Megan Moore
Buff ID: 1014735

Legality and Ethics of Web Scraping – Homework 3

Gathering data can be quite an intensive and tedious manual process which exponentially becomes more difficult the larger of a data set that is sought. The data, and vast amount of it, available using the internet is comprised of all sorts of formats from structured to unstructured and both quantitative and qualitative data. This data comes from web pages, html tables, databases, emails, messages, blogs, photos, etc. Due to this extensiveness, a new method was needed. Web scraping became an automatic method to gather large amounts of data from websites. There are a variety of ways to perform web scraping to obtain data from websites. Most commonly, the data is unstructured and the format must be converted into structured data within a spreadsheet or database in order to be able to be used.

The web scraping process involves two parts – crawling and scraping. Analyzing a website requires a thorough review of the website structure or database repository in order to understand how the data one seeks is currently stored. To do this requires a base understanding of internet or database architecture, i.e. technical knowledge of coding languages. The crawler part of webs scraping can be best described as an artificial intelligence algorithm built to browse websites searching for specific data by following a string of links across the internet. This process involves building and running a script which automatically searches the website or database and retrieves the desired data. These types of crawler applications are usually built in R and Python language due to the overall popularity and use of these particular languages. The scraper part of the web scraping process is a tool created to extract that data from the website. In more recent availability, a large amount of “point and click” scraping tools allow automation for some steps of the scraping process without the base understanding of website and database architecture previously mentioned. Not surprisingly, these automated tools do not work as well

as they simply aren't smart enough to analyze some web pages. It is becoming more common that these tools need advanced scraping capabilities and custom programming to keep pace with the ever evolving web content technologies and user interactions. There is a wide availability of both technical papers and literature for web scraping topics. Some of these are more information while others are to be used in a technical tutorial application. All articles however have one goal – to position the argument for the need of an integrated, social-technical approach to the web scraping process that combines not only the technical aspect but also the legal and ethical ties surround this growing and emerging practice.

One might not be aware but a large aspect to consider in the web scraping process is the legal and ethical issues that can arise. The classic consideration of "just because one is capable of doing something does not mean they should" comes to mind. Having only the technical skills needed for web scraping is not enough for producing publishable research or data driven products. Unlike prior years, now researchers must determine the level of legal and ethic aspects involved before beginning a web scraping project. This is due to many legal frameworks that have arisen: illegal data access, data breaches of contract, copyrights and disclosures of trade secrets. Ethics is even more of a grey area that researchers must navigate. A research project that potentially needs to rely on data that is derived from compromising individuals privacy or violation of rights should be avoided. The potential also for research projects to lead to erroneous decisions or attribute to bias or discrimination are ethical concerns as well. Individuals and researchers involved in the web scraping process must now be mindful of the legal and ethical issues to avoid lawsuits or publicized controversies which can impact projects and reputations of both companies and individuals. While there are a number of available types of literature for reviewing the technical aspect of web scraping, there is a noticeable gap for pieces pertaining to

the softer issues that surround the data gathering process. This is due to web scraping still being a newer practice and the application of legal frameworks is inconsistent. Because of this inconsistency, it becomes obviously difficult for practitioners to determine the legality of their actions. Social acceptability is still an evolving and emerging component for web scraping that relies on data collection to answer research or industry questions.