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

<table>
<thead>
<tr>
<th>Day</th>
<th>Session</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thursday</td>
<td>10-12pm</td>
<td>The structure of webpages. Introduction to BeautifulSoup</td>
</tr>
<tr>
<td></td>
<td>1-3pm</td>
<td>Web data scraping with BeautifulSoup</td>
</tr>
<tr>
<td>Friday</td>
<td>10-12pm</td>
<td>Authentication and cookies Introduction to mechanize</td>
</tr>
<tr>
<td></td>
<td>1-3pm</td>
<td>Alternative data scraping and text processing tools</td>
</tr>
</tbody>
</table>