Web Scraping

Fundamentals of Computing and Data Display

10/2/2024

APIs

- Google Trends data will “change” if you pull at two different times. This is because Google adds noise to hide the exact numbers!
- **API documentation: Key aspects to look for**
  - **Base URL**. Sometimes, it is easier to look at examples to find what the base URL is.
  - **Query parameters**. Again, it might be easier to look at examples to guide you.
  - **API Key**. This will determine whether you need an API key to access it or not.

Web Scraping

- Think of this as collecting data by writing down observations about the world.
- Sometimes, the data you want is in tables already … but many times it is not.
- Many times, it is very raw data and needs to be cleaned.

APIs vs. Web Scraping

- Though both are web data, the two are actually quite different in how they collect data.
- **APIs**
  - Curated by a data provider to share data
  - Standard method of accessing data
  - Relatively clean (after converting formats)
- **Web Scraping**
  - Can be used on websites generally, not just from data providers
  - Need to identify parts to scrape from HTML structure
  - Lots of cleaning needed

Downsides to Web Scraping

- Websites can change
- Hard to generalize sometimes
  - If the pages are similar from the same source, then it should be possible to automate. Harder to do for websites from many different sources (such as grabbing mission statements from webpages of nonprofit organizations).
- Hard to clean (sometimes)

Storing Data

- Long term, the best solution for storing large amounts of data is to use a **database**.
- R stores all data in **memory**. This means that you are limited by your **RAM**.

- Ideal workflow:
  - Use R (or Python) to scrape/pull/collect a portion of the data.
  - Store within a database such as PostgreSQL/SQLite/etc.
  - Manage data using SQL
  - Query the database and get data that you need for analysis within R for further analysis.

Dynamic Web Pages

- More difficult to do
- Can use Selenium
- Basically navigate the webpage and scrape data as you navigate.

Privacy and Restrictions

- Many times, websites will limit what you can scrape to protect their own interests.
  - Sales data being easily scrapable might lead to undercutting prices.
- Generally, if you are able to find data on the web, that means it is publicly available (unless, of course, you had to log in or access it somehow).
- Fewer concerns … but more when you link with other datasets.