Bruce Wen
CMSC 25900
Scraping and Profiling Data Brokers

Abstract

Goals

The **goal** of this project is to build a tier-structured, modular scraper that has the following features: **(1)** Easily add on new data-broker websites, and future users whose data will be scraped, **(2)** Provide an efficient method for the user to verify the data collected on them from data-brokers. Also, we conduct **preliminary studies** by scraping the data of ~20 consenting individuals to determine some preliminary findings. The intrinsic difficulty here to answer broader questions is the lack of consistency and accuracy among the different data brokers, which therefore requires users to verify their data; given the limited timeframe of the project, we can only survey a small sample of individuals and come up with preliminary findings.

Code Base

Scraper. 7 data broker websites (peekyou, thatsthem, findpeoplesearch, facebook, google, linkedin, twitter) were scraped using a variety of methods, including Twitter API, Selenium, and dynamically generated HTML/CSS class names in Facebook was dealt with. The above data broker websites were shortlisted from an initial, longer list of websites. Data broker sites that I experimented but didn't scrape included Pipl (trial application denied), zabasearch.com (blocked by distilnetworks.com). See *notes.md*, in code base, for some interesting issues I encountered while scraping.

MongoDB. A cloud database, MongoDB Atlas (512MB free storage), is used to store the scraped data and to communicate the requested information to the server for the front-end. The database also stores the user's input to verify accuracy of the data. The demo of live version is provided in the video.

Frontend Vue.JS. A front-end website is built so that users can verify the accuracy of their data, because initial preliminary analysis suggested that much of the information produced through data scraping on the data broker sites is inconsistent and thus it will be difficult to centrally verify the accuracy of the data. The UI allows users to mark each result if it represents accurate information about them and add comments if some of the information is right while others are not. Live weblink provided below. The front-end is hosted on Heroku. Initially I attempted Firebase hosting but ran into issues with MongoDB integration, so I switched to Heroku.

Privacy Considerations

Users' consent were obtained before scraping. This is done through a survey that only collects the user's first name, last name and email. The survey is attached as a pdf file. Furthermore, the stored data on mongoDB will be deleted at the conclusion of the project.

Preliminary Findings

Preliminary analysis on about 20 participants (age range 19 to ~70, adults/students not just in Chicago) was conducted. Much publicly available data on data broker sites are unreliable. Certain data on a same individual from a data broker website may be accurate while others may refer to another person of a similar name, such as a family member with very similar first and last names. Aggregation and comparison across different data broker sites can allow one to reasonably guess one's personal details (provided you already know something about that person!), such as the example on Blase discussed in the video. Also, certain relatively unique names that have fewer search results on major search engines like Google, Facebook, seem to generate more accurate results on data broker sites, such as Patrick in the video.

For future work, run the scraper on a significantly expanded user base, to generate more insights across larger age range and other demographics, and explore other data broker sites that may require paid access such as Pipl. A login system on the front-end will be useful especially when there is a large number of users surveyed. This larger user base will allow us to implement more sophisticated methods and possibly machine learning models to generate predictions of individuals' personal details from just knowing their names and scraping a list of data broker sites.

Front-End Weblink: https://blooming-dusk-68777.herokuapp.com/
3 Min Video Link: https://youtu.be/mc9i1VFWIhc
Full 13 Min Video Link: https://youtu.be/rDea9AGjjMc