PROJECT: YELP RESTAURANT RECOMMENDATION ALGORITHM

Vanshika Chaddha

1. main.rs

   The main.rs file serves as the entry point for the project. It orchestrates the following tasks:

   - Reads the input datasets (reviews and business names).
   - Constructs two types of graphs:
     - Good Ratings Graph: Restaurants with ratings >= 3.5.
     - Bad Ratings Graph: Restaurants with ratings < 3.5.
   - Analyzes the graphs to:
     - Identify the top reviewers (users with the most reviews).
     - Identify the top restaurants (restaurants with the most reviews).
     - Generates restaurant recommendations for a target user using cosine similarity scores.

   Key Functions in main.rs:

   - **good_make_graph and bad_make_graph:** Generate graphs based on good and bad ratings.
   - **top_nodes:** Extracts the top n nodes (users or businesses) based on degree.
   - **create_adjacency_matrix:** Creates an adjacency matrix where each row represents a user and columns represent businesses.
   - **restaurant_recommender:** Computes recommendations based on user similarity.

2. **make_graphs.rs**

   This file implements the core graph-building and analysis logic. It processes the reviews to construct user-business relationships and provides tools for graph analysis.

   Graph Representation

   Nodes: Users and Businesses.

   Edges: Connections between users and businesses based on ratings.

   Key Functions:

   - **read_reviews:** Reads a CSV file containing user reviews. Extracts user IDs, business IDs, and star ratings.
   - **good_make_graph and bad_make_graph:** Construct graphs by filtering reviews based on the star rating threshold (3.5). Users and businesses are stored in a graph structure using HashSet and HashMap.
   - **top_reviewers and top_restaurants:** Calculate the degree of nodes (users or businesses) by counting connections.

- **create_adjacency_matrix:** Generates an adjacency matrix to represent user-business relationships. Each entry indicates whether a user has interacted with a business.
- **Cosine Similarity Functions:**
    - **precompute_norms:** Computes the vector norms for each user.
    - **cosine_similarity:** Calculates similarity between two users based on their adjacency vectors.
    - **calculate_user_similarity:** Precomputes similarity scores between pairs of users who have reviewed the same businesses.
    - **restaurant_recommender:** Uses cosine similarity scores to recommend businesses to a target user. Businesses already reviewed by the user are excluded.

3. **tests.rs**

   This file includes unit tests to verify the correctness of the core functions. The tests cover:
   - Graph construction for good and bad ratings.
   - Calculating top reviewers and top restaurants.
   - Generating adjacency matrices.
   - Validating the restaurant recommendation logic.

   Example Test:
   - In test_good_make_graph, the function verifies that only users and businesses with ratings >= 3.5 are included in the graph.

4. **data_preprocessing.ipynb**

   The Jupyter Notebook handles the data preprocessing tasks. Specifically:
   - Filters raw datasets to extract only relevant columns (user ID, business ID, and star ratings).
   - Cleans and prepares the datasets (filtered_reviews.csv and filtered_restaurants.csv) for input into the Rust program.
   - Outputs clean CSV files that can be directly used by main.rs.
   - I am from Toronto so I only used data from Ontario!

How Restaurant Recommendations Are Calculated

1. **Graph Construction**

   The program constructs a bipartite graph:

   Nodes: Users and businesses.

   Edges: Connections between users and businesses based on reviews.

   Two graphs are created:

   Good Ratings Graph: Reviews with ratings >= 3.5.

Bad Ratings Graph: Reviews with ratings < 3.5.

2. **Adjacency Matrix**
   An adjacency matrix is created where:
   - Rows represent users.
   - Columns represent businesses.
   - Entries are 1 if a user has reviewed a business, otherwise 0.

3. **Cosine Similarity**
   Cosine similarity measures the similarity between two users' review histories. High similarity means users have reviewed similar businesses.

4. **Recommendation Logic**
   To rank businesses for recommendation, the system begins by identifying users similar to the target user through cosine similarity, calculated using their review patterns in the adjacency matrix. Each similarity score reflects how closely aligned the target user's preferences are with others, forming a similarity score map. For each similar user, the system filters out businesses already reviewed by the target user, ensuring that recommendations are novel. It then assigns a cumulative score to each business, summing the similarity scores of all users who reviewed it. Businesses are ranked in descending order based on these cumulative scores, creating a sorted list of recommendations. Finally, the system maps the ranked business indices back to their names, producing a user-friendly output of top recommendations tailored to the target user's preferences. This approach ensures personalized and dynamic recommendations that evolve as new data is added.

Degree Distribution of Nodes

The degree of a node represents the number of connections (edges) it has in the graph.

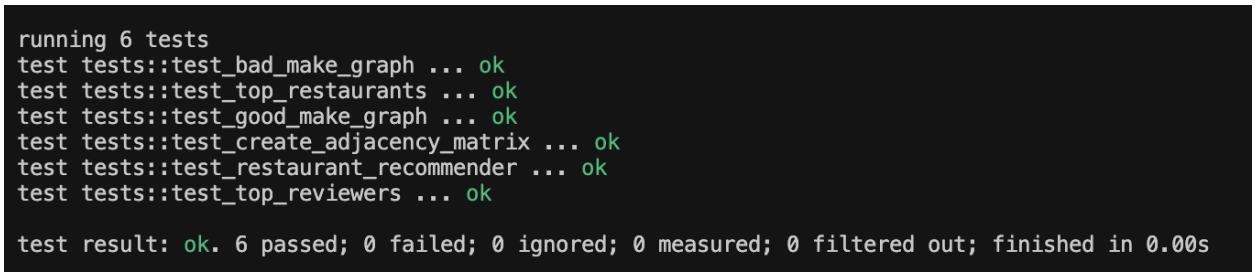
- **Top Reviewers:** Calculated using top_reviewers, which counts the number of businesses reviewed by each user.
- **Top Restaurants:** Calculated using top_restaurants, which counts how many users have reviewed each business.

Outputs

```
Top 20 Restaurants for Bad Ratings:
Restaurant: zgQHtqX0gqMw1nlBZl2VnQ, Reviews: 544
Restaurant: RwRNR4z3kY-4OsFqigY5sw, Reviews: 458
Restaurant: N_2yEZ41g9zDW_gWArFiHw, Reviews: 400
Restaurant: a8pmtlVKf7NiSLI-4KejIw, Reviews: 396
Restaurant: -av1lZI1JDY_RZN2eTMnWg, Reviews: 396
Restaurant: aLcFhMe6DDJ430zelCpd2A, Reviews: 382
Restaurant: QGTqGNLZbBA1QD8L_fO9ZA, Reviews: 334
Restaurant: O1TvPrgkK2bUo5O5aSZ7lw, Reviews: 305
Restaurant: 73_UT7fZ7mzXcguX8-oSuQ, Reviews: 299
Restaurant: mG71gQ5HWl7ut9OCN6NJKA, Reviews: 293
Restaurant: r_BrIgzYcwo1NAuG9dLbpg, Reviews: 292
Restaurant: RtUvSWO_UZ8V3Wpj0n077w, Reviews: 285
Restaurant: L3R7y3_tUKQlM9kor4-Ccw, Reviews: 270
Restaurant: BUcTdN-rNE8urCCQuxSOQA, Reviews: 257
Restaurant: nBl_4gw5ecGzNkHyzfii8g, Reviews: 256
Restaurant: htVvtLIFftBLqzRISjReDw, Reviews: 252
Restaurant: DE89UdHFMCN6DtYWZuer5A, Reviews: 233
Restaurant: Cp3YRVZojrCGeQS41Hf1pw, Reviews: 229
Restaurant: 8I5U8OQ06nSxX2y4PPOWzQ, Reviews: 224
Restaurant: MS-hfug4QDXqb_Mws3qlzA, Reviews: 219
```

```
Top 20 Restaurants for Good Ratings:
Restaurant: r_BrIgzYcwo1NAuG9dLbpg, Reviews: 1885
Restaurant: RtUvSWO_UZ8V3Wpj0n077w, Reviews: 1140
Restaurant: aLcFhMe6DDJ430zelCpd2A, Reviews: 1085
Restaurant: iGEvDk6hsizigmXhDKs2Vg, Reviews: 1028
Restaurant: N93EYZy9R0sdlEvubu94ig, Reviews: 927
Restaurant: uAAWlLdsoUf872F1FKiX1A, Reviews: 752
Restaurant: e41TP5cXZqSrz50xCBJqZw, Reviews: 665
Restaurant: B70iTJjcPkuYn8ouUewWgw, Reviews: 654
Restaurant: f5O7v_X_jCg2itqacRfxhg, Reviews: 642
Restaurant: Yl2TN9c23ZGLUBSD9ks5Uw, Reviews: 627
Restaurant: mZRKH9ngRY92bI_irrHq6w, Reviews: 581
Restaurant: h_4dPV9M9aYaBliH1Eoeeg, Reviews: 579
Restaurant: k6zmSLmYAquCpJGKNnTgSQ, Reviews: 561
Restaurant: DE89UdHFMCN6DtYWZuer5A, Reviews: 531
Restaurant: kOFDVcnj-8fd3doIpCQ06A, Reviews: 525
Restaurant: _xAJZOKBMPOe47p1MphB2w, Reviews: 518
Restaurant: RwRNR4z3kY-4OsFqigY5sw, Reviews: 515
Restaurant: -av1lZI1JDY_RZN2eTMnWg, Reviews: 507
Restaurant: 0a2O150ytxrDjDzXNfRWkA, Reviews: 501
Restaurant: JMiaNitMzMbJm6Kh0RbT5A, Reviews: 480
```

```
Top 20 Reviewers for Good Ratings:
User: CxDOIDnH8gp9KXzpBHJYXw, Reviews: 1059
User: YBT3EKUNN4IP8m4x7sGu1g, Reviews: 515
User: WeVkkF5L39888IPPlRhNpg, Reviews: 511
User: Q9mA60HnY87C1TW5kjAZ6Q, Reviews: 505
User: U5YQX_vMl_xQy8EQDqlNQQ, Reviews: 397
User: JrXC_MDp38BWwLn2SFdNsA, Reviews: 362
User: FREeRQtjdJU83AFtdETBBw, Reviews: 351
User: 2e5V6M4GNufEnbGJpVdCjw, Reviews: 339
User: ic-tyi1jElL_umxZVh8KNA, Reviews: 333
User: 0BBUmH7Krcax1RZgbH4fSA, Reviews: 328
User: O3pSxv1SyHpY4qi4Q16KzA, Reviews: 315
User: orh0HRUNCWuQMt9Iia_osg, Reviews: 308
User: pMefTWo6gMdx8WhYSA2u3w, Reviews: 304
User: JnPIjvC0cmooNDfsa9BmXg, Reviews: 297
User: Nq6e5N8bjgD9B46O4va_zA, Reviews: 296
User: yT_QCcnq-QGipWWuzIpvtw, Reviews: 288
User: V4TPbscN8JsFbEFiwOVBKw, Reviews: 275
User: EiP1OFgs-XGcKZux0OKWIA, Reviews: 271
User: PGeiszoVusiv0wTHVdWklA, Reviews: 264
User: j6wLUT0ZXi-x0otelYIFpA, Reviews: 259
```



Top 20 Restaurants for Good Ratings
The program identifies and displays the top 20 restaurants that received good ratings, where good ratings are defined as star ratings >= 3.5. These businesses are ranked based on the number of reviews they received.

Top 20 Restaurants for Bad Ratings
The program also identifies the top 20 restaurants that received bad ratings, where bad ratings are defined as star ratings < 3.5. These businesses are ranked by the number of negative reviews.

Top 20 Reviewers
The program identifies and ranks the top 20 reviewers based on the number of businesses they have reviewed. Reviewers (users) are ranked by their degree in the graph, which represents the number of connections (reviews) they have to businesses.

Recommended Restaurants for a User
The program generates personalized restaurant recommendations for a target user. The target user's identifier (e.g., gvXtMj3XuPr0xHjgmlmtng) is manually provided as input. The system calculates these recommendations using cosine similarity to find users with similar preferences.
You will have to manually change the user's identification number to get other outputs.

Implications

1. **Personalized Recommendations**
    - The project delivers personalized restaurant recommendations tailored to individual user preferences using cosine similarity and user-business interaction graphs.
    - Users benefit from discovering new restaurants that align with their tastes, enhancing user satisfaction and engagement.
2. **Improved Customer Decision-Making**
    - By identifying highly-rated and poorly-rated restaurants, the project empowers users to make informed decisions about where to dine.
    - Highlighting top businesses helps users quickly identify quality options, while showcasing poorly rated businesses alerts them to potential issues.
    - Top reviewers and their behaviors can be leveraged to identify influential users in the platform. These users can act as ambassadors for businesses or help identify trends in preferences.

This project has far-reaching implications in improving recommendation systems, understanding user behavior, and leveraging graph-based models for personalized suggestions. By analyzing both positive and negative user preferences, it not only enhances decision-making for consumers but also provides actionable insights for businesses to improve their offerings. The flexibility to extend cosine similarity for user matching further broadens its utility in social platforms and community-building applications.