Yelp Business Data Analysis and Recommender System

Omkar Pradhan, Yashvi Bhandari
pradhan.o@northeastern.edu , bhandari.ya@northeastern.edu

Problem Definition

- To summarize user generated restaurant reviews into broad topics to help business owners make improvements in their service.
- To compare the performances of LDA and LSA algorithms
- Develop recommender systems to recommend relevant restaurants to users based on their previous activity
- Demonstrate effectiveness of deep learning-based collaborative filtering compared to traditional matrix factorization-based collaborative filtering

Existing Methods

- Topic Modeling via Latent Dirichlet Allocation and Latent Semantic Analysis
- Recommendation via Content-Based and Collaborative Filtering Based (Matrix Factorization) Approaches

Proposed Method

- Topic Coherence score measures degree of semantic similarity between high scoring words in a single topic
- The number associated with the highest coherence score is chosen to implement the LDA, and LSA algorithms
- Content based filtering: Dataset already had features corresponding to each restaurant. We exploited this to form an item-feature matrix. We used KNN to recommend similar restaurants.
- Item-Item collaborative filtering: Used SVD based Collaborative filtering. Items were mapped to latent factor space to get item-latent factor matrix.
- Neural Collaborative filtering: NCF replaces user-item inner product in matrix factorization with a neural network. Thus, it can be thought of as a nonlinear generalization of traditional matrix factorization.

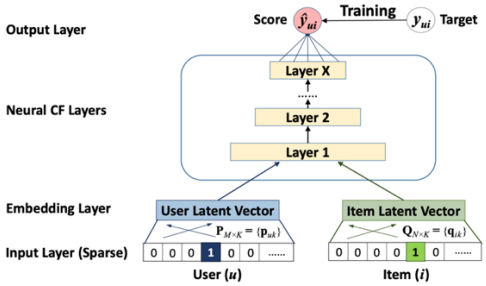

Fig. 1 – Neural Collaborative Filtering Framework

Data Description & Experimental Setup

- All-purpose learning dataset provided by Yelp Inc. consists of businesses, reviews, and user json files
- Boston was among the top 5 cities with the highest number of reviews, hence our choice for the project
- Topic Modeling performed for restaurant "Pok Pok" based on topic coherence score
- Calculated RMSE for our Item-Item collaborative filtering and Neural Collaborative filtering models

Results

- Highest coherence score for the restaurant "Pok Pok" was associated with number of topics = 7, for greater values we observed a major drop in coherence score, hence the choice for our topic modeling

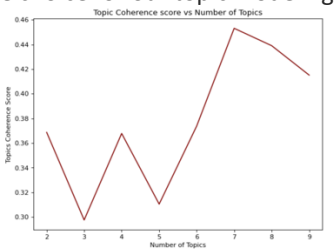

Fig 2. Topic coherence score vs Number of topics for 'Pok Pok' restaurant

Topic Number	Top Words
1	Waitstaff, story, hearing, pokpok, similar, rude, street
2	Pok, seated, hour, america, wit, ordered, restaurant
3	Pok, food, portland, place , better, service, great
4	Food, pok, wing, place, thai, good, like
5	Food, wait, wing, good, place, chicken, dish
6	Food, pok, reservation, table, family, thai, wait
7	Cash, america, place, food, star, good, pok

Table 1: Prominent words corresponding to each topic of LDA for 'Pok Pok' restaurant's reviews.

- As observed in table major concerns with "Pok Pok" include rude staff, long wait time, food quality, etc.
- The RMSE significantly decreased for Neural Collaborative filtering in comparison to Item-Item collaborative filtering

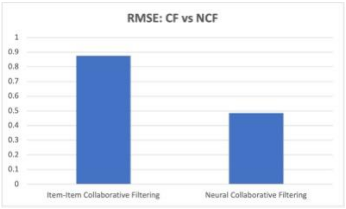

Fig 3: RMSE of CF vs NCF

Discussion of Results

- **Effectiveness**: On comparing the results of LSA and LDA we found that LDA performed better for some data while LSA for another
- We discovered that Topic Modeling depends on the type of input given by the user
- **Better Performance:** Neural Collaborative filtering performed exceedingly well in comparison to matrix factorization with much lower value of RMSE

Takeaway Points & Future Work

- In future, we would like to explore Probabilistic Latent Semantic Analysis and Parallel Latent Dirichlet Allocation
- Recommender System for Restaurant is created for 1 City i.e., Boston There is total 836 cities in the dataset. Hence project is expandable for future explorations
- In future, we would like to explore Graph Learning-Based Recommender Systems for better recommendations