Restaurant Recommendation System

Vaidehi Parikh
parikh.v@northeastern.edu

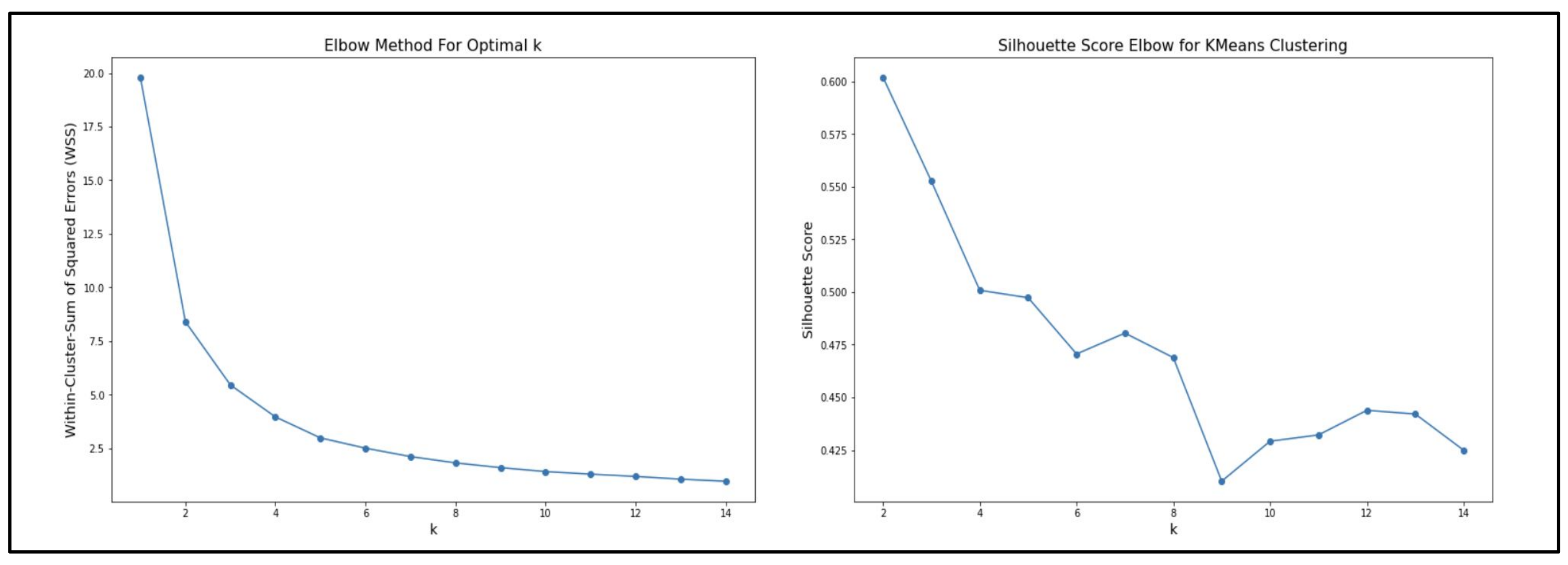
Shaival Shah
shah.shaivals@northeastern.edu

Problem Definition

- Yelp has a vast amount of data available that has a potential of generating high revenues. The highest number of ratings and reviews given by the user are in the "**Restaurant**" business. The project is an Application project with the objective of yielding recommendations to the user, based on the ratings and reviews provided by the user.
- If the user finds it accurate, it would increase the Yelp site usage. They would rate new restaurants leading to a positive feedback loop generating revenues, user information and more insights.

Existing Methods

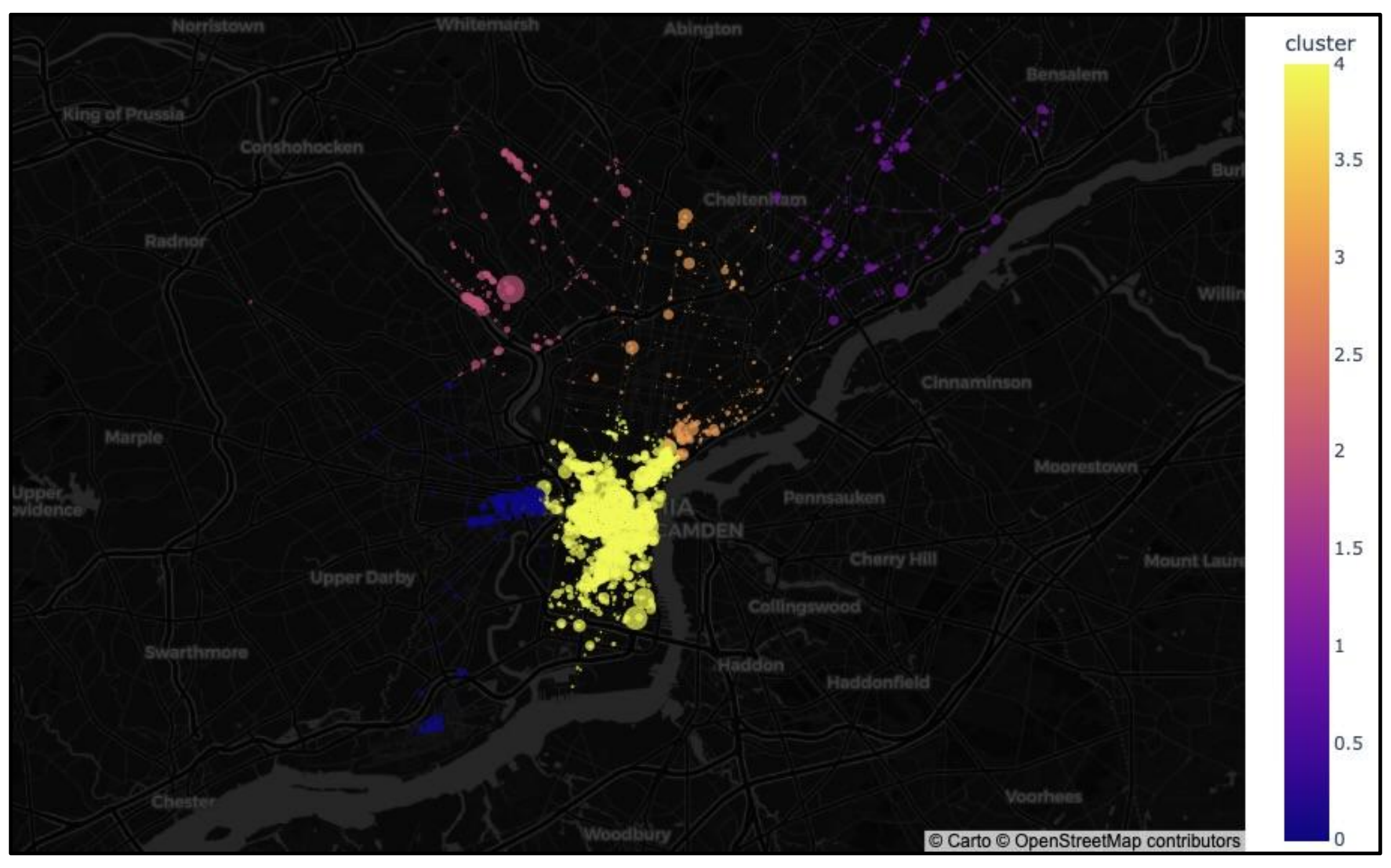
- Collaborative Filtering[Goldberg, David, et al. 1992]
  - User-Based Collaborative Filtering
    - Recommendations provided based on similar user preferences.
  - Item-Based Collaborative Filtering
    - Recommendations provided by calculating similarity of items based on User ratings.
- Content Based Filtering[Prem Melville, et al. 2002]
  - Recommendations provided based on similarity of items after extracting item features.
- Hybrid Approach [Robin Burke 2002]
  - Combination of more than one approaches.



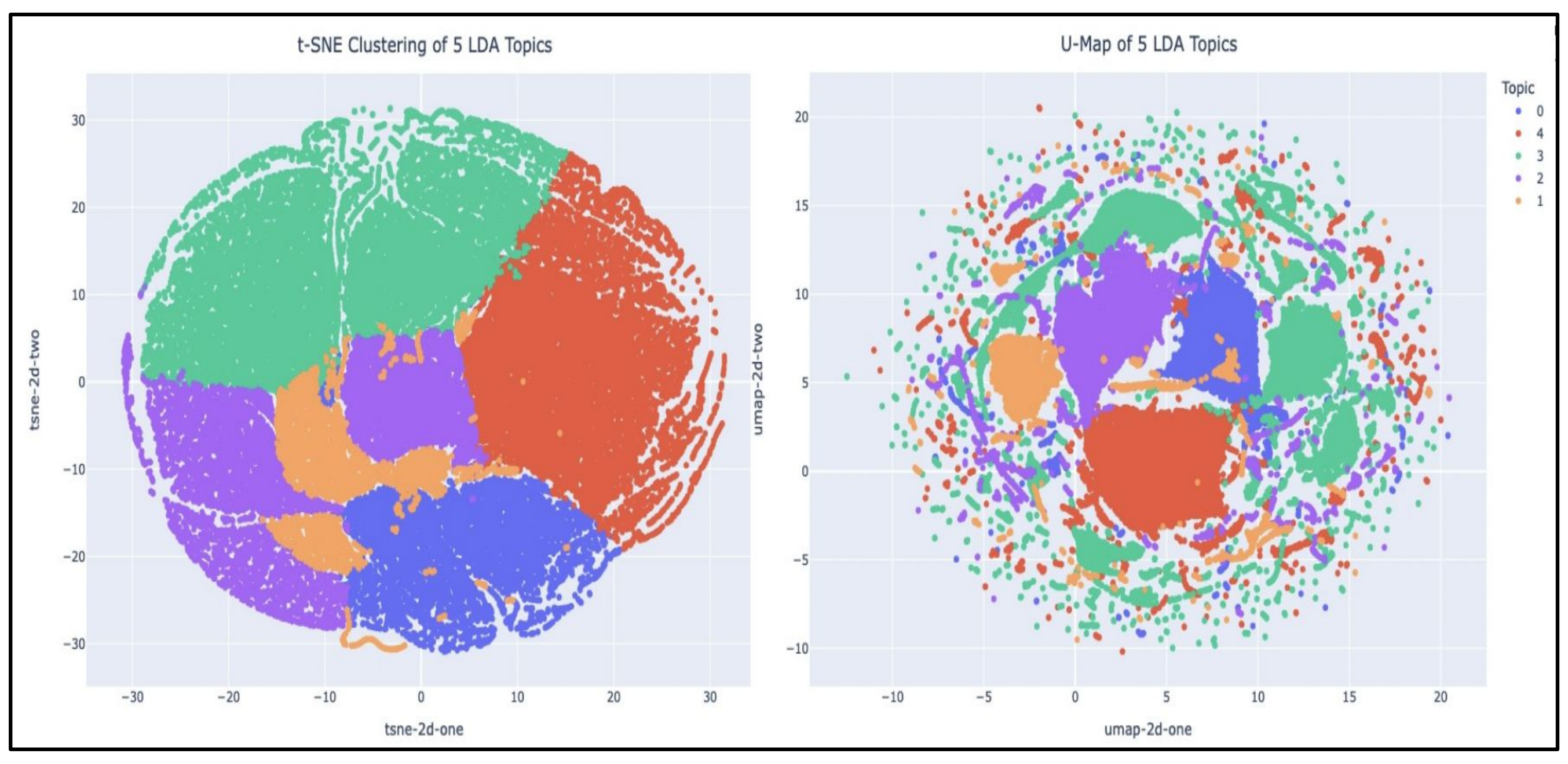
Elbow method to choose optimal k and their respective Silhouette scores.

Proposed Method

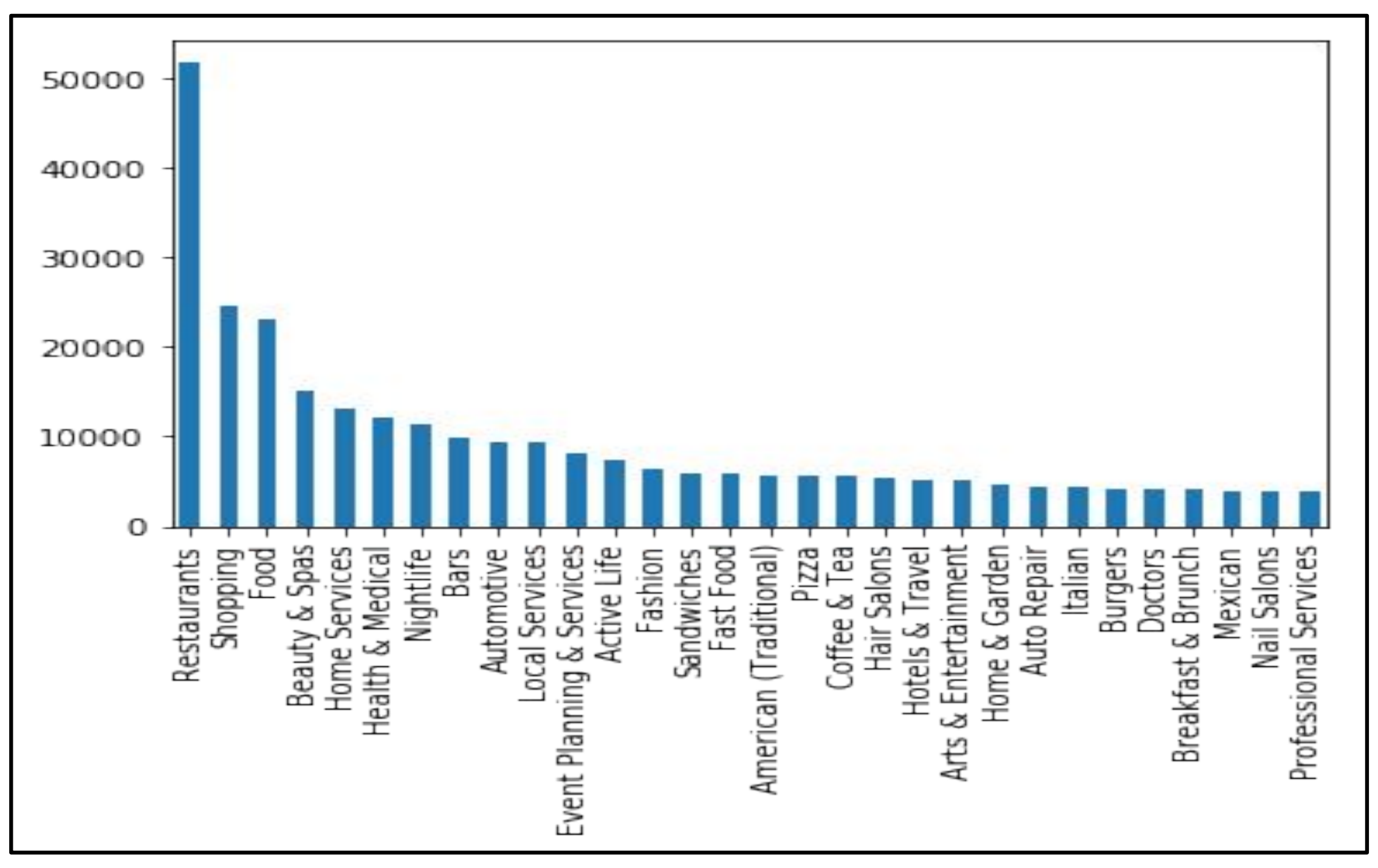
- Location Based Filtering (To handle **cold-start** problem)
  - Provides Top k recommendations based on user's location.
    - **K-means** Clustering is used to create clusters based on latitude and longitude.
    - Selection of K is based on **Elbow-method** and **Silhouette scores**.
    - **Assumption: Always recommend highly rated restaurants to a new user.**
- Item-Item based Collaborative Filtering
  - Top k recommendations provided by calculating similarity of restaurants based on the ratings given by the user.
    - Performed sentiment analysis using Textblob and Vader to calculate super scores i.e. **Super Score = User Ratings + (Textblob Score x Vader Score)**
    - **Normalized** the **ratings** in the **user-item matrix** by subtracting mean ratings for each restaurant.
    - Performed **Matrix Factorization** using **truncated SVD** to retrieve latent features.
    - Created item-item similarity matrix using **cosine similarity**.
- Content Based Filtering
  - Top k recommendations provided based on similar restaurant categories and dominant keywords.
    - Implemented **LDA (Latent Dirichlet Allocation)** for topic modeling and extracted five different topics and their most dominant 10 keywords.
    - Created a bag of words for each restaurant and **count-vectorizer** to convert the text into the vector of token counts.
    - Created item-item similarity matrix using cosine similarity.
- Hybrid Approach
  - Combined collaborative and content-based filtering methods to make a robust model.
  - Recommend top k restaurants based on **weighted average** (60% to content-based and 40% to collaborative filtering).



Restaurants clusters received after performing k-means using latitude and longitude



t-SNE Clustering and UMAP for 5 LDA Topics



Distribution of Business in the dataset

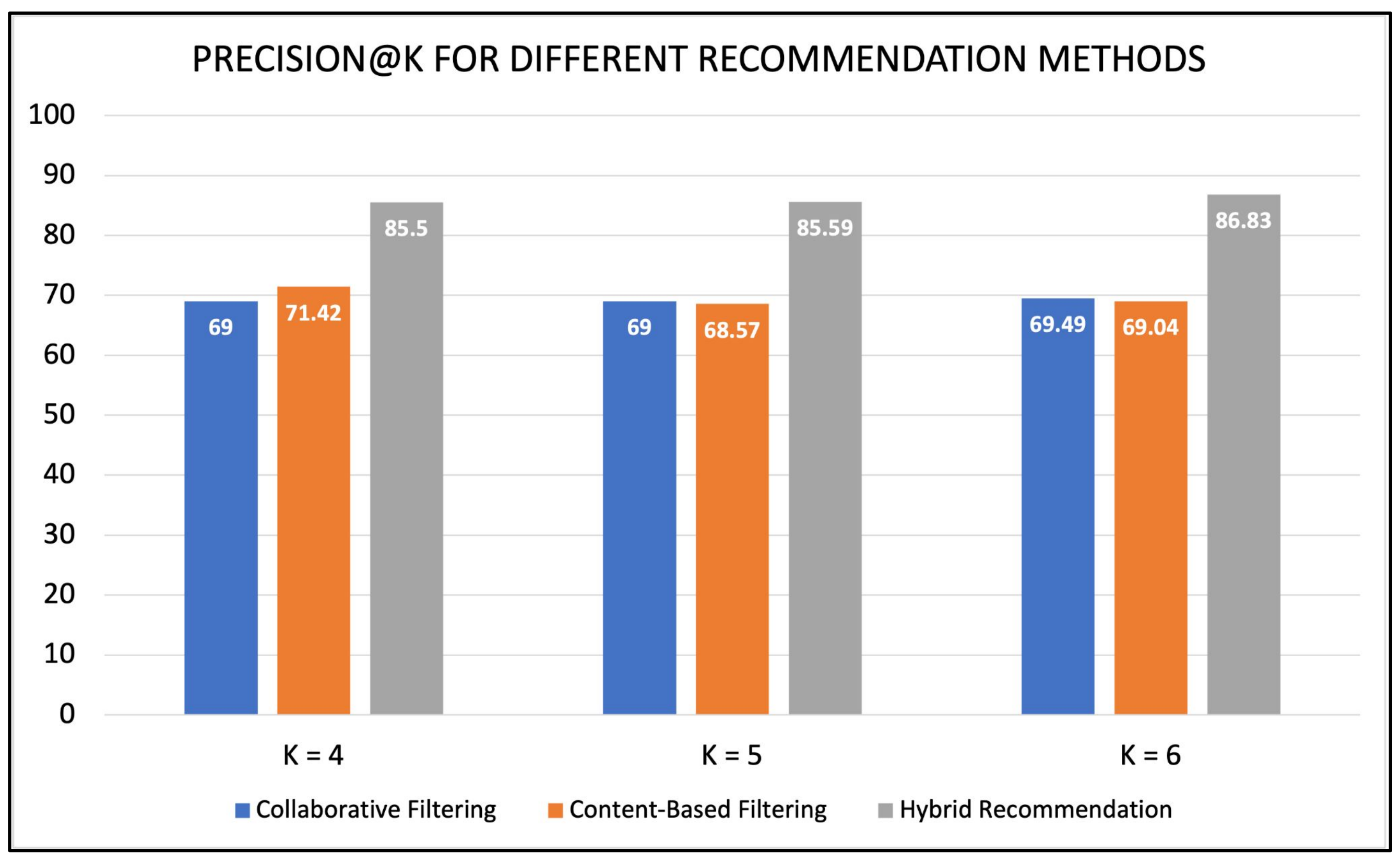
Data Description & Experimental Setup

Yelp Dataset contains 4 json Files.

- **Business**: **209K** local businesses with features: user_id, business_id, stars, country, state, pincode etc.
- **Review**: **8M** reviews with features: review_id, stars, review etc.
- **User**: user details with features: user_id, review_count, name, useful_votes etc.
- **Check-in**: check-in counts

Recommendation Experimental setup:

- Selected "Restaurant" out of all the businesses.
- Implemented recommendation system for **Philadelphia** out of all metropolitan areas.
- Report **precision@k** for the recommendation methods.
- Higher the precision@k : Better the Recommendation
- **Ground Truth (Collaborative): If the recommended restaurant has a Yelp Rating greater than or equal to 4, then it is a valid recommendation.**
- **Ground Truth (Content-Based): Calculate precision based on the similarity of restaurant category. Similar Restaurant Categories -> Higher Precision**



Precision@K for Different Recommendation Methods

Results

Recommendation Methods	Precision@4	Precision@5	Precision@6
Collaborative Filtering	69	69	69.49
Content Based Filtering	71.42	68.57	69.04
Hybrid Recommendation	85.5	85.59	86.83
Location Based Recommendation	100	100	100

Discussion of Results

- When compared to traditional approaches, the **Hybrid Approach** performed consistently better.
- The **content-based** recommendation method performed a great job at extracting item characteristics. Almost 70% of the recommended restaurants had similar keywords.
- **Collaborative filtering** performed well in recommending nearly 70% of restaurants with Yelp ratings of 4 or above.
- Including **Demographics** in the recommendation is a wonderful way to cope with the cold start issue.
- It always suggests highly rated restaurants that are close to the user's current location. As a result, a precision@k of 1 was obtained.

Takeaway Points & Future Work

Use a **hybrid approach** to make the model robust against the flaws of collaborative and content based filtering.

Future Work:

- Incorporate **Graph Theory** for location-based systems to optimize traveling routes.
- Try and implement prediction of ratings for recommended restaurants and make use of k-fold cross validation to evaluate using **RMSE**.
- Incorporate **Deep Learning and Neural Network** architectures for collaborative filtering
- Try **Bi-Grams** and **Tri-Grams** for sentiment analysis, and **pretrained BERT** weights for topic modeling.
- Deploy the work on a website and create a user interface.