1-Data Loading and Preparation

November 15, 2024

1 Data Loading and Preparation

In this section, we load the Yelp dataset, focusing on the business and review information for Asian cuisine restaurants in Canada.

1.1 Goals

- Load the business and review datasets efficiently.
- Filter the data to target mid-range rating (3-3.5 stars) for Asian cuisine restaurants.

1.2 Steps

1. **Data Import**: Load JSON data from Yelp, including business and review files.
2. **Filtering Data**: Filter the dataset to include only Canadian restaurants that offer Asian cuisine.
3. **Initial Data Exploration**: Inspect data structure, column types, and missing values to understand the dataset better.

[ ]:
```
import pandas as pd

# Load the Yelp business data
businesses = pd.read_json('yelp_academic_dataset_business.json', lines=True)

# Define Canadian provinces and Asian cuisine keywords for filtering
canadian_provinces = ['ON', 'QC', 'BC', 'AB', 'MB', 'SK', 'NS', 'NB', 'NL',␣
 ↪'PE', 'NT', 'NU', 'YT']
asian_cuisine_keywords = ['Chinese', 'Japanese', 'Korean', 'Thai',␣
 ↪'Vietnamese', 'Indian', 'Filipino', 'Malaysian', 'Asian Fusion', 'Sushi␣
 ↪Bars', 'Dim Sum', 'Ramen']

# Filter businesses located in Canada and in the Asian cuisine category
canadian_asian_businesses = businesses[
    (businesses['state'].isin(canadian_provinces)) &
    (businesses['categories'].str.contains('|'.join(asian_cuisine_keywords),␣
 ↪case=False, na=False))
]

# Further filter businesses with an overall rating between 3 and 3.5 stars
```

```
mid_range_businesses =␣
 ↪canadian_asian_businesses[(canadian_asian_businesses['stars'] >= 3) &
                                                    ␣
 ↪(canadian_asian_businesses['stars'] <= 3.5)]

# Get the list of business IDs for restaurants with an average rating of 3 to 3.
 ↪5 stars
business_ids = mid_range_businesses['business_id'].tolist()
```

[ ]:
```
# Load review data in chunks to filter for relevant business IDs
reviews = pd.read_json('yelp_academic_dataset_review.json', lines=True,␣
 ↪chunksize=100000)

# List to store filtered reviews from each chunk
filtered_reviews_list = []

# Filter reviews for the selected business IDs
for chunk in reviews:
    filtered_chunk = chunk[chunk['business_id'].isin(business_ids)]
    filtered_reviews_list.append(filtered_chunk)

# Combine all filtered review chunks into a single DataFrame
filtered_reviews = pd.concat(filtered_reviews_list, ignore_index=True)

# Show a sample of the filtered reviews to confirm the filtering worked
print("Sample of reviews for mid-range businesses:")
print(filtered_reviews[['business_id', 'stars', 'text']].head())
```

[ ]:
```
# Save the filtered reviews for future use
filtered_reviews.to_csv('filtered_mid_range_business_reviews.csv', index=False)
```