Summary of Tutorial: Legality and Ethics of Web Scraping

Mike Pendleton

West Texas A@M University

Prior to explaining web scraping, the explanation of where the data is collected was important for the authors to begin the discussion. Big Web Data is the collection of data from a multitude of sources. Big Web Data included examples of HTML, databases, emails, blog posts and more. The most important aspect within the sources due to the amount of data change and updates is the veracity. In other words there is two types of data that can be collected and organized, both determinations would be either quality or the lack there of.

Web Scraping is a definitive way of extracting the data from the determined sources and then to provide it in a way determined for the purpose of the project. The processes involved are website analysis, website crawling and data organization. There can be several ways that web scraping can be done. Programmatic languages such as R and Python are the most popular for programmers. Individuals can acquire an abundant amount of other tools that can automate this for those in which programming is not a strong suit in relation to coding.

When the tools determined in order to begin a web scraping project, there are several legalities and ethics which should be considered. Regarding legalities and ethics stating “should be considered”, I did intend to state this loosely because the legality and ethical pool has a large amount of grey areas situationally. Due to the technology being considered relatively new, there is not much litigation where it would be considered strong for a concerned impacted party to be able to win either a civil or legal case.

The purpose of grouping legality and ethics into one group is due to the amount of grey area legally, this would also seem legitimate with ethics as ultimately the individual makes the choice.

Those that are aware of these topics, then could state that they did not intend to impact the party where the suspected cause has occurred. Prior to reading the article I was expecting a different result and would have thought of the topic being more structured, to these are the laws that would be infringed. “Computer Fraud and Abuse Act (CFAA)”, is the basis where several opinions have been made regarding what can be done legally.

The sections in the CFAA where an individual could be held responsible are breach of contract, Copyrighted Material, trespass to chattels, and trade secrets. Web site developers should leverage a robots.txt file, as this does provide a defined authorization on what can or cannot be used for web scraping projects. The only potential problem to robots.txt file is if the web site owner does not own the data directly, as the license for the file would not be created.

Through the use of Internet research ethical guidelines, which is what the authors followed for the paper is the direct relation to how web scraping should be processed in the intended project. This leads back to my original concerns for the legality of impacted parties as guidelines do not necessarily protect or warrant actions legally. The proposed questions that a person should ask him or herself are as follows: ( (Vlad Krotov, 2022))

* Can my actions in relation to Web data produce harm to individuals, organizations, or

communities?

* What can we do to provide reasonable assurance that this unintended harm does not happen?”

The questions alone propose a systematic problem on web scraping as there are not any truly defined laws, which would give precedence on if there was a malicious intent for the web scrapping. The only defined reasoning from my interpretation was if a loss has occurred, but my question to this is how would the individual performing web scraping know if it would negatively impact the web server? In order for an individual to know, they would have to possess intimate knowledge of the resources assigned to the system. Additionally, the amount of resources are leveraged during the base line for the reasoning per the web servers firm to allocate x amount of cpu and ram. It would be extremely difficult to prove that someone outside the organization has these types of details.

I think that answering the above questions are a good start in order to provide truth to attempting in the web scrap project legitimately. It would be better if the legality was not so grey and thus could help with a more defined structure. For example, if such a defined rule would break “ x name” law, then if included in the robots.txt file and a person would know legally and ethically what can or cannot be done. Ethics is defined as what is morally right, and I believe it would be difficult to follow an ethically constraint on if there were not any structured legalities surrounding web scraping do’s and dont’s through litigation.

# Works Cited

Vlad Krotov, L. R. (2022). *Tutorial: Legality and Ethics of Web Scraping.*