Recommendation System using Yelp data

Name:	Biswas Nandamuri
Semester:	Spring 2023
Course Number:	DSA 5900
Credit Hours:	4
Faculty Supervisor:	Triet Tran

Introduction

The project involves building a recommendation system that suggests businesses to users based on their previous ratings and reviews dataset available at yelp.com/dataset. This project is of interest and value because personalized recommendations have become crucial in today's market. People seek more relevant and tailored recommendations that match their unique tastes and preferences. The recommendation system will provide a personalized experience to users by suggesting businesses based on their past ratings and reviews.

The project addresses the problem of providing relevant and accurate recommendations to users in a crowded and competitive market. With vast options available, it is challenging to determine the most suitable businesses for a user based on their tastes and preferences. The recommendation system will help to solve this problem by suggesting businesses based on the user's past ratings and reviews.

The project will contribute to these goals by developing a recommendation system that uses the user's past ratings and reviews to generate suggestions, along with the utilization of MLFlow to build an end-to-end machine-learning process to train, test and produce the models. The effectiveness of the recommendation system will be evaluated through user testing performed using a minimal web API and user interface, along with the model accuracy and relevance analysis.

Objective

Technical Project Objectives:

1. Data Collection and Cleaning: Collect and clean the data from the business ratings and reviews dataset available at yelp.com/dataset.
2. Recommendation System Development: Build a recommendation system that suggests businesses to users based on their previous ratings and reviews.
3. End-to-end Machine-learning Process: Utilize MLFlow to build an end-to-end machine-learning process to train, test and produce the models.
4. User Testing: Evaluate the effectiveness of the recommendation system through user testing performed using a minimal web API and user interface.
5. Model Accuracy and Relevance Analysis: Analyze the model accuracy and relevance of the recommendations generated by the system.

Individual Learning Objectives:

1. Data Manipulation: Gain hands-on experience in collecting and cleaning large dataset either using Pandas or PySpark for a recommendation system.
2. Machine-learning Process: Understand the end-to-end machine-learning process and implement it using MLFlow.
3. User Interface Development: Develop skills in building user interfaces for evaluating the effectiveness of the recommendation system (FastAPI and Vue JS).

4. Problem Solving: Develop critical thinking and problem-solving skills by analyzing and interpreting the results of the recommendation system.
5. Communication Skills: Develop communication skills by presenting the results and recommendations of the project in a clear and concise manner.

Plan

Data description:

The data that will be used in this project is the business ratings and reviews dataset available at yelp.com/dataset. The size of the dataset is approximately 5GB, and the format is in JSON. The fields that the dataset contains are: business_id, user_id, stars (rating), review_id, text (review), date, etc. Based on my current plan and knowledge about the data, the fields that I will be using in the project are the business_id, user_id, stars (rating), text (involves NLP tasks like vectorization, etc.) and business_categories. The surface features of the dataset are very well documented at yelp.com/dataset/documentation/main.

Process description:

1. Data Collection and Cleaning: The data will be collected from the yelp.com/dataset and cleaned using either Pandas or PySpark. The data will be preprocessed to handle missing values, outliers, and to format the data into a usable form for the recommendation system.

2. Recommendation System Development: Collaborative Filtering and/or Matrix Factorization techniques, along with NLP will be applied to build the recommendation system. The system will suggest businesses to users based on their previous ratings and reviews.

3. End-to-end Machine-learning Process: MLFlow will be used to build an end-to-end machine-learning process. The process will include the following stages:
    1. Data preprocessing
    2. Model training
    3. Model testing and
    4. Model production

4. User Testing: The effectiveness of the recommendation system will be evaluated through user testing performed using a minimal web API and user interface. User testing will involve presenting the recommendations generated by the system to users and analyzing the results.

5. Model Accuracy and Relevance Analysis: The accuracy and relevance of the recommendations generated by the system will be analyzed. This will involve evaluating the performance of the model, such as precision, recall, F1 score, and AUC.

Deliverables

The outcome of the practicum will be the successful delivery of a recommendation system that suggests businesses to users based on their previous ratings and reviews. The recommendation system will provide relevant and accurate recommendations to users in a crowded and competitive market.

Technical Deliverables:

1. An end-to-end machine-learning process that trains, tests, and produces the models using MLFlow.
2. A minimal web application using FastAPI and Vue JS to see the recommendations generated by the model for a particular user.
3. A report on the model performance metrics like accuracy, recall, f1 score and etc.

Anticipated Visualizations:

To communicate the findings, I anticipate using visualizations, done using PowerBI, such as bar graphs, pie charts, and histograms to represent the distribution of data and the results of the user testing. Additionally, I plan to use scatter plots to represent the relationship between the features and the predicted results, and heat maps to represent the results of the model accuracy and relevance analysis. These visualizations will help in understanding the effectiveness of the recommendation system and the areas that need improvement.

Schedule

Plan:

1. Data Collection and Cleaning: February 15th
2. Recommendation System Development: February 30th
3. End-to-end Machine-learning Process: March 15th
4. API and UI development: March 25th
5. User Testing: March 30th
6. Model Accuracy and Relevance Analysis: April 6th

Project Completion Date: April 6th, 2023

References:

1. McAuley, Julian, and Jure Leskovec. "Hidden factors and hidden topics: understanding rating dimensions with review text." In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165-172. 2013.
2. Rendle, Steffen, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. "BPR: Bayesian personalized ranking from implicit feedback." *arXiv preprint arXiv:1205.2618* (2012).
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.