COMM 620 Data Retrieval and Processing Techniques

Time: Thu 3:30-6:20p

Classroom: ANN 408

Instructor: Professor Lian Jian

Office: ANN 414A

eMail: ljian at usc

Office hours: 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.

The final paper (double-space, 25 pages max) requires a lit review, hypotheses/RQ, method (describing your data collection in a little more detail than usual), and preliminary results.

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 (1/12): Introduction

- Introduce different sources of data: SQL, API, Scraping HTML
- Requirement for term project

Homework

- Get familiar with iPython Notebook environment, by reading this page
- Go through Unix tutorial (Mac) if you use a Mac
- Go through DOS tutorial (PC) if you use a PC
- Lesson 1 on Udacity. Please do the exercises, quizes, they are well designed.

Week 2 (1/19): Python Basics --- Strings and Lists

- Operations on strings: concatenation, search, find, replace, converting integers to strings, output
- Operations on lists: creating lists, indexing, looping through lists, adding and deleting elements

Homework

- Lesson 2 and 3 on Udacity. Please do the exercises, quizes, they are well designed.

Week 3 (1/26): Python Basics --- Lists and Dictionaries

- Operations on lists: reading data files in CSV format into lists, and loop through lists
- Operations on dictionaries: creating dictionaries, adding elements, looping through dictionaries, nested dictionaries

- Assignment 1 due, including exercises practicing string operations

Week 4 (2/2): Python Basics --- Functions and Loops

- Functions and loops on lists and dictionaries: counting in lists and dictionaries, using dictionaries as lookup tables

- Assignment 2 due, focusing on operations on lists
- Check-In Report #1 Due (a pargraph or two via email about what you have accomplished so far, where you are stuck, etc.)

Week 5 (2/9): Python Basics --- Lists and Dictionaries Continued

- Hands-on practice on counting with dictionaries, with data files from IMDB

- Assignment 3 due, containing more exercises with real data files from Reddit

Homework

- Go through W3schools MySQL tutorial
- Go through a more advanced MySQL tutorial

Week 6 (2/16): Review/Practice of Python Basics

- Review questions from all assignments
- Discuss data structure that might be needed in students' projects

- Assignment 4 due, focusing on combining lists and dictionaries to construct datasets using the top 250 movies on IMDB

Week 7 (2/23): APIs

- Using Python to access APIs such as those for Twitter, Facebook, Tumblr etc.
- Using Python process data from APIs

- Assignment 5 due, containing a complex set of data manipulation with IMDB data files on countries, actors, genres etc.

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

- HTML file format introduction
- Using Python scrape websites, in particular using the Beautiful Soup module
- Conceputalize the data structure used for storing data from websites
- Scheduled, periodic scraping

- Assignment 6 due, API

Week 9 (3/9): MySQL Database

- Conceptual understanding of relational databases
- Practice using Terminal to access MySQL database
- Practice using GUI tools to access MySQL database
- Practice using Python to access MySQL database

- Assignment 7 due, webscraping

Week 10 (3/16): Spring Break

Week 11 (3/23): Review/Practice of Python Basics

- Answer students' questions
- Special topics such as dealing with dates, counting, unicodes

- Assignment 8 due, SQL database

Check-In Report #2 Due

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

- Exercises on debugging techniques

Check-In Report #3 Due

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

- Advanced MySql queries, such as subqueries, statistics, sampling, indexing

Check-In Report #4 Due

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

- Writing customized functions to perform special tasks

Check-In Report #5 Due

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

- Regular expressions, to be used with text parsing and natural language processing

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 at 5pm on May 9, 2017