WEB SCRAPING

CP255 Fall 2015

HOW TO GET WEB DATA

- Download formal data set from a web site or FTP server
- Use an API
- Web scraping

TERMINOLOGY

- **Spider** – crawls the web by following links
- **Crawler** – just another name for a spider
- **Data Scraper** – generic computer program that extracts human-readable data
- **Web Scraper** – a data scraper specifically for web pages

WHAT IS A WEB SCRAPER?

- Small computer program that:
  - accesses web pages
  - finds specified data elements on the page
  - extracts them (and transforms them if necessary)
  - compiles this data into a coherent data set

- Compare this behavior to that of a web browser

WHAT IS A WEB SCRAPER?

- Can be run iteratively over many web pages

- Data spread across thousands or millions of pages

- Construct large, robust data sets out of otherwise messy text that would only appear in your web browser

DIY WEB SCRAPING

- Stuff from the *server* side
  - HTML and XML
  - Javascript (beware of AJAX)

- Stuff on *your* side
  - Python
  - Scrapy framework
  - Xpath query language

Scrapy

- 2 key components for you to customize to run your own scrapy scraper: the data model and the spider

- Data model
  - defines the fields/columns for our data set

- Spider
  - URLs to fetch
  - data cleaning functions
  - parser

SCRAPY PROCESS

- Scrapy visits each URL in the list
- It uses XPath to find each data element you want
- It extracts the data and saves in memory
- It repeats process until all URLs are visited
- Lastly, it saves the whole data set as a CSV file