate the Boring Stuff with Python](#)

Homework Policy

Turning in means uploading to Blackboard before the next class. You are encouraged to work in groups, but you are required to write up the scripts, and run them independently.

Research Project

This course will be a project based course, so you will be required to complete a research project that involves collecting, manipulating, and analyzing large scale data. You are required to submit a one-page proposal (not graded) by Week 6, and a final paper due at the end of the term.

Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism --- presenting someone else's ideas as your own, either verbatim or recast in your own words --- is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in [SCampus in Section 11, Behavior Violating University Standards](#). Other forms of academic dishonesty are equally unacceptable. See [additional information in SCampus and university policies on scientific misconduct](#)

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the [Office of Equity and Diversity](#) or to the [Department of Public Safety](#). This is important for the safety whole USC community. Another member of the university community --- such as a friend, classmate, advisor, or faculty member --- can help initiate the report, or can initiate the report on behalf of another person. The [Center for Women and Men](#) provides 24/7 confidential support, and the sexual

assault resource center webpage sarc@usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the [American Language Institute](#), which sponsors courses and workshops specifically for international graduate students. The [Office of Disability Services and Programs](#) provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, [USC Emergency Information](#) will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.

Weekly Topics

Week 1 (1/13): Introduction

Week 2 (1/20): Strings and Lists

Week 3 (1/27): Lists and Dictionaries

Assignment 1 due

Week 4 (2/3): Functions and Loops

Assignment 2 due

Check-In Report #1 Due (a paragraph or two via email about what you have accomplished so far, where you are stuck, etc.)

Week 5 (2/10): Lists and Dictionaries Continued

Assignment 3 due

Week 6 (2/17): Database

Assignment 4 due

Week 7 (2/24): APIs

Assignment 5 due

Week 8 (3/2): HTML & Web Scraping

Assignment 6 due

Week 9 (3/9): Review/Practice of Python Basics

Assignment 7 due

Week 10 (3/16): Spring Break**Week 11 (3/23): Review/Practice of Python Basics**

Assignment 8 due

Check-In Report #2 Due

Week 12 (3/30): Review/Practice of Python Basics

Check-In Report #3 Due

Week 13 (4/6): 'Advanced' Review: Databases 1 hr / Project Time 2 hrs

Check-In Report #4 Due

Week 14 (4/13): 'Advanced' Topics: Functions 1 hr / Project Time 2 hrs

Check-In Report #5 Due

Week 15 (4/20): 'Advanced' Topics: Regex 1 hr / Project Time 2hrs

Check-In Report #6 Due

Week 16 (4/27): Presentations

For the presentation, each person gets about 20 minutes. Tell use what your hypotheses or RQ were, where and how you got the data, and your preliminary results.

Final paper due on Blackboard on the day of final exam

It requires a lit review, hypotheses/RQ, method (describing your data collection in a little more detail than usual), and preliminary results.