Web scraping and text processing with Python

Hubert Jin
hubertj@princeton.edu
Will Lowe
wlowe@princeton.edu
Q-APS, Department of Politics

AM session: 10:00-12:00
PM session: 1:00-3:00
May 26-27, 2016

This workshop introduces basic tools and techniques in the Python language for automated extraction of content from the web, natural language parsing, and other data-handling tasks that are commonly encountered in data-intensive research projects. We will show how to use the library `beautifulsoup` and the browser emulator `mechanize` to extract web content, including authentication and cookie handling. In addition, we will also give brief introduction to other alternative data scraping and text processing tools, as well as some NLP packages.

Workshop website Google sign-up

Prerequisites

The course requires the basic knowledge of Python programming and access to a laptop.

Setup

You will need to be able to install software on a laptop. We will be using the Anaconda python distribution and will provide installation instructions before the course.

All materials will be available on the course Blackboard.

Structure

Day	Session	Details
Thursday	10-12pm	The structure of webpages. Introduction to `beautifulsoup`
	1-3pm	Web data scraping with `beautifulsoup`
Friday	10-12pm	Authentication and cookies Introduction to `mechanize`
	1-3pm	Alternative data scraping and text processing tools