Syllabus

THE UNIVERSITY OF CHICAGO
The Harris School of Public Policy

PPHA 30520
Text Mining in Public Policy
Spring 2018
Tuesday 3-5:50

Robert M. Goerge, Ph.D.
Senior Fellow, Harris School and Computation Institute
Executive Director, MSCAPP
rgoerge@uchicago.edu

Jonathan Ozik, Ph.D.
Senior Fellow, Computation Institute
Computational Scientist, Argonne National Laboratory
jozik@uchicago.edu

Office hours by appointment

The purpose of the class is to provide the MPP student a set of skills that can be used in their professional career to compile and analyze unstructured data. If a public policy analyst is asked by the chief of staff to compile everything that is known about managed care for elderly patients, the analyst should be able to mine the available resources on the web to pull out a set of abstracts, data, information, or policy recommendations so that he or she can have a report done in hours rather than weeks. If a congressman asks his aid to pull information on earmarks on a particular topic from the United State Public Laws, they should be able to do that in a few hours after taking this class.

The class will address the collection of data (webscraping) as well as basic and more advanced mining of the text. The first weeks of the class will include lectures and exercises building datasets that are relevant for public policy. The middle part of the class will focus on writing code to mine and analyze current policy questions. The final weeks of the class will focus on student projects, individual or group, making use of the datasets developed throughout the class. These projects will be presented to the entire class and a paper will be required.

A group project will be available to students. The topic will be the Temporary Assistance of Needy Families(TANF)--the major US policy for poor families. Students will also have the opportunities to choose their own policy areas.

Prerequisites

We require some experience with Python programming. Ideally, this would be a formal class, such as Introduction to Python Programming 35550. We will accept other experience. Those students who are concerned about their preparation should contact the instructors.

Students will need to bring a laptop to every class. We will use the Anaconda distribution (https://www.anaconda.com).

Grading

Pass/Fail allowed. The final grade will be based on: 1) class attendance and participation in discussions (20%); 2) one page prospectus of the final project (due on 4/24) (10%) ; 3) programming exercises (30%); 4) classroom presentation (15%); and 5) final project (25%).

Course Reading and Materials

The course will utilize primarily online materials.

For webscraping, we will be using standard Python libraries (request: http://docs.python-requests.org and lxml: http://lxml.de) as well as the Scrapy (https://doc.scrapy.org) framework.

For natural language processing, we will be using the Python NLTK library (http://www.nltk.org/), materials from the online NLTK book (http://www.nltk.org/book/) and additional sources for more advanced topics.

Course Schedule

March 27	Introduction Syllabus Big Data in Public Policy How data- and text-mining are appropriate for public policy Public Policy Analysis, Research and Decision Making Technical requirements for the class Hands on: Computing setup Introduction to scraping web content
April 3	Corpora in Public Policy and the Social Sciences Introduction to TANF Hands on:

	Scraping web content HW: Programming Exercises 1
April 10	Hands on: More scraping web content Automatic Natural Language Understanding and Accessing Corpora Readings: NLTK Chapter 0, Chapter 1, Chapter 2 (Sections 1, 2) HW: Programming Exercises 2 Due: Programming Exercises 1
April 17	Guest speaker: TBD Hands on: More Python and Lexical Resources Accessing Text and Regular Expressions Readings: NLTK Chapter 2 (Sections 3-end), Chapter 3 (Sections 1-4) HW: Programming Exercises 3 Due: Programming Exercises 2
April 24	Guest Speaker: TBD Hands on: More Regular Expressions Writing Structured Programs Readings: NLTK Chapter 3 (Sections 5-end [skip 8]), Chapter 4 (Sections all [skip 5, 7])

	HW: Programming Exercises 4 Due: Final Project Prospectus Programming Exercises 3
May 1	Lecture: Text-mining of social program data Hands on: Tagging Readings: NLTK Chapter 5 (Sections all) HW: Programming Exercises 5 Due: Programming Exercises 4
May 8	Hands on: Classifying text Readings: NLTK Chapter 6 (Sections all) HW: Programming Exercises 6 Due: Programming Exercises 5
May 15	Hands on: Advanced topics I Readings: TBD Due: Programming Exercises 6

May 22

Hands on:
Advanced topics II

Readings:
TBD

May 29

Student Projects and Presentations
