COGS 9 – A05 Discussion

Deadlines

- Reading Quiz 2 – October 13th (Yesterday), Late submission till 15th
- Assignment 1 – October 14th (Today), Late submission till 16th
- Reading Quiz 3 – October 20th (Next Thursday)
  - Readings for data cleaning (Goes in hand with Lecture 07)

Getting data

- Download button (Easy)
  - Find your dataset, click download and you're done
- APIs (Medium)
  - Choose method
  - Build URL
  - Get Authorization/Authentication
- Web Scraping (Hard)
  - Configure a crawler/spider to pull required HTML pages
  - Look through the page for your required information
  - Tidy the data

API DEMO

- Basic idea on terminal/shell and shell commands
- curl -X get http://files.rcsb.org/download/10mh.pdb
- Methods: get, put, post, patch, delete (For web)
- URL: The URL from where you want information
- Authorization/Authentication: Bearer tokens, access tokens
- APIs do not necessarily mean web APIs, can be APIs of a python module as well

Web Scraping Demo

- Require a good amount of configuration
- Scrapy is an example of a scraping framework for python
- It uses a spider that systematically goes through the websites you have configured in an ordered manner
- A spider is just a bot that gets you the required data, it is your job to parse and clean the data

Wrangling data with Pandas Demo