What are the ethics of web scraping? How do we know when it's allowed vs not allowed? What are possible legal consequences of unauthorized web scraping?

Web scraping is an activity that hasn't been given much attention by laws and regulations around the country as of yet. It is still a relatively new method of grabbing data from the internet and using it to gain insights and make conclusions. However, there are certain ethical concerns when pulling or scraping data from the internet, and these concerns must certainly be paid attention to.

One consideration of ethical web scraping is that the data scraped must be publicly available. There is data on the internet which has been 'leaked' or is actually private information, and this is certainly not OK to use for a project that involves web scraping. Private information or 'leaked' information is not available for public use.

Another concern of the ethics of web scraping is that the purpose of the web scraping. Certain bots or web scraping applications are used nefariously, for email spamming, competitive data mining, data theft, spam, and more. These are all unethical uses for web scraping and should be reported if seen anywhere.

Web scraping has been slowly and slowly getting more and more scrutinized by the court system. In 2000, eBay filed an injunction against "Bidder's Edge" for using bots to gain an unfair advantage during the bidding process of eBay sales. This lawsuit ended up getting settled out of the court system, but began the court system cracking down on the activity. However, there was a case "Intel v. Hamidi" which reversed the decision of the eBay case, and made it free reign again. Since then, more and more cases have been ruled out of these "fair use clauses" and have been convicted of unethical web scraping. Individuals can also be convicted of hacking through unethical web scraping practices. There are many legal ramifications of unethical web scraping, so it's best to ensure the data found is publicly available and not accidentally leaked.

What are some of the technologies we can use for web scraping?

There are many tools out in the world which help to facilitate web scraping and make it easy for everyday data scientists to easily get publicly available data from the web. Usually the easiest solution to getting data from the web is utilizing an API, or Application Programming Interface, to execute specific commands and requests on a website or web application. ScrapingBee is a web scraping API like this and assists in getting data without getting blocked. ScrapingBee works well with JavaScript. Another web scraping tool is DiffBot. DiffBot offers multiple APIs that can scrape data from products, articles, discussions, and other web pages. However, DiffBot is quite expensive. According to a ScrapingBee article, the cheapest plan begins at $299 per month. For SEO (Search Engine Optimization) professionals, there is a tool specifically catered to them called ScrapeBox. It is a desktop software and is low-cost. ScrapeBox is said to have many capabilities related to SEO tasking. For those without much or any programming abilities, there is an enterprise web scraping platform called Import.io which is a no-code or low-code solution. It is said to have a great and easy to use user interface (UI). Although it is easy to use, it is expensive and doesn't come with much support tools.

There are many many other tools which can be used for web scraping that exist. The ScrapingBee website lists many tools so please reference this page for more information. Many tools are catered to a specific need such as SEO, non-programmers, JavaScript pages, visual scrapers, and more. It all depends on what is needed for the specific task at hand.

Technical Section

The technical portion of this assignment was primarily completed using Jupyter Notebooks. A Jupyter Notebook file and corresponding pdf was included with the submission of this assignment. Please see those two (2) attachments for more information regarding what was done during the technical portion of the assignment. I included two (2) screenshots below to show the MongoDB 'news' database and collections. There is a 'denver' collection for the Lab portion of this week's work (not necessary to include, but I did anyway), and a 'cdphe' collection for this week's assignment. Data was scraped from Colorado Department of Public Health and Environment.

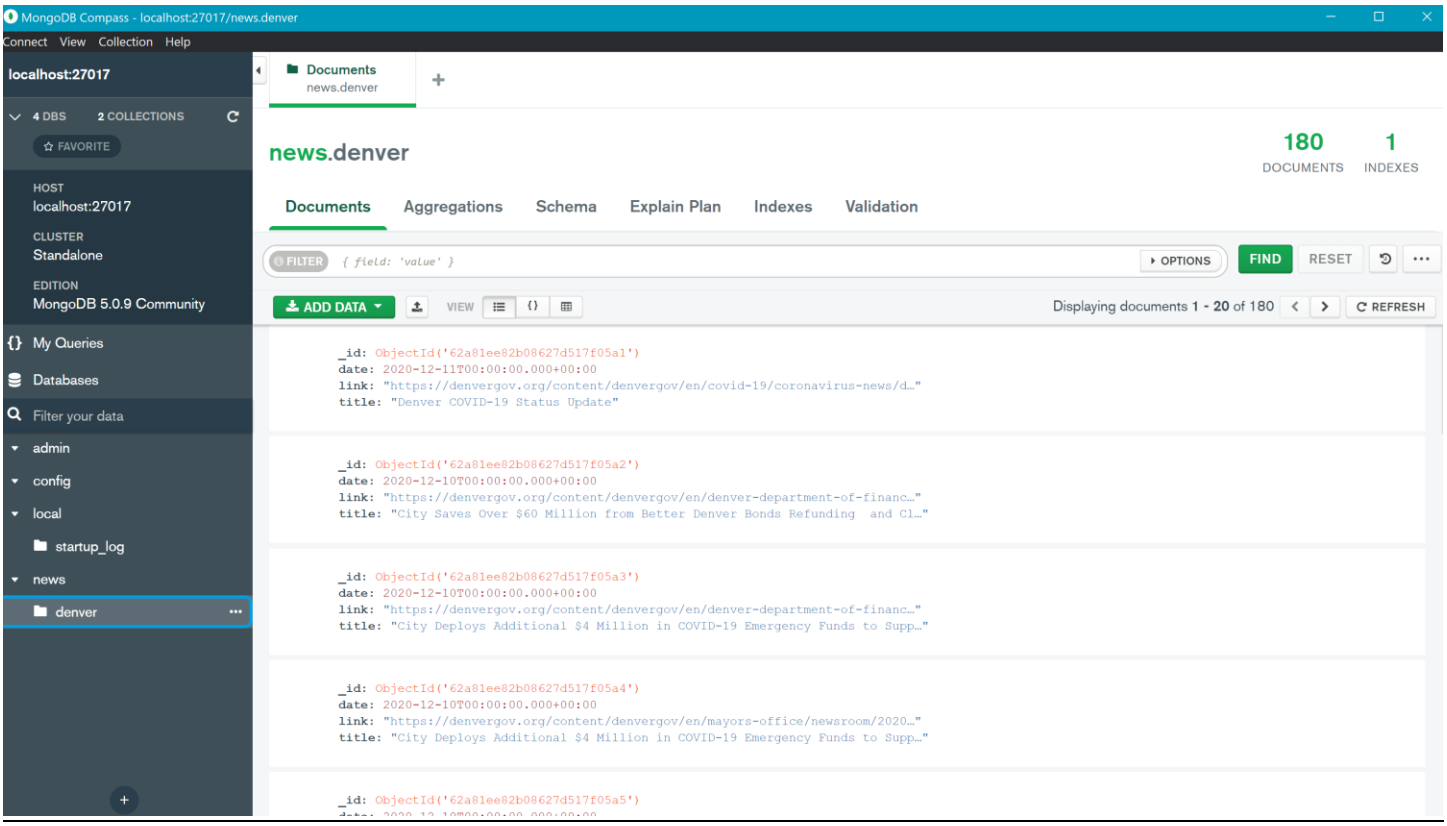

Image 1: Lab Portion of Week 6, MongoDB, Scraping from denvergov.org

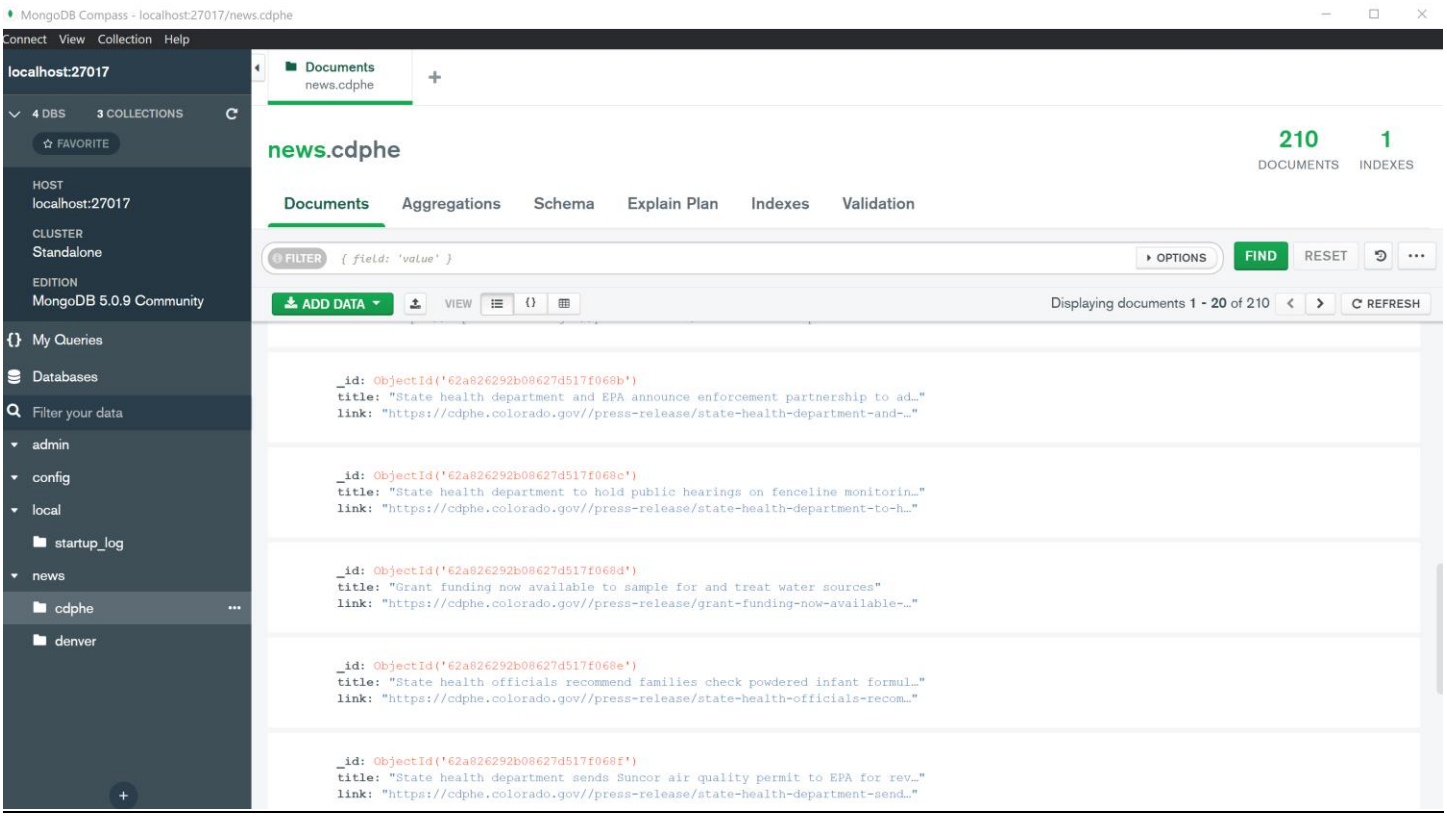

Image 2: Assignment, Collection in MongoDB, Scraped from CDPHE

Thank you!

Jeremy Beard

References

1) Roberts, E., Lynch, B., Hewitt, N., & Johnston, D. (2022, March 11). Is Web Scraping Illegal? Depends on What the Meaning of the Word Is | Imperva. Retrieved June 12, 2022, from https://www.imperva.com/blog/is-web-scraping-illegal/
2) Perez, M. (2021, August 12). Is Web Scraping Ethical? | ParseHub. Web Scraping Blog (Tips, Guides + Tutorials) | ParseHub. Retrieved June 12, 2022, from https://www.parsehub.com/blog/web-scraping-ethical/
3) Sahin, K. (2021, June 11). The Best Web Scraping Tools for 2021. ScrapingBee. Retrieved June 12, 2022, from https://www.scrapingbee.com/blog/web-scraping-tools/