Pillow

Case Study – Data Science/Data Engineer

Overview

This case study is intended to test candidates on basic programming skills in python and SQL, as well as data science basics and fields of interest. Please review the instructions below and the code sample.

Deadline & Submission

Complete and return the case study within 3 days. Please return your source code and a simple write-up in pdf. (Jupyter notebook is preferred; you can use Markdown block to complete the write-up and export your results to PDF; please include both .ipynb and .pdf file in a zip file). We will review the results and run the code.
Note: You should name your zip file as Pillow_case_study_your_name_date.zip

Part 1 Keep Learning:

Data science is a relatively new field and it is changing rapidly. A good candidate should be passionate about at least one or two topics in data science and actively be learning new tools and algorithms. Answer the following questions and add it to your final write-up:

1. List 5 of your favorite website/blog/podcast/RSS/MOOC/Github related to data science/engineering
2. List 3 amazing papers you've read recently on your favorite topics
3. List the best book you've read related to data science/engineering and briefly explain why

Part 2 Exploratory Data Analysis

Pillow is dedicated to promoting short-term rentals. Therefore, understanding the current short-term market plays a vital role in Pillow's business. In this part of the case study, you will need to explore a small Airbnb data set and answer the following questions carefully:

- Use Postgresql client or psycopg2 package in Python and database information given below to connect the database
- Select 14 features (id,bedrooms, bathrooms, city, country, is_location_exact, lat,lng,price, description, picture_urls, picture_captions, star_rating, recent_review) from ***case_study_data_short_term_rentals*** table in **public** schema
- Use python (jupyter notebook prefered) to do exploratory data analysis with visualizations and answer the following questions:

    - How many listings are unique? (Filter out duplicates if any)
    - What country has the highest median price for a one bedroom?

Pillow

- Find two unique listings that have the shortest distance (euclidean distance/orthodromic distance)
- Visualize all US listings on a map
- Find out the total number of active listings (Active means you can still find it on Airbnb at the time you do these exercises. Hint: Id is a unique identifier of Airbnb, https://www.airbnb.com/rooms/13051179 will lead you to an Airbnb listing. You will get an error if the page is no longer active)

Database information

Host address:	ec2-52-53-200-58.us-west-1.compute.amazonaws.com
Port:	5432
Database:	postgres
User:	luke
Password:	luke_pillow
Schema:	public
table:	case_study_data_short_term_rentals

(Hint: full path of table should be public.case_study_data_short_term_rentals)
Please DO NOT create any table or relations in the database.

Part 3 Fields of Interest:

You can choose one of the following optional questions to complete the case study with the same dataset.

- **(Option 1) Nature Language Processing**
  - Find the most useful and meaningful tags or embeddings, or sentiment analysis to predict their rating
  - If we want to use descriptions as a feature to predict the price of one Airbnb listing, what is your approach? (Implement it if you have time)

- **(Option 2) Image processing**
  - Use one pre-trained deep learning model to identify objects in all of the images
  - If we want to use an image as a feature to calculate similarity between two listings, what is your approach? (Implement it if you have time)

- **(Option 3) Web scraping**
  - Build a crawler to crawl the calendar information for all the given listings.
  - If we want to make sure it is automatic, scalable, and not blocked by other websites, what is your approach? (Implement it if you have time)