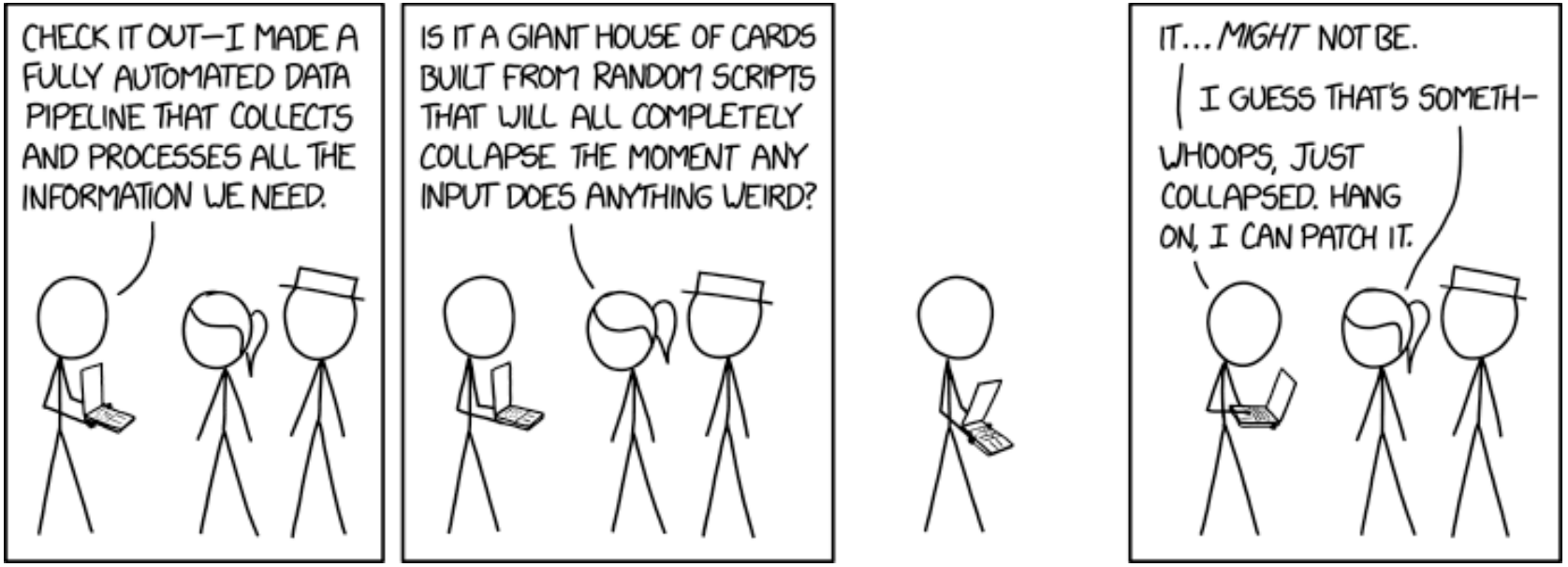
CS 506 Guest Lecture

Web Scraping

Today's Agenda

- Web scraping overview
- When *not* to scrape
- Common web scraping
- Advanced web scraping
- Diabolical web scraping
- Scraping when all other techniques fail
- A note on the legality of web scraping and robots.txt (be nice!)

Web Scraping



When not to use web scraping

- Data available from other sources (e.g. data portals, data aggregators, metadata portals, etc.)
- Underlying DB/API available (e.g. Venmo, Twitter, Reddit, even BU)
- Let us look at an example with BU's course bulletin

Quick Overview of Web Scraping

- What exactly is it?
- Why do we care?
- How do we do it?

Simple Scraping

- With bot-friendly and sensible websites, scraping is trivial. See: https://scrapy.org or https://webscraper.io/
- With forums and other applications running on known technologies:
  - Figure out what it is built with
  - Use GitHub/GitLab/BitBucket to find scrapers for those types of forums
- Let us look at a demo:
  - https://forums.nexusmods.com/ (what is it built with? Wappalyzer, builtwith)
  - Can we find scrapers for this?

Advanced Scraping

- Sometimes, websites are not happy until you are not happy
- Intent notwithstanding, accessibility is often sacrificed for fancy rendering (oh hello, Facebook), bot detection (hello again, Facebook), or because of sheer incompetence
- How do we scrape these websites?

Dharmesh's Law of Scraping

"Every website that uses computers to verify the authenticity of human users can be scraped."

Corollary: All human browsing activity can be convincingly emulated by a clever program, given enough time.

Advanced Scraping

- First understand that scraping websites that do not want to be scraped is an arms race
- Second, try your best to procure the information through other methods (e.g. a lot of websites that dissuade scraping will readily make data available to academics)
- If asking nicely doesn't work, start by doing cursory reconnaissance:
  - What framework/libraries is the website using?
  - Are there extant scrapers you can use? Try looking at robots.txt and obtaining a sitemap if you can
  - Does web.archive.org contain the page you need? The Web Archive might be easier to scrape. Have you tried other search engines?

Advanced Scraping: Recon

- How are simple scrapers being identified?
  - First line of defence is robots.txt. Respectful bots will not crawl endpoints declared out of bounds
  - Second line of defence is dynamic page rendering and simple bot detection
  - Advanced anti-scraping techniques will use anomaly detection ML techniques to identify bots and ban them
- Let us see how we can beat advanced anti-scraping techniques

Screen Scraping

- Formally, collection of visual data from a source
- Difficulty can range from laughably simple to practically impossible
- Let us see a simple demo (IMDB Reviews)

Screen Scraping Difficult Websites

- Websites detect unusual activity through fingerprinting. Advanced scraping requires that you try and “blend in” to the cornucopia of regular user activity
- Practically, this might mean:
  - Using constantly changing IP addresses if the website does not authenticate users (use AWS Lambdas or scrape over TOR)
  - Configuring your scraper to emulate user activity (think about all the things you do when browsing the internet on your computers and code that behavior into your program)

Diabolical Scraping: The reCAPTCHA Bugbear

- As a general rule, accessibility is your friend. Websites are now required by law to be accessible to users with disabilities. This works in your favour because all you now have to do is scrape those versions of the website
- Practically, reCAPTCHA challenges are difficult/almost impossible for visually impaired users to solve. This has led to several people building modules that can get around that problem. Can we leverage these modules/techniques? Let us try a demo
- PS: Speech transcription is *vastly* easier than image recognition
- Sometimes, disabling JS completely can work wonders

Scraping when all else goes wrong

- If after a lot of effort you still cannot scrape a website, what other options are at your disposal?
- Use humans:
  - Option 1: show up to office hours. Pros: Free and easy. Cons: Quality of help may vary depending on how much you understand what your instructors are talking about
  - Option 2: hire humans on services like Amazon's mTurk. Pros: relatively inexpensive. Cons: time consuming and quality of data gathered may suffer somewhat
  - Option 3: hire a developer to scrape websites for you. Pros: guaranteed quality and relatively quick turnaround. Cons: expensive

How can I scrape Facebook?

- After Cambridge Analytica, Facebook have gotten very good at restricting API access
- If you have a legacy app from 2017 or earlier, you're in luck, since legacy APIs have still not been deprecated!
- Alternatively, https://pypi.org/project/facebook-scraper/

Resources and Questions

- Set of scrapers: http://www.schrenk.com/nostarch/webbots/DSP_download.php
- Questions?