Lecture 12 – Data Collection: Webscraping
Announcements – Assignments

- Readings 04:
  - link posted to course site later today
  - due Sunday

- HW 02:
  - Due Wednesday night

- Tutorial 4.1:
  - Releasing tonight or tomorrow
  - Topic Modeling for today’s data

- Office hours
  - Today 1:30-3:00 pm
Each post in this channel will correspond to a specific assignment. Please add any reaction to the post if you want me to grant you access to the solutions for that assignment.

Tutorial 1.1
Tutorial 1.2
Tutorial 1.3
Tutorial 2.1
Homework 01
Course Feedback - Optional

- Mid-semester anonymous brief survey

- What have you learned so far and how comfortable do you feel with the material?

- What has been going well in the course so far? What are things you are enjoying about the course?

- What has not been going well in the course so far? What are things you are not enjoying about the course?

- What can we (the course staff) be doing better?
Course Outline

- **Python Overview**
  Week 1

- **Lexical based analysis methods**
  Week 2 - 3
  - Text Processing
  - Document Representation
  - Topic Modeling

- **Data Collection**
  Week 4
  - Web Scraping
  - APIs

- **Machine Learning**
  Week 5
  - Regression & Classification
  - Clustering

- **Advanced Topics & Final Projects**
  Week 6
  - Dimensionality Reduction
  - Word Representations
Final Project – Deliverables

- Project ideation – Friday May 28\textsuperscript{st}
- Project proposal – Friday June 4\textsuperscript{th}
- Project presentations – Monday June 14\textsuperscript{th}
- Project submissions – Friday June 18\textsuperscript{th}

http://coms2710.barnard.edu/final_project
Project ideation – Friday May 28th

- [https://www.overleaf.com/read/yzpgxcgsqdvp](https://www.overleaf.com/read/yzpgxcgsqdvp)

- roughly 250 word overview of what you are interested in
Newspapers

  - [https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/GZC6PL](https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/GZC6PL)

- **LexisNexis: need an account**
  - You can experiment with: [https://github.com/ahalterman/cloacina](https://github.com/ahalterman/cloacina)

Copyright © 2016 Barnard College

Slide from Federico Nanni
Political Speeches

- EuroParl:
  - http://www.talkofeurope.eu/data/

- UK
  - https://www.hansard-corpus.org/

- US Congress
  - https://www.congress.gov/

Slide from Federico Nanni
Manifesto’s, Speeches, Archived Collections

- Manifesto project: https://manifestoproyect.wzb.eu/
- US Presidency http://www.presidency.ucsb.edu/
- Web Archives https://www.loc.gov/collections/united-states-elections-web-archive/

Slide from Federico Nanni
Books

- **Project Gutenberg**
  - https://www.gutenberg.org/

- **HathiTrust Digital Library** - https://www.hathitrust.org/
  - Python tool
    - https://github.com/htrc/htrc-feature-reader
  - Tutorial
  - Examples
The web:

- largest & most diverse collection of information in history
- Rich corpus for scientific research, technological advancement, and innovative new businesses
- A digital copy of our world

Using the web to:

- insight into politics, art, economics, health, culture and almost every other aspects of life.

Purpose of Common Crawl:

- Make the web be openly accessible to anyone who desires to utilize it
Interacting with the Internet
HTTP

- HyperText Transfer Protocol
Interact with websites via Request & Responses:

- **Request:**
  - an operation to be performed on a URL

- **Response:**
  - Message from server based on a request
Web Scraping & Crawling
Definitions

- **Scraping:**
  - *Using tools to gather data you can see on a webpage*

- **Crawling:**
  - *Moving across or through a website in an attempt to gather data from more than one URL or page*

- **HTML:**
  - *HyperText Markup Language*
  - The standard markup language on the Web

Definitions from [John Little](#)
HTML: *HyperText Markup Language*

- **HTML tags** to represent different elements on a web page

- Structured as a tree

```html
<html>
  <head>
    <title>Google</title>
    <style>...</style>
  </head>
  <body>
    <div>...</div>
    <center>...</center>
    <div>...</div>
  </body>
</html>
```
## HTML Tags

<table>
<thead>
<tr>
<th>HTML Tag</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;!DOCTYPE&gt;</td>
<td>Defines document type</td>
</tr>
<tr>
<td>&lt;html&gt;</td>
<td>Defines HTML document</td>
</tr>
<tr>
<td>&lt;head&gt;</td>
<td>Main information about document</td>
</tr>
<tr>
<td>&lt;title&gt;</td>
<td>Title for document</td>
</tr>
<tr>
<td>&lt;body&gt;</td>
<td>Document body</td>
</tr>
<tr>
<td>&lt;h1&gt; to &lt;h6&gt;</td>
<td>Headings</td>
</tr>
<tr>
<td>&lt;p&gt;</td>
<td>Paragraph</td>
</tr>
<tr>
<td>&lt;br&gt;</td>
<td>Line break</td>
</tr>
<tr>
<td>&lt;!--comment here--&gt;</td>
<td>Comment</td>
</tr>
<tr>
<td>&lt;img&gt;</td>
<td>Image</td>
</tr>
<tr>
<td>&lt;a&gt;</td>
<td>Hyperlink</td>
</tr>
<tr>
<td>&lt;ul&gt;</td>
<td>Unordered list</td>
</tr>
<tr>
<td>&lt;ol&gt;</td>
<td>Ordered list</td>
</tr>
<tr>
<td>&lt;li&gt;</td>
<td>List item</td>
</tr>
<tr>
<td>&lt;style&gt;</td>
<td>Style information for a document</td>
</tr>
<tr>
<td>&lt;div&gt;</td>
<td>Section in a document</td>
</tr>
<tr>
<td>&lt;span&gt;</td>
<td>Section in a document</td>
</tr>
</tbody>
</table>

Table from Melanie Walsh’s textbook
• Selectors:
  • Class – can apply to multiple elements
  • Id – unique to an element

• Attributes
  • url: <href>
  • Image
  • Style
BeautifulSoup

- Python library for parsing HTML (and XML)

- Documentation: