Project 2 Questions

Sofia Gryzenia

a. 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 I would use JSON to verify that the business license is valid for short term rentals in San Francisco. I would use JSON to store all the data with the listing ID and policy numbers in a dictionary. I would then go through each of the ID’s checking if it is valid. One argument I might get against adopting my system is that it cannot be used on the actual Airbnb website because they forbid web scraping. However, if we create static html files like for this project, we will be able to use the data. But, if these static html files are created another argument may be that we could be missing newer listings with an invalid policy number.

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](https://en.wikipedia.org/wiki/Housing_insecurity_in_the_United_States) 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.

If I was a data scientist working with a hosing activist organization to fight against housing insecurity, a research question that we could explore using data from the housing insecurity Wikipedia page is: what is the percent different between the home prices by county and the average cost home cost for each state? I would explore this research question by using the data from the home prices by county map, grabbing that data, and using that data and totaling it for each state to calculate the average cost. Then, I would compare the averages for eeach state to the prices by county.

1. 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](https://en.wikipedia.org/wiki/Web_scraping#United_States) and [this article about the legal issues with the Computer Fraud and Abuse Act](https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-crime-violate-websites-terms-service), and describe at least one factor you believe is important to consider when discussing the legality of web scraping and why.

When discussing the legality of web scraping, it is important to consider the terms of use of the website you are hypothetically web scraping from. The enforceability of the terms of use regarding web scraping is unclear, in the past some cases were ruled against a company when terms of use were broken by web scraping and other cases it was exactly the opposite. So, it is critical to review these terms to save yourself from the possible repercussions of violating a websites terms of use when web scraping.

1. 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?

When we are deciding when to use or not use public data, it is important to consider a few guidelines. First, it is important to make sure there is a valid intent and purpose for using this public data. After using the data, it is important to keep the data within the group of people who are using and viewing the results of the data. If the data is spread and shared to the public, it could cause issues with people not liking their information being spread around and analyzed. So, making sure web scraping on public data has good reason and is kept controlled will prevent any issues that could result from it.