Steven Weil, Aaron Anderson

Professor Adjei

DSC 200

11/26/2024

Web Scraping Report

Web scraping is one of the most important skills that every data scientist should master. There are often times when you are tasked with solving research questions by finding data on the internet. Scraping can make that task substantially more efficient. It is a process where you design a script to parse the HTML content of a page and extract the necessary information. Automating this process allows data scientists to download and collect exponentially more data for their research. For the web scraping in this assignment, we will be using different python libraries such as BeautifulSoup, Selenium, and requests. These libraries make web scraping through python possible and are some of the most impactful tools we have learned in this class. However, web scraping also comes with a lot of responsibility. Many sites disallow access to certain webpages, so you must be sure to look at the robots.txt file and terms of service before scraping. Also, to avoid overloading the server, make sure to add pauses into your scraping script. Scraping is an extremely powerful tool, but just like anything, it must be done correctly and thoughtfully.

The first webpage that we were interested in scraping was CollegeSimply’s Kentucky college tuition page. This page can be found at <https://www.collegesimply.com/colleges/kentucky/college-cost/>. Overall, CollegeSimply is a website that provides information about colleges all over the country for prospective students and their families. We selected the Kentucky tuition webpage because it contains information that was extremely valuable to us back when we were searching for colleges. However, it is all separated in contained to each university, so being able to scrape the data and put it into one csv would make comparing universities much more efficient. The web scraping policies that CollegeSimply included on their site could be found in the robots.txt file and on a terms of use page. The robots.txt file is a page that outlines the domain paths that are disallowed for scraping. Almost all websites have a robot.txt file and checking to make sure your page is allowed is a crucial step of the process. CollegeSimply indicates that the use of certain scraping bots is not allowed, but not any of the ones we learned in class. Also, it disallows scraping from any domain that includes /sort, but we also do not have that in our page. Finally, the terms of use page did not include any information about scraping, so it is clear to scrape in a responsible way.

The second website that we are interested in scraping is Amazon. With the holiday season approaching, it could be valuable to scrape data about the different products you could buy as presents to store and compare them yourself. For this example, we chose Kentucky basketball gear as the search term, so the domain of the page was <https://www.amazon.com/s?k=kentucky+basketball+gear&crid=349JLIBRA51GS&sprefix=%2Caps%2C106&ref=nb_sb_ss_recent_1_0_recent>. The robots.txt file was much more intricate for Amazon, but there were no disallows for the path we chose. Similarly, to the CollegeSimply robots.txt file it also noted that certain bots were not allowed, such as GPTBot, and CCBot. The terms of use page did not mention that scraping was prohibited, so we are able to scrape responsibly.

Overall, this assignment provided a great opportunity for us to learn more about web scraping and how powerful it can be. However, it also taught us the limitations and regulations around web scraping and why it is so important to follow them. There were many sites that we planned on scraping, only to later find out they did not allow scraping. Our first idea was to scrape IMDB for movie data, but they had a strict policy that prevented any data from being scraped from their website. The only permitted use was limited data being used personally that was collected manually. It is important to respect the rules around scraping as digital citizens and data scientists. Web scraping can be one of the best tools we have access to, but it must be used correctly. This assignment was a great way to learn that and continue practicing our skills in web scraping!

Appendix

**CollegeSimply robots.txt file:**User-agent: \*

Disallow: /\*sort\*

User-agent: MJ12bot

Disallow: /

User-agent: AwarioRssBot

User-agent: AwarioSmartBot

Disallow: /

User-agent: NTENTbot

Disallow: /

User-agent: serpstatbot

Disallow: /

User-agent: YandexBot

User-agent: Yandex

Disallow: /

User-agent: Bleriot

Disallow: /

User-agent: Cliqzbot

Disallow: /

User-agent: PetalBot

Disallow: /

User-agent: dotbot

Disallow: /

User-agent: BLEXBot

Disallow: /

User-agent: adbeat\_bot

Disallow: /

User-agent: MojeekBot

Disallow: /

User-agent: AccompanyBot

Disallow: /

**Amazon robots.txt file:**

User-agent: \*

Disallow: /exec/obidos/account-access-login

Disallow: /exec/obidos/change-style

Disallow: /exec/obidos/flex-sign-in

Disallow: /exec/obidos/handle-buy-box

Disallow: /exec/obidos/tg/cm/member/

Disallow: /gp/aw/help/id=sss

Disallow: /gp/cart

Disallow: /gp/flex

Disallow: /gp/product/e-mail-friend

Disallow: /gp/product/product-availability

Disallow: /gp/product/rate-this-item

Disallow: /gp/sign-in

Disallow: /gp/reader

Disallow: /gp/sitbv3/reader

Disallow: /gp/richpub/syltguides/create

Disallow: /gp/gfix

Disallow: /gp/associations/wizard.html

Disallow: /gp/dmusic/order

Disallow: /gp/legacy-handle-buy-box.html

Disallow: /gp/aws/ssop

Disallow: /gp/yourstore

Disallow: /gp/gift-central/organizer/add-wishlist

Disallow: /gp/vote

Disallow: /gp/voting/

Disallow: /gp/music/wma-pop-up

Disallow: /gp/customer-images

Disallow: /gp/richpub/listmania/createpipeline

Disallow: /gp/content-form

Disallow: /gp/pdp/invitation/invite

Disallow: /gp/customer-reviews/common/du

Disallow: /gp/customer-reviews/write-a-review.html

Disallow: /gp/associations/wizard.html

Disallow: /gp/music/clipserve

Disallow: /gp/customer-media/upload

Disallow: /gp/history

Disallow: /gp/item-dispatch

Disallow: /gp/dmusic/order/handle-buy-box.html

Disallow: /gp/recsradio

Disallow: /gp/slredirect

Disallow: /dp/shipping/

Disallow: /dp/twister-update/

Disallow: /dp/manual-submit/

Disallow: /dp/e-mail-friend/

Disallow: /dp/product-availability/

Disallow: /dp/rate-this-item/

Disallow: /gp/registry/wishlist/\*/reserve

Disallow: /gp/structured-ratings/actions/get-experience.html

Disallow: /gp/twitter/

Disallow: /ap/signin

Disallow: /gp/registry/wishlist/

Disallow: /wishlist/

Allow: /wishlist/universal\*

Allow: /wishlist/vendor-button\*

Allow: /wishlist/get-button\*

Disallow: /gp/wishlist/

Allow: /gp/wishlist/universal\*

Allow: /gp/wishlist/vendor-button\*

Allow: /gp/wishlist/ipad-install\*

Disallow: /registry/wishlist/

Disallow: /review/common/du

Disallow: /gp/registry/search.html

Disallow: /product-reviews/B0069IY63Y

Disallow: /gp/orc/rml/

Disallow: \*/gcrnsts

Disallow: /gp/gc/widget

Disallow: /gp/dmusic/mp3/player

Disallow: /gp/entity-alert/external

Disallow: /gp/customer-reviews/dynamic/sims-box

Disallow: /review/dynamic/sims-box

Disallow: /gp/redirect.html

Disallow: /gp/twister/ajaxv2

Disallow: /ss/twister/ajax

Disallow: /b?\*node=7454917011

Disallow: /b?\*node=7454927011

Disallow: /b?\*node=7454939011

Disallow: /b?\*node=7454898011

Disallow: /gp/customer-media/actions/delete/

Disallow: /gp/customer-media/actions/edit-caption/

Disallow: /gp/dmusic/

Allow: /gp/dmusic/promotions/PrimeMusic

Allow: /gp/dmusic/promotions/AmazonMusicUnlimited

Disallow: /gp/offer-listing/

Disallow: /b?\*node=9052533011

Disallow: /lm/R1XIHQVKXSKBNJ

Disallow: /lm/R3HQ5WJSZK6QSO

Disallow: /surprise/

Disallow: /local/ajax/

Disallow: /gp/socialmedia/giveaways

Disallow: /gp/b2b-rd

Disallow: /gp/aw/so.html

Disallow: /reviews/iframe

Disallow: /gp/switch-language

Disallow: /ga/p/

Disallow: /gp/profile/

Disallow: /giveaway/host/setup/

Disallow: /ss/customer-reviews/lighthouse/

Disallow: /ospublishing/story/\*

Disallow: /gp/aw/ol/

Disallow: /gp/promotion/

Disallow: /hz/leaderboard/top-reviewers/

Disallow: /creatorhub

Disallow: /creatorhub/\*

Disallow: /slp/s$

Disallow: /-/

Allow: /-/es/

Allow: /-/en$

Allow: /-/zh\_TW/

Allow: /-/zh\_TW$

Allow: /-/he/

Allow: /-/he$

Disallow: /hz/help/contact/\*/message/$

Disallow: /gp/aw/shoppingAids/

Disallow: /rss/people/\*/reviews

Disallow: /gp/pdp/rss/\*/reviews

Disallow: /gp/cdp/member-reviews/

Disallow: /gp/aw/cr/

Allow: /gp/offer-listing/B000

Allow: /gp/offer-listing/9000

Disallow: /gp/aag

Allow: /gp/aag/main?\*seller=ABVFEJU8LS620

Disallow: /gp/pdp/profile/

Disallow: /gp/help/customer/express/c2c/

Disallow: /slp/\*/b$

Disallow: /hz/contact-us/ajax/initiate-trusted-contact/

Disallow: /hp/video/

Disallow: /gp/video/api

Disallow: /gp/video/settings

Disallow: /gp/video/library

Disallow: /gp/video/watchlist

Disallow: /gp/video/mystuff

Disallow: /gp/video/profiles

Disallow: /gp/video/search

Disallow: /gp/video/auth

Disallow: /gp/video/salp/i

User-agent: EtaoSpider

Disallow: /

User-agent: GPTBot

Disallow: /

User-agent: CCBot

Disallow: /

User-Agent: PerplexityBot

Disallow: /

Links to terms of use pages:

<https://collegesimply.com/terms/>

<https://www.amazon.com/gp/help/customer/display.html?nodeId=202140280>