Sarah Wojtowicz

Project 2 Questions

a. Throughout this project, we acted as investigators to uphold the system of accountability created by the San Francisco lawmakers: listers must register with the city's planning office and put the business license's number on Airbnb's website, Airbnb must display some effort in validating these policy numbers, and third parties can register a complaint of illegal short-term rentals with the city planning office. We used web-scraping to do the latter using several hours of our personal time. Imagine you're a software developer at either the San Francisco Planning Office (SFPO) or Airbnb.com. Describe a different system that verifies that the business license is valid for short term rentals in San Francisco and list at least two arguments you might hear at your organization (either SFPO or Airbnb.com) against adopting your system.

> If I was a software developer at the San Francisco Planning Office, another system that could verify that the business license is valid for short term rentals would include scraping all the policy numbers from Airbnb.com and matching them to the San Francisco Planning Office's database of business licenses that have been registered. This way, it would validate that not only the lister published a policy number on their Airbnb listing, but it would also validate that it is legitimate based on if it is in the San Francisco Planning Office's records. One argument I might hear from the SFPO is that they are concerned with legality issues regarding scraping Airbnb lister data and using that to form an investigation against the Airbnb lister. Another argument they might have is that if we create this new system to target illegal listers without a valid policy number, that an investigation into why they don't have a valid policy number might backfire on the SFPO. For example, an investigation might point out flaws in our system such as the lister applied for a policy and has been on the waitlist for an unreasonable amount of time etc.

b. The database we've created through web-scraping is a great data source of information for data scientists in order to answer and explore research questions. Skim through the Housing Insecurity in the US Wikipedia page and describe at least one research question that you could answer or explore using this data if you were a data scientist working with a housing activist organization to fight against housing insecurity.

> One research question that you could explore with the data found in this project is, is it cheaper to rent for short term housing, such as a few nights or weeks, or rent for long term housing? This is a relevant question because when you look at the per night rate on the different Airbnb listings, many are above $100 and even $200 per night for one bedroom. In contrast, when you look at the average 1 bedroom rent by year graph featured on the US Wikipedia page, the average cost for long term rent in California is

about $1,200 per month. This means that it would be very cost inefficient to rent short term housing through Airbnb.com.

c. As discussed in the introduction, the legality of web scraping is still uncertain in the US. Skim through the Legal Issues section of Web Scraping in the US on Wikipedia and this article about the legal issues with the Computer Fraud and Abuse Act, and describe at least one factor you believe is important to consider when discussing the legality of web scraping and why.

One factor I believe is important to consider when discussing the legality of web scraping is noting the difference between business data and personal data. For example, in the FareChase and American Airlines case mentioned on the Wikipedia page, FareChase was scraping and utilizing data that was already publicly available on the American Airlines website and it involved the business as a whole. In contrast, I believe it is different to discuss the legality of web scraping when it involves scraping user data from social media sites such as Facebook. This is because user's aren't putting their information online to be profitable like a business is, thus they shouldn't be treated the same when it comes to web scraping.

d. Scraping public data does not always lead to positive results for society. While web scraping is important for accountability and open access of information, we must also consider issues of privacy as well. Many argue that using someone's personal data without their consent (even if publicly provided) is unethical. Web scraping requires thoughtful intervention, what are two or more guidelines that must we consider when deciding to use or not to use public data?

Something to consider when deciding to use or not use public data is where that data will be stored. Having a secure and organized data storage system should be a requirement when using public data. If for some reason there is not a reliable storage system for the data, then public data should not be able to be used. In addition to storage, another requirement to guide whether public data can or cannot be used is allowing the public to know who the information will be shared, bought, or given to. This could be in the form of relaying a data release policy to the public so they know where and how their data would be shared, if shared at all.