Dotlas - Part-Time Student Data Engineer Assessment

About

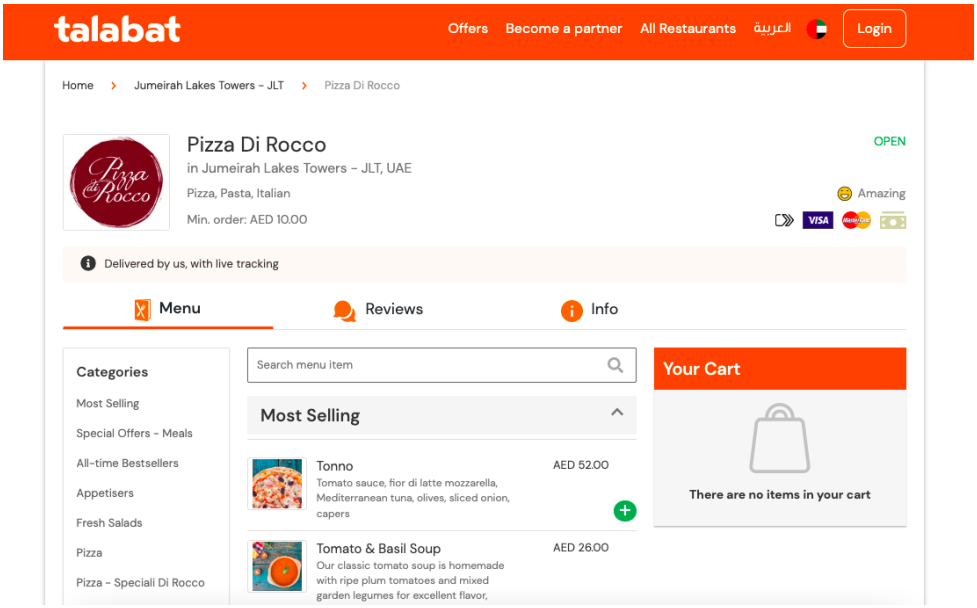
Dotlas is a data-first intelligence firm where the information we collect influences the decisions we make for our clients. In this exercise, you will scrape a public website for information as an evaluation of your programming expertise.

There is a separate document titled `Help.md` or `Help.pdf` that contains resources to learn / get started with web-scraping basics or python usage

Your Mission, Should you choose to Accept it

- This assessment involves web-scraping a small set of webpages using `python` and `BeautifulSoup`. The goal is to build a parser that can extract certain data features from the page.
- The webpages used will be from `talabat.com` website (referred to as `sample_data`), which is a food-delivery website containing listings of restaurants and their menu items.
  - A list of webpages is provided as `sample_data` within `data/sample.json`. You can use these URLs to build the parser.
  - All restaurants listed in `sample_data` are very awesome Italian restaurants btw 🍝

Fields to Scrape



Restaurant Details

Feature Name	Type	Description
`restaurant_name`	`str`	The name of the restaurant
`restaurant_logo`	`str`	The URL of the logo
`latitude`	`float`	The latitude of the location of the restaurant
`longitude`	`float`	the longitude of the location of the restaurant
`cuisine_tags`	`list`	The list of cuisine tags associated with a restaurant. Ex: `Pizza, Pasta, Italian`
`menu_items`	`list`	A list of menu items where each item in the list is given in the table below (`Menu Details`)

Menu Details

Feature Name	Type	Description
`item_name`	`str`	The name of the dish
`item_description`	`str`	Description of the dish
`item_price`	`float`	The price in AED of the dish
`item_image`	`str`	The image URL of the dish

Tooling

- Use `Python v3.10`
  - You can download it from the Python website
  - Or, you can install Anaconda and run the following command in your terminal:

```
$ conda create --name dotlas python=3.10 -y
$ conda activate dotlas
```

- Use `BeautifulSoup` library for parsing
  - `pip install bs4`
- `requests` library for fetching webpages
  - `pip install requests`

Evaluation

At the end of the exercise, you will need to share:

- **Source code**: A link to the GitHub repository where your scraper code is hosted.
- **Data:** A copy of the output (table containing restaurant details) for the 5 URLs in `sample_data` and 5 other Talabat URLs of your choosing (10 in total).

The results will be evaluated in 2 ways - qualitative and quantitative

Qualitatively

based on the following criteria:

- Readability / maintainability of code bases for the web scraper.
- Choice of tool(s) used for scraping.

Quantitatively

based on the following criteria:

- Performance of the scraper (as a function of time).
- Performance of the scraper on new Talabat URLs that are not part of the `sample_data`.
  - Make sure that your scraper solves for edge cases by trying as many talabat URLs as you can

Contact

Feel free to reach out any of the following persons at Dotlas, should you have any questions.

- Eshwaran Venkat
- Ala Mani

The solution can be iterative so feel free to ping us anytime for support. Remember to be nice to the Talabat servers, so don't overburden them with web-requests.

May the force be with you!