ETL Project – San Francisco Restaurants and Parking Meters

Background

San Francisco, the [second-most densely populated](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population_density) major American city, covers an area of about 46.89 square miles, with more than 358,000 household, 70% of which owns vehicle (about 1 car per household).

There are about 4,400 restaurants in San Francisco and about 275,450 street parking spaces citywide. Finding parking spots is always a challenge during dinning hours in San Francisco.

Business Need

We perform the collection, integration and transformation of large volumes of parking meters data from San Francisco Open Data to and 1,000 restaurants’ data by calling yelp API. Our goal is to locate parking spots of each restaurant base on the restaurant’s longitude and latitude, select the 10 closest ones to each specific coordinate set and store the data into a new table.

Steps.

* Parking Meter Data

1. Data Extraction: We download 35,000 parking meters data set, randomly selected from total available 275,450 parking spaces from San Francisco Open Data.
2. Data Cleaning: We get rid of all rows with invalid data, including the ones with blank cells, error, or the ones with different format, and duplicate ones. We select columns that are relevant to our project only.



1. Data Transformation: After the data cleaning process, there are 32,865 rows and 7 columns left which later will be saved in a new parking meter data file.

* Restaurant Data

1. Data Extraction: We connect to Yelp API to obtain San Francisco restaurant data. By reading the API documentation, we understand that we are only allowed to query 50 business data per request with 20 requests limit. We extract 1,000 total restaurant data from Yelp API. 
2. Data Cleaning and transformation: We parse JSON structure data into CSV. The main columns in this CSV file are names of the restaurants, food category, coordinates and Yelp pages.



* Pair-Data

1. Data Collection: In order to collect the pairing data of the parking meters and restaurants, we import two python modules called math and heapq. The built-in function of math and heapq allow us to calculate distance between each two coordinates and push 10 closest matches in sorted order among 32,865 meters and 1,000 restaurants.

Heap is a module for implementing heaps on ordinary lists, allowing us to extract and append the 10 smallest distance values into the new list. Each restaurant has a pairing result in the format of list of ten list of tuples which contain the distance between that particular restaurant to its 10 closest parking meters, the coordinates of these 10 meters, and the meter IDs. 

1. Data Transformation:

What we need in the new pairing dataset is the restaurant’s names and their yelp IDs along with their closest meter IDs. We export the data into CSV file. 

1. Data Loading

The last step is to load the clean restaurant data, parking meter data, and pairing data into SQL database for data query applications.