Is Web Scraping Ethical?

Anthony Leonel Carvalho

Florida International University

CGS3095 Section U0X – Fall 2022 **Abstract**

In 2019, I created a web scraper that helped myself and many other students discover answers and explanations for homework topics. I was later sent with a cease-and-desist letter from a huge institution that conducts a similar company, saying that I had downloaded data that belonged to them and was previously not permitted to do so; was my activity unethical; had I violated any terms; and is web scraping unethical?

Is Web Scraping Ethical?

1. INTRODUCTION

In order to help students get unstuck on homework, a student created bots to scrape data shared online by websites and institutions, which presumably violated the copyright and privacy of the websites that once shared this information online. To understand whether the practice of better, the scrapping for this issue is ethical or not, we must first understand what web scraping is and what it is used for. Web scraping is a method of gathering data that is as old as the internet itself. It has been used by many, most notably search engines, which were designed to scrape websites and create indexes to fuel search algorithms.

2. BACKGROUND AND SIGNIFICANCE

Web scraping dates back to around the creation of the world wide web in 1989 because you can't scrape the web if there's no web to scrape from so after he invented the world wide web, Tim Berners-Lee also created the first web browser that converted html code into a nice formatted hyperlink and media rich document a couple years later in 93 Matthew Gray at MIT created the Wanderer with the intention of measuring the size of the world wide web. The Wanderer was most likely the first web robot and had the potential to become a general-purpose WWW search engine with its index.

The main character in this ethical issue is actually me; in 2019, I wrote a script to help me find flashcards and quick answers to questions online; I used selenium and python to scrape the data and Google as my primary source of data; and the script worked as follows: The user would enter an input, a question, and the script would utilize that input on Google and record the first results that appeared. Another component of the script was

responsible for using the query supplied by the computer and using Dom elements to discover elements that would match the terms from the question, thereby finding the solution to the query, repeating these steps on each link google returned would give the user a list of ways to solve that problem in many different ways found.

2.1 ACTION BEING USED FOR AN ETHICAL GOOD

This tool specifically assisted me and my friends in locating valuable information online that would assist us in studying and comprehending more about subjects in class. I named this tool homework king, which means "king of the homework," because it is similar to Google in that it returns a page with the exact location of the explanation to the user without the need for googling.

The tool ran over three thousand inquiries over a two-week period (the length of time I had the tool up and running), and I was able to make a little over two hundred dollars from the $0.99 price I charged for an account on my tool (the first five searches were free, then $0.99 a month). The "success" was unexpected since I made it while only trying to pursue a project while learning a library called Selenium and Requests. I was also studying HTML more in detail and coding my own CSS, which drove me even more to design everything on my own and devote a significant amount of time to this project. One day, I received a written message from a website called Chegg, which apparently discovered content that belonged to them on many of my recent results, and as a result, they wrote me a Cease-and-Desist letter, which, being young and not wanting to pursue this as an actual platform or business, I decided not to fight over and abandoned the idea.

With all the history being said, to discuss this ethical issue, we have to answer an important question is web scraping illegal? Well, this can be a bit of a murky area because

some websites prohibit it, while others allow it with conditions. In my situation, the websites I was scraping from were not attacked in any manner, thus I was not accessing previously forbidden information; my script would index the page, which non-logged people could read. In this method, no websites were hacked; instead, the script viewed and saved their pages.

Crawlers search the internet for matches, in this example, a match to the supplied query followed by the most likely response, with the outcome most likely being the answer to the given question. According to the ACM 1.5 privacy code of ethics, this conduct is technically immoral because the data holder, the website, and the website owner have not consented to freely release this information/material to be searched by bots. In this case, however, the crawler uncovered the steps for finding the explanation on a specific PDF file scraped from a public university database. As a result, the script was allowed access to the data because it is public. In my opinion, the script writer (me, the computer professional) who wrote the code to scrape websites from Google is ethical in some way, given that the website owners openly distributed the material (pdf and textbooks) on Google to be searched and parsed by other engines and crawlers.

3. CONCLUSION

While it can be invasive on the website owner and privacy stands of the institution who decided to publish these documents online, it is totally acceptable as it was used to improve other student's lives whom before, were stuck on questions that they were not able to find answers otherwise. In my point of view, even though web scrapping is considered grey area when it comes to privacy and even copyright infringement's (as it is basically harvesting website's data) I used my skills to create an incredible tool to help other students like himself to be more educated and get "unstuck" on questions they were once stuck and

not able to find solutions and explanations. Additionally, the action and good also follows the CFAA act (Computer Fraud & Abuse Act) where, the data was deliberately published online by the owner or institution that operates the website and data. Finally, my invention operates in the legal side of the law as it does not infringe the CFAA act as well as it offers a benefit to a group of people that benefits from the easy access of that information (solutions and explanations for academic exercises).

3.1 Ethical Actor

In this case, the computer professional who developed the web scraper or bot, the Ethical Actor is found to be me, as the person who developed the tool who presumably breached the privacy of those who alleged me of doing so.

3.2 Action being Analyzed

In this case, the act creating and utilizing bots or automated scripts to parse data from a private's organization website, which presumably would breach terms of services provided by the same. The script in this case would go over websites all over the web that somehow matched the question the user using my tool entered and return an indexed result of that particular page for better visualization of the user, in order to help him better find useful information to solve the question searched.

3.3 Ethical Standard

Standards used to test this ethical action are the following:

ACM 1.5 according to the ACM 1.5 privacy code of ethics, because the data holder, the website, and the website owner have not consented to freely allow this information/material to be searched by bots this behavior is considered to be unethical in that sense.

However, moreover following the standards given by the CFAA act (Computer Fraud & Abuse Act) which states that if the owner or institution that manages the website and data purposefully shared the data online without a required log in page it also authorized the utilization of the same for to be viewed and searched, thus parsed which was the case on my Ethical issue.

3.4 Specific Provision(s) of Ethical Standard

In order to understand the provisions of the Ethical Standard, we first must apply the two standards used in this Ethical Issue, ACM 1.5 states that the data holder must have given consent to allow bots to search the website, however according to CFAA, consent is given one data is shared online without a required log in page which in this case authorizes the bot to index that content, only being unethical or illegal when the bots are to slow systems down or hack and breach security of the third party websites.

3.5 Comparison of Actions with Specific Provisions: Ethical

To compare the actions of this ethical issue we can utilize the Wanderer program as an example, the Wanderer program was technically the first web index ever invented, which in 1993 was created to go over all pages of the world wide web to create an index of all the content of the internet, similar to Google nowadays, the creator of the Wondered program MIT student and Google employee for over 15 years, Matthew Gray was not found to violate any standard or law regarding the use of data and his bot utilization. Even thought at that time there was no laws specifically noting use of bots or crawlers utilizing the CFAA act and following the Rules of ACM more specifically ACM 1.5, we can assume that his actions were also not unethical in that sense. Therefore, a similar issue in this case presented by me, should similarly not be invasive by

any means when it comes to breaching privacy, non-authorized use of data, and hacking violation as the data was publicly shared by institutions.

3.5 Comparison of Actions with Specific Provisions: Unethical

On the other hand, one might argue that the usage of any information without prior consent might be technically unethical. According to the ACM 1.5 privacy code of ethics, this conduct is technically immoral because the data holder, the website, and the website owner have not consented to freely release this information/material to be searched by bots, thus making it technically unethical. Utilizing this approach, we can conclude that the action I took was unethical because the website that sent me a cease-and-desist letter discovered that their information was specifically found in one of my queries and given that they had not previously given me permission to share that data. Therefore, allegedly breaching their privacy and usage terms, that itself can conclude that the action and good were found to be unethical.

REFERENCES

[1] Matthew Gray develops the World Wide Web Wanderer. is this the first web search engine? Matthew Gray Develops the World Wide Web Wanderer. Is this the First Web Search Engine? : History of Information. (n.d.). Retrieved November 27, 2022, from https://www.historyofinformation.com/detail.php?id=1050

[2] What is scraping: About price & web scraping tools: Imperva. Learning Center. (2019, December 29). Retrieved November 27, 2022, from https://www.imperva.com/learn/application-security/web-scraping-attack/#:~:text=Web%20scraping%20is%20the%20process,replicate%20entire%20website%20content%20elsewhere.

[3] Scraping Robot. (2022, April 8). Web scraping history: The origins of web scraping. Scraping Robot. Retrieved November 27, 2022, from https://scrapingrobot.com/blog/web-scraping-history/#:~:text=It%20was%20developed%20by%20Jonathon,275%2C000%20entries%20spanning%201%2C500%20servers.

[4] S. Burr Eckstut | Erin Hanson. (2022, April 22). Web scraping, website terms and the CFAA: Hiq's preliminary injunction affirmed again under van buren. White & Case LLP. Retrieved November 27, 2022, from https://www.whitecase.com/insight-our-thinking/web-scraping-website-terms-and-cfaa-hiqs-preliminary-injunction-affirmed-again

[5] The code affirms an obligation of computing professionals to use their skills for the benefit of society. Code of Ethics. (n.d.). Retrieved November 27, 2022, from https://www.acm.org/code-of-ethics