Alexandra Nisenoff

Data Broker Project Proposal

Project Goal/Research Questions (subject to change):
What types of data are the most likely to be included in free public previews of the data collected by data brokers? How consistent is this data across different data brokers? For people that have used different names (nicknames, maiden names) how consistent are the search results?

~~~~~~

Plan:
So My plan is to get have a really lofty goal where only accomplishing part of it will still be pretty cool (read as “enough to count as a project”) and just have the stages of the project build on each other and see how far I get. The lofty goal would be to create an infrastructure that would be able to take in a name and create an interface that would aggregate the data and allow a person to quickly go through and label the data as being accurate or not (as well as allow for comments to capture the nuances in the ways the data can be wrong in a survey sort of way).

I figured that in the process of crawling of the data broker sites, the types of information that they provide in the free previews would become apparent. Building towards a system for a human to label the data as being accurate or not it would make sense to only show them information that shows up on multiple sites once. This process would essentially be a byproduct of figuring out how consistent the information is across different sites. This stage is probably going to end up being a bit tricky and touching on a lot of entity resolution techniques/issues. Finally finding a way of presenting it to the individual to label, recording that information, and maybe creating some sort of visualization would be the final step that I probably won’t get to do.

~~~~~~

Tools:
- I think that I will probably end up just using Selenium (Firefox and Python) to actually interact with the websites.
- Probably Python for any data analysis
- If I end up getting to do the extra bits that seem fun I guess I would probably end up with some HTML/CSS/JavaScript

~~~~~~

Timeline:
Week 5:
Project Proposal & First Meeting with Shepherd
Background Reading
Set up any other technical tools beyond selenium/python
Compile a list of Data Brokers that provide public previews of collected data
Figure out specifically what data I want to collect from the pages
Week 6:
~~~~~~

Set up code that will collect data from specific sites
Set up script that will do comparison of the data
Try to find people who are willing to let me search them with my tool

Week 7:

Second Meeting with Shepherd
Finish working on collection code
Try to make selenium code more generalizable
Do more testing of code
Maybe set up interface for human labeling of accuracy

Week 8:

Clean up code (also finish it up because I'm not fooling anyone with this timeline)
Test EVERYTHING!!!!!!!
Film video
Do writeup

Week 9:

Project Due

~~~~~~

Questions

1. How should we be testing our tools? It definitely isn't good enough to just test it out on ourselves, but just searching people without their permission doesn't seem like it would be okay either
2. How robust should our tool be to changes in the formatting of the site?
3. The instructions say that our tools should be privacy preserving, how far should we take this idea since we are kinda just building a really good stalking tool?
4. What does it mean by abstract? Are we talking like 150 word thing you would see at the beginning of a research paper or trying to fit an entire report in one page?

~~~~~~

Other

1. It would be really interesting to keep track of the trackers that are on the different websites (I guess to do this I would have to figure out BrowserUpProxy or something)
2. Something that could be really useful would be a tool to help people exercise their rights under GDPR/CCPA it seems like on these sites but this isn't really in the scope of this project
3. Could be interesting to look at if the accuracy of the data changes with the age of the data (previous residences)