CHIN370 Assignment 3 Writeup

Carlo Mehegan
April 3, 2023

If data exists online that you would like to analyze, but it is only available through a webpage, webscraping allows for this data to be gathered. When deciding whether or not to scrape a source, you should consider if the data is meant to be publically available, if the data is commercial, or if using the data puts any parties involved at risk. In these cases, it may be unethical or even illegal to take this data and recompile it for personal use.

There may also be easier, more legitimate ways to access this data. Before scraping, you should find out if the website/service has an API that allows developers to easily interface with the data in an approved way. You can also contact the owners and they may be willing to send the data to you directly. Both of these methods remove the need for webscraping and makes sure that the data is approved for your use.

That being said, there are also ethical ways to webscrape. Many websites recognize that information may be scraped, and will provide their own guidelines for scraping. Many websites have a *robots.txt* file, added by convention, that establishes rules for web scrapers that want to scrape the site. This file typically contains information like how long to wait between queries and which files should not be scraped. By complying with the guidelines set out in this file, you can webscrape without risk of being blocked by the website or obtaining data that shouldn't be scraped.

Reference

Wuxia. Wikipedia, Wikimedia Foundation. Accessed April 3, 2023. https://en.wikipedia.org/wiki/Wuxia (Wikipedia page used for assignment)