SMM750 Digital Technologies and Value Creation (PRD1 A 2022/23)

Group assignment - Deadline 04/11/2022 at 4 pm

Description of the project

You have decided to open an e-commerce shop for wine. As a savvy analyst, you don't want to just start the business blindly. Rather, you want to understand the market, and how to make a big profit. Hence, you decide to collect information on competitors and prices using your newfound scraping skills.

In your groups:

1. Sketch a plan for your research. In particular, discuss what kinds of analyses are most relevant (e.g., understanding market sizes for different wines, prices, possibilities of arbitrage from price differences across geography, …)
2. Discuss the data sources you can use to gather the information that enables your analyses and how you can systematically gather data from these sources (e.g., using APIs, scraping)
3. Prioritize your data gathering activities, considering for each data source (i) it's importance for your analyses, and (ii) the complexity of gathering the data
4. In order of priority, for some of the identified data sources, create a pipeline to (i) collect data, (ii) clean the data, and (iii) engineer relevant features. You are free to use APIs, BeautifulSoup/Requests, Selenium, or more advanced tools. However, it is essential that there be a significant web scraping component (e.g., navigating to a page and downloading a pre-made dataset will not be enough. Similar, while the use of APIs is encouraged, this should not be the only way to access data). Be reasonable in the number of sources you collect data from – start with one source, and if you have time remaining add further sources that bring a high degree of information per effort.
5. At the end of the pipeline your code should output one or more Pandas DataFrames. For these final datasets, generate lexica: Each lexicon should be a table containing the names of the features of the dataset in the first column, the descriptions of the features in the second column, and the units of the features in the third column. See *Table 1* from *chimera_exercise.pdf* (Week 5 materials) for an example.
6. Using Python, produce summary statistics and visualizations of the key features in your datasets relevant to completing your analyses.
7. Produce a companion document that describes and comments your solutions to the previous points in plain English.

Deliverables

By November 4 (4pm), you are supposed to submit the following package:

- The complete Python code used, which should be fully executable and understandable by a non-expert with coding experience (this can be a notebook, but it doesn't have to be)
- A companion document in PDF-format. The format and length are up to you, but a high added value per page is beneficial for assessment. The following contents should be included: (i) Your research plan that describes the different analyses you envision and how they contribute value to your new business. (ii) The data sources necessary to run your analysis, including a brief discussion of their prioritization. (iii) Your approach to scraping the data, the challenges you encounter, and the way you overcome these challenges. (iv) Your approach to pre-processing the data, including the reasoning behind the features you engineer. (v) The lexica of your datasets. (vi) Visualizations of the most relevant features of your data.
- A slideshow with at most five slides that summarizes the project. Some groups will be selected at random to present their work in class. You should aim to present less than five minutes.

Assessment

Your submission will be evaluated against four criteria:

- appropriate use of concepts and frameworks discussed in class
- effectiveness of the proposed answer/solution
- originality and creativity of the proposed answer/solution
- organization and clarity of submitted materials