Mid-Semester Progress Report

DSA5900 – Spring 2023

Biswas Nandamuri

03/18/2023

Introduction

This practicum project aims to develop a recommendation system for Yelp Dataset using two popular methods: Collaborative-Based Filtering and Content-Based Filtering. The Yelp Dataset contains information about local businesses (such as restaurants, hotels, and bars), user information, and their reviews and ratings. The goal of the recommendation system is to provide personalized recommendations to users based on their preferences and behaviors.

The business context for this project is to improve the user experience on Yelp's platform by providing relevant recommendations to users, increasing engagement and loyalty, and ultimately driving revenue growth for the company. Yelp's success depends on its ability to provide valuable and accurate recommendations to users, which drives business traffic and enhances its visibility.

The recommendation system developed in this practicum project will contribute to Yelp's business goals by improving the relevance and quality of its recommendations to users. The recommendations' relevance and quality will be improved by leveraging Collaborative-Based Filtering and Content-Based Filtering methods to analyze user behavior and item characteristics and recommend items most likely to interest the user. By doing so, the recommendation system will enhance the user experience on Yelp's platform and provide a competitive advantage in the highly competitive online recommendation space.

Objectives

This practicum project aims to develop a recommendation system based on Yelp Dataset using Collaborative-Based Filtering and Content-Based Filtering methods.

The technical project objectives are as follows:

- Preprocess the Yelp dataset to clean and prepare it for analysis and modeling
- Implement Collaborative-Based Filtering (User-User and Item-Item) to analyze user behavior and generate recommendations
- Implement Content-Based Filtering to analyze item characteristics and generate recommendations
- Evaluate the performance of both methods using appropriate metrics

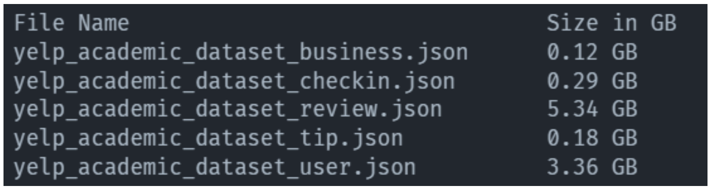
In addition to these technical objectives, the primary individual learning objectives I concentrated on are:

- To gain a deeper understanding of recommendation systems and their applications in the industry
- To develop skills in data preprocessing, data analysis, and machine learning concentrating on using PySpark, Databricks, Google Colab, and Power BI
- To gain experience working with large datasets using the principles of Resilient Distributed Dataset (RDD) using Spark, PySpark, and Databricks
- To improve the communication skills needed as a Data Professional by presenting my work and findings to the team

Data

Ingestion

The first step in building a recommendation system using Yelp Dataset is acquiring the source data. The Yelp Dataset is available for download on the Yelp website, and it contains many reviews, ratings, and other information related to businesses listed on the platform. After downloading the dataset, the data is available in five JSON files.



```
File Name                                 Size in GB
yelp_academic_dataset_business.json       0.12 GB
yelp_academic_dataset_checkin.json        0.29 GB
yelp_academic_dataset_review.json         5.34 GB
yelp_academic_dataset_tip.json            0.18 GB
yelp_academic_dataset_user.json           3.36 GB
```

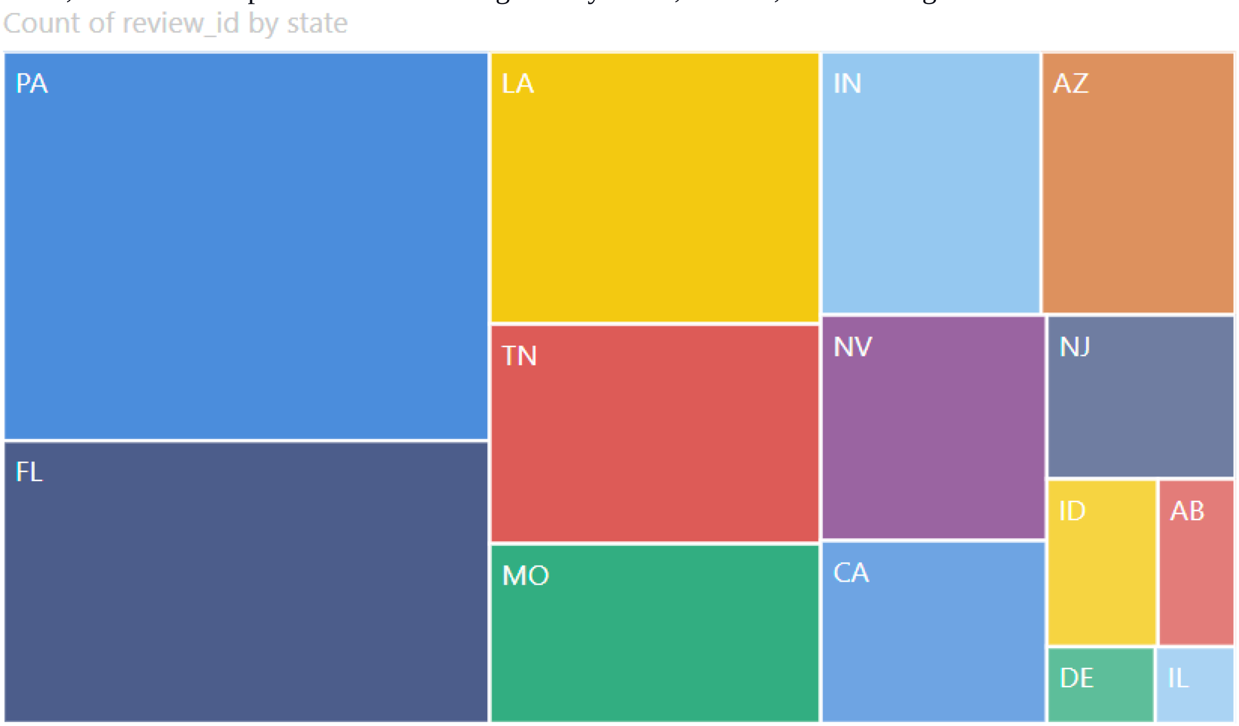
The Yelp dataset contains over 7 million reviews, 150,000 businesses, and almost 2 million user information. The three JSON files used for this project are yelp_academic_dataset_business.json, yelp_academic_dataset_review.json, and yelp_academic_dataset_user.json, summing up to the size of 9 GB of data. The data ingestion process for this volume of data was challenging. Thus, I tried using the Databricks community edition, but the community edition worked once, and subsequent logins kept showing login issues. So I had to use Google Colab to ingest and process the dataset. Another challenge is the nested JSON data available about the businesses. To handle this, I had to wrangle the nested JSON objects to convert them into single-dimensional JSON objects suitable to be stored as a CSV file. This method took a significant amount of computational resources.

Exploration

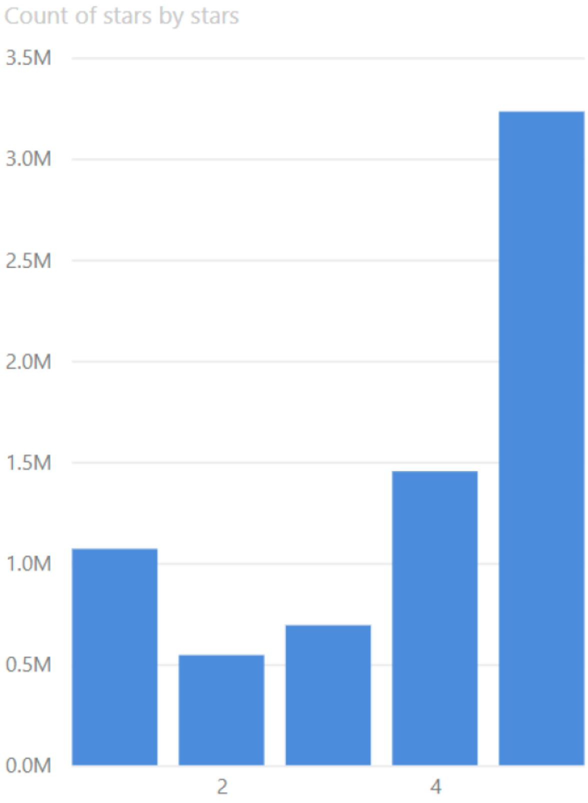
The Yelp dataset is a rich source of information that can be used to build a recommendation system. The dataset includes several types of data, including businesses, users, reviews, check-ins, and tips. However, as the above section informs, the project only uses data regarding the businesses, users, and reviews.

- **Businesses:** The dataset contains information about over 150,000 businesses. Each business has a unique identifier, a name, an address, a latitude and longitude, a phone number, a list of categories, and various attributes. The attributes can include information such as whether the business is open 24 hours, has free Wi-Fi, has outdoor seating, etc. Thus, the attribute contains both boolean-type key-value pairs and complex object types.
- **Users:** The dataset includes information about 2 million users. Each user has a unique identifier, a name, a list of friends, a review count, and an average rating. Users also contain other information, such as their location, Yelp elite status, personal interests, and compliments received by a particular User from other Users.
- **Reviews:** The dataset includes 7 million reviews written by Yelp users. Each review includes a unique identifier, a rating, a text description, and the review date. Each review is associated with a specific business and user.

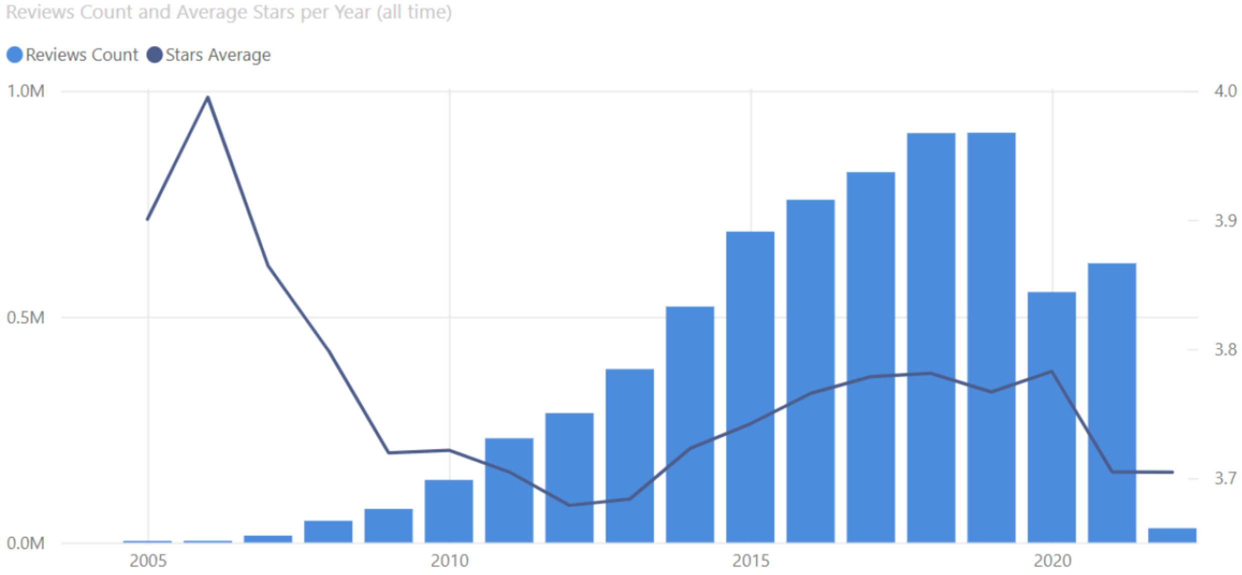
The dataset contains data from 14 cities in the United States. Most reviews are from the United States, with the most prominent cities being Pennsylvania, Florida, and Las Vegas.



The distribution of starts does not follow a normal distribution, with almost 3.4 M users giving five stars and approximately 1.1 M users giving one star. Based on the bar plot below, there is a skewness toward positive ratings, with most businesses receiving at least three stars.



Considering the number of reviews received each year, and the average rating users gave that year, the below chart shows the number of reviews received each year using the bar chart and the average stars using the line chart. The combination chart shows a positively skewed bar plot and a negatively skewed line chat. The chart shows that the users were very active in 2018 and 2019, with a significant drop in reviews in 2020. I suspect this might be due to COVID-19.



Preparation

Several data cleaning and transformation steps were used to prepare the Yelp dataset for analysis and modeling.

First, the data in the business.json file was wrangled to convert multi-dimensional JSON objects into boolean columns. Specifically, the attributes such as "BusinessParking" and "Ambience" were converted into binary indicators with the names "attr_businessparking" and "attr_ambience," which allowed me to filter and search for businesses based on their characteristics quickly.

Next, I examined the dataset for missing data and implemented imputation methods as necessary. While most of the Yelp dataset had no missing values, some businesses in the wrangled business data needed specific attributes, such as whether they had a wheelchair-accessible entrance. In these cases, I imputed the missing values with "false" since only businesses with specific attributes had those attributes in the initial multi-dimensional JSON data.

Lastly, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization method was used to convert the text reviews in the reviews.json file into a numerical format that could be used for modeling. I chose TF-IDF because it provides a way to weigh the importance of each word in a document(review), which can help to identify meaningful patterns in the data.

By performing these data preparation steps, I was able to clean and transform the Yelp dataset so that it could be used to build an effective recommendation system.

Methodology

Techniques

I intend to use content-based and collaborative filtering for this project to build a recommendation system for the Yelp dataset. Content-based filtering is a supervised approach that uses characteristics of items to recommend similar items to users, while collaborative filtering is an unsupervised approach that identifies patterns in user behavior to recommend items. In the case of the current project, the items are the businesses.

To implement content-based filtering, I will use natural language processing techniques to analyze the text reviews left by users in the reviews.json file. Specifically, I will use the TF-IDF vectorization method to create a matrix of numerical features representing each word's importance in each review. I will then use cosine similarity to find the similarity between items and recommend items most similar to those a user has already liked.

For collaborative filtering, I will use matrix factorization techniques to decompose the user-item matrix into lower-dimensional matrices representing user and item factors. I will then use these factors to predict how a user would rate an item and recommend items with the highest predicted ratings. Specifically, I will use Singular Value Decomposition (SVD) and Alternating Least Squares (ALS) algorithms to factorize the user-item matrix.

To evaluate the performance of the recommendation system, I plan to use both model accuracy and coverage metrics. Model accuracy will be measured using standard evaluation metrics such as precision, recall, and F1 score. These metrics will help me to assess the system's ability to predict user preferences and recommend relevant items accurately.

Coverage metrics, on the other hand, will be used to measure how well the system can recommend items to a diverse range of users. The coverage metrics that I will use include catalog coverage, which measures the percentage of businesses in the catalog that the system can recommend, and user coverage, which measures the rate of users in the dataset to which the system can recommend items.

Additionally, I will perform cross-validation to validate my findings and ensure that my models are not overfitting the data. I will use k-fold cross-validation to split the data into k-folds and evaluate the performance of my models on each fold. I will also perform hyperparameter tuning to optimize the performance of my models.

Process Validation

To validate my approach, I talked with my sponsor and TA, Triet Tran. Triet provided valuable insights and recommendations regarding the use of coverage metrics in addition to standard model evaluation metrics. This feedback helped me better understand the importance of recommending various items (businesses) to users and ensuring the system can provide recommendations for a wide range of users.

I chose the techniques described above, namely content-based and collaborative filtering, because they have been widely used and have shown promising results in building recommendation systems. Content-based filtering is a good approach when explicit features or attributes are associated with the recommended items, such as categories or tags, in the case of the Yelp dataset. Collaborative filtering, on the other hand, leverages the preferences and behaviors of similar users to make recommendations.

Thus, I will incorporate coverage metrics to ensure that the system can recommend items to a diverse range of users, which is essential in creating a successful recommendation system, along with model evaluation metrics like precision, recall, and F1 score.

References

1. Recommender Systems: Machine Learning Metrics and Business Metrics - neptune.ai
2. Evaluation Metrics for Recommender Systems | by Claire Longo | Towards Data Science
3. Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." *Computer* 42, no. 8 (2009): 30-37.
4. "Yelp Open Dataset." Yelp, Inc., 2023, yelp.com/dataset.