Introduction

With the continual advancement of computer technology and the proliferation of the Internet, the amount of criminal justice-related information being placed online has dramatically increased over the last decade. As a result, public access to certain types of criminal justice data and statistical information on the Internet has rapidly expanded, presenting new and fundamentally different data access opportunities for criminal justice researchers. One method researchers are using to harness these new data access opportunities is web scraping. Web scraping is essentially an automated tool for searching and extracting data from websites and other online sources. Pioneered in the fields of data science and e-commerce, web scraping provides a user with an automated way to find and collect data of interest from online sources that is more efficient and economical than techniques traditionally used in the past, and it arguably holds great promise for researchers working in the criminal justice community (Levy, 2017).

This brief is intended to: introduce criminal justice researchers to web scraping and explain what web scraping is and how it works; provide examples of how web scraping has been used in criminal justice research; and describe several issues one should be aware of if thinking about using this type of data collection method for criminal justice research purposes.

What is Web Scraping?

Web scraping is an automated tool for finding and extracting data from online sources. It utilizes computer programming software and customized software code to mine data or other information from online sources in order to remove a copy of the data and store it in an external database for analysis. Typically, the data harvested through web scraping is analyzed to answer questions that could not be answered, or answered inefficiently, using the data as it was originally presented online. Essentially, web scraping is a way to pull information from particular web pages and re-purpose it for customized analysis (Marres & Weltevrede, 2013). Web scraping is also referred to as automated data collection, web extracting, web crawling, or web content mining. Web scraping has arguably been around since the inception of the World Wide Web, but it has primarily been utilized in the field of data science and is commonly associated with e-commerce (Marres & Weltevrede, 2013). Indeed, a form of web scraping is often used by travel-related websites readers may be familiar with, specifically those that allow consumers to compare prices for airline tickets or hotel room offers by different companies. In the past decade, however, the use of web scraping has emerged in several other fields including journalism, marketing, policy analysis, and psychology research (Baker & Yacef, 2009; Marres & Weltevrede, 2013; Youyou, Kosinski, & Stillwell, 2015).

How Does Web Scraping Work?

Web scraping involves the development and use of two customized software programs – a crawler and a scraper. The crawler systematically downloads data from the Internet; then the scraper systematically pulls the relevant information (unstructured, semi-structured, or structured) from the downloaded data, codes it, and relocates it in a database or file based on a pre-determined structure and format defined by the user. This new external database or file – populated with data originally presented online – is subsequently analyzed in ways the original online presentation of data did not support. Common software programming
languages like R and Python are typically used to write the software code for both the crawler and the scraper. Hence, software programming skills are essential for building and deploying a web scraper. The software code, however, is constructed based on specific search and data extraction criteria established by the researcher based on his/her understanding of the online data source(s) of interest and the research questions the analysis will attempt to answer. In practice, a data source theory, developed by the researcher, guides the programmer’s development of the crawler and scraper. This theory describes the researcher’s and programmer’s assumptions about the information source and its content, as well as their understanding of how the available data is maintained and how key measures are operationalized.

Web Scraping as a Criminal Justice Research Tool

The use of web scraping by criminal justice researchers is a relatively new phenomenon. In search of the literature for criminal justice-related research employing web scraping as a data collection tool, only a handful of studies were found in which web scraping was utilized. One of these studies was conducted by the Urban Institute (2017) as part of a larger exploration of how criminal background checks by employers may create barriers to employment among residents of the District of Columbia (D.C.). Background checks are utilized by potential employers, in D.C. and around the nation, to screen job applicants and to identify those with a criminal record. Having a criminal history record, however, does not necessarily mean an individual has been convicted of a crime. While a criminal history record is generated when someone is arrested, an arrest does not always result in a criminal charge; and a charge does not always result in a criminal conviction. Hence, it is possible for someone who has not been adjudicated to have engaged in criminal behavior to still have a criminal record, and this information can be, and sometimes is, used by employers to screen out job applicants, arguably unfairly limiting employment opportunities for D.C. residents with such records.

One of the key information needs in understanding the extent of this problem in DC requires determining what percentage of individuals with criminal records were and were not charged or convicted of a criminal offense. Researchers have attempted to answer this question in the past; but due to data fragmentation across law enforcement agencies and the courts, the ability to accurately answer this question for D.C. has been a challenge (Council for Court Excellence, 2011; Duane, Reimel, and Lynch, 2017). According to the Urban Institute researchers, web scraping provided a viable way to overcome some of the existing data access and analysis issues that resulted from this data fragmentation. Specifically, Urban Institute researchers used a web scraper to collect publicly available criminal history record data for Washington, D.C. residents over a 10-year period. These data were then used to estimate how many D.C. residents had a criminal record yet had not been convicted of a crime. The researchers determined that of the 68,000 D.C. residents who were flagged as having an arrest during the 10-year period examined, about half had not been convicted of a crime during that time span. This use of web scraping allowed Urban researchers to pull information off the web to produce more accurate estimates of the number of residents with criminal records who had not been convicted of a crime. This, in turn, better informed policy discussions regarding employment barriers for D.C. residents.

Another recent example of how web scraping has been used for criminal justice-related research involves the work being done by journalists from ProPublica Illinois, a non-profit news agency. In an article published in July 2017, David Eads describes ProPublica’s efforts and ultimate failure to obtain certain information on the Cook County jail population from the Cook County Sheriff’s Department through a Freedom of Information Act (FOIA) request. To overcome the data access obstacles encountered, Eads worked with computer programmers proficient in writing software code to create and deploy a web scraper for extracting publicly available data from the Cook County jail website (maintained by the Sheriff’s department), including inmate names, their date of birth, and the location of the jail in which an inmate was held. The information extracted from the website using web scraping will be utilized as one part of a larger project aimed at tracking the flow of inmates through the entire criminal justice system in Illinois.

A third example comes from a National Institute of Justice-funded study currently in progress at JRSA. The study is exploring how the characteristics of various on-line advertisements for escorts, such as those posted on CraigsList and other on-line sources, can potentially be used to identify human trafficking cases.