Python Final Project

Adrian Guzman afg30

November 2019

1 Introduction

I have decided to create my own python project.

2 Python in my Future

I hope to use python for web scraping and data mining. I am interested in data analysis, but I also like to code sophisticated programs. I think it's amazing that we have the ability to scrape the internet for any data we want, and do anything with said data. This opens up so many potential cool projects that can have significant impact on business, personal development/knowledge, etc.

3 Python Modules

A great library for web-scraping in python is BeautifulSoup. It allows users to send http request to receive the HTML content of the requested page extremely easily. It then provides many methods to parse the HTML and find specific tags, classes, etc.
Another library that will prove useful is pandas. This will be useful for storing the data I scrape, and prove useful in the future when cleaning the data. Matplotlib and Seaborn will also be essential in the data analysis portion of the project for plotting and structuring the cleaned data to find intersting patterns and associations. Potentially I can use scipy.stats to calculate more statistics on the data such as mutual information to find statistically significant associations.

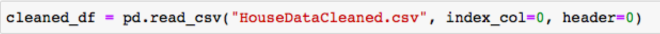
4 My Project

I will utilize all of the aforementioned modules in my final project. For the project I will scrape real estate data from a website called 'Realtor.com' for houses in the area of Austin, TX, as this is where I live. The site has over 5500 listings for houses in Austin. There are over one hundred attributes for each house, so it will be my job to comb through each house page and find and store the correct data paired with the correct attribute. There are different types

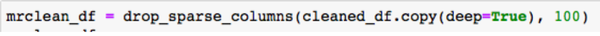
of real estate (house, condo, land, hotel) so not all attributes are consistent across all listings, so data consistency will also be a problem. Once the data is gathered, I will clean it by ensuring all data columns are appropriately filled out with correctly typed values that match the meaning of the associated attribute, and to remove all attributes that do not appear for the majority of real estate. Once the data is cleaned, I will do some statistical analysis to find interesting associations between different variables.
For future work I would like to create a statistical model that can predict the price of a house given certain attributes.

5 Explanation and Results

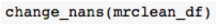
I managed to completely finish my project, although I will admit the statistical analysis could have been expanded more. The main difficulty I encountered was in scraping data. I had to limit the amount of requests made per minute, so the program takes up to a few hours to run. Also, some attribues had different identification tags across real estates, and the data types were often not the same across listings. It made for a difficult process of determining which attributes were which and then making sure the data was in the correct format. Also accounting for mising data proved to be a pain, as I had to end up putting a lot of my code into try catch statements to account for BeautifulSoup queries erring out because of a missing HTML tag or identifier.
The program is split up into 3 jupyter notebooks; one for scraping, one for cleaning, and one for analyzing. The scraping notebook is easy to use. Simply initialize all the functions, and call the crawl_realtor_com() function with the start page set to 0. This will parse through all pages of real estate in the Austin area and store attributes for each house in a CSV file. It appends to the csv file because in case the program crashes on page 67 of 105, you don't have to start over you can just start from the page you left off on by entering it as a parameter to the function. This was mainly for testing and shouldn't crash now. The cleaning notebook then reads in the unclean csv output by the scraper notebook, removes lowly populated columns, removes nans, etc. and outputs the clean data frame into a new csv using the following commands.



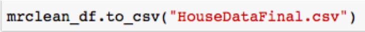
```
cleaned_df = pd.read_csv("HouseDataCleaned.csv", index_col=0, header=0)
```



```
mrclean_df = drop_sparse_columns(cleaned_df.copy(deep=True), 100)
```



```
change_nans(mrclean_df)
```



```
mrclean_df.to_csv("HouseDataFinal.csv")
```

This new csv containing the clean data is then read in by the analyzer notebook. The analyzer is kind of just a tool to mess around with the data, providing a few plots that can plot the associations between variables, the distribution of a variable, etc. It also includes the beginnings of some NLP on property

descriptions, but this has yet to be fully implemented.
As I mentioned earlier, I would like to create a statistical model that can predict the price of a house given some other attributes. Also storing the data I gathered and running the program periodically would give some interesting time series on the house prices in Austin.

I provided the clean data excel file in the submission in case you are interested and don't want to spend hours running the scraper program.