**Yelp Recommendation System and Sentiment Analysis**

**Shrikant Mudholkar | Varsha Bhanushali | Monas Bhar**

**ABSTRACT**

Using yelp data, we built a restaurant recommender system for individuals and organizations. For each of 50K Yelp users, we create a ALS model to recommend restaurants to user using users’ preference and previous user selected data. We propose a solution that would recommend restaurant to users and perform sentiment analysis on restaurant review data.

Two main features of the application

* Using user profile, we would recommend restaurants using geospatial location in conjunction with collaborative filtering
* Using the user reviews, we would perform sentimental analysis on the recommended restaurant reviews.

**INTRODUCTION**

Inspired by our recent experience searching Yelp for restaurant recommendations, we built a restaurant recommendation system. Recommendation systems provide personalized, relevant recommendations to users and have been used in various domains, such as retail, movie-going, etc.  
Currently, Yelp, the leading publisher of reviews of local businesses in the world, does not provide recommendations. Instead, users should filter, sort, then read reviews to determine whether a business can provide them with what they want. A personalized recommendation system will provide a better user experience by incentivizing users to review and rate more in return for better restaurant recommendations; this in turn gives Yelp more data that can be used to further improve the recommendation system.

Businesses often want to know how customers think about the quality of their services to improve and make more profits. Restaurant goers may want to learn from others' experience using a variety of criteria such as food quality, service, ambience, discounts and worthiness. Yelp users may post their reviews and ratings on businesses and services or simply express their thoughts on other reviews. Bad (negative) reviews from one's perspective may influence potential customers in making decisions, e.g., a potential customer may cancel a service and persuade other do the same.

**DATA AND METHODS**

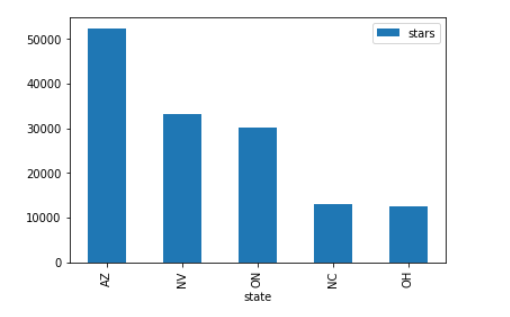
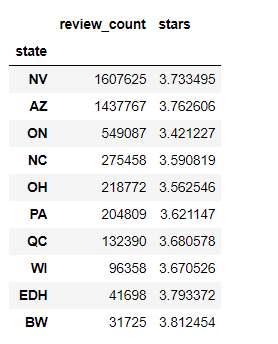
**A. Data Description**

The dataset was originated from the online Yelp Dataset Challenge, consisting of five parts which provides us with 566,000 basic business information (e.g., hours, address, ambience), 2.2 million customer reviews as well as 519,000 tips by 552,000 users. The total size of data is about 2.39GB. For this analysis, we focused on reviews for restaurants and used the customer reviews and business attributes data. These two datasets are both in json format. After filtering restaurants out of all business, there were 1,363,242 customer reviews collected from 77,445 different restaurants. Most of those restaurants are in Arizona, Nevada and North Carolina and the dataset includes a huge variety of cuisine types

Attributes in review data include business id, full address, price range, business categories and etc. Attributes in review data include review content, rating, business id and etc. The attributes we used were business id, business categories, review content and rating. Specifically, review content was the corpus for our analysis; rating was the identifier for discriminating positive or negative sentiment; business id served as the key for data munging and the business categories served as the key for grouping.

**B. Data Cleaning**

The Dataset originally consisted of different categories of business. We clean and filtered the data by cherry picking the categories related to restaurants and included only those business which are related to restaurant category. Further we calculated the mean stars and evaluated the top states with most review count and mean stars.



We eliminated the states with significantly low number of review count and stars. We segregated and stored the data in different files with the respective to states. We also performed join on the three different dataset of users, review and business to form a combined file consisting of users related to a particular business with the user review data.

**C. Code with Documentation**

<https://github.com/shrican/yelp-recommendations-sentimental-analysis/tree/master/src/scripts>

**D. Methods and Algorithms**

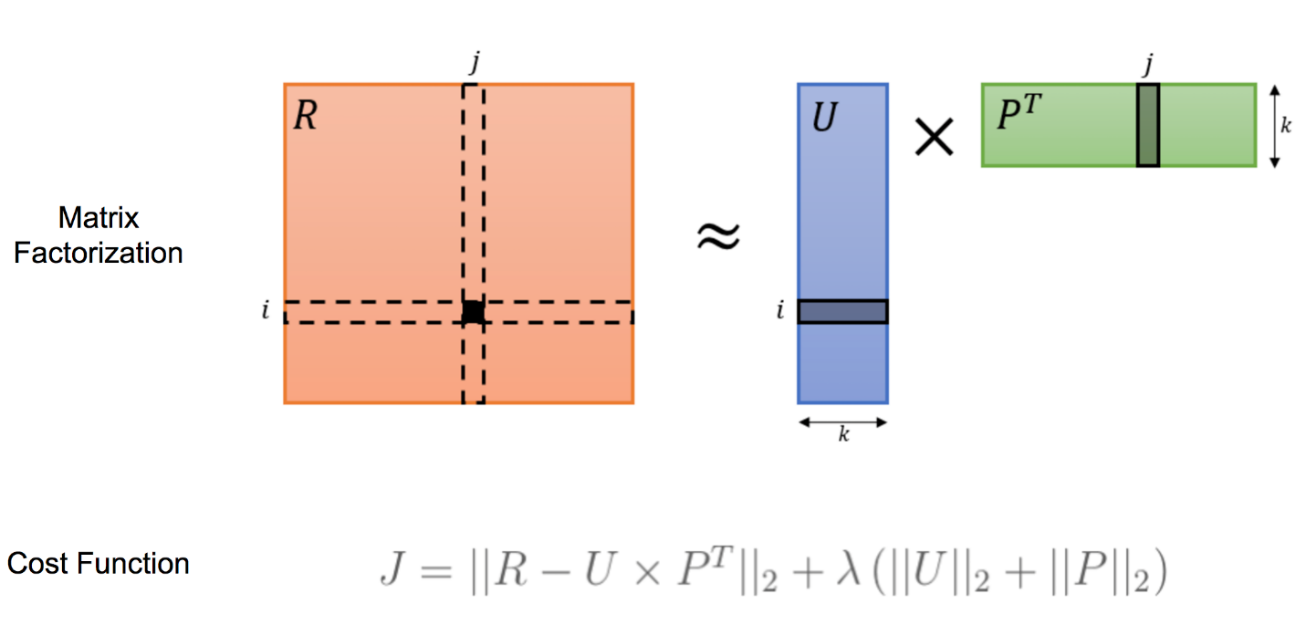
Task Definition

We use data provided by Yelp, described further in the Data section, to build a recommendation system that provides personalized restaurant recommendations to users. Since different people have different food preferences and dietary restrictions, we perform careful feature selection to take advantage of the information reflected in a Yelp user’s reviews.

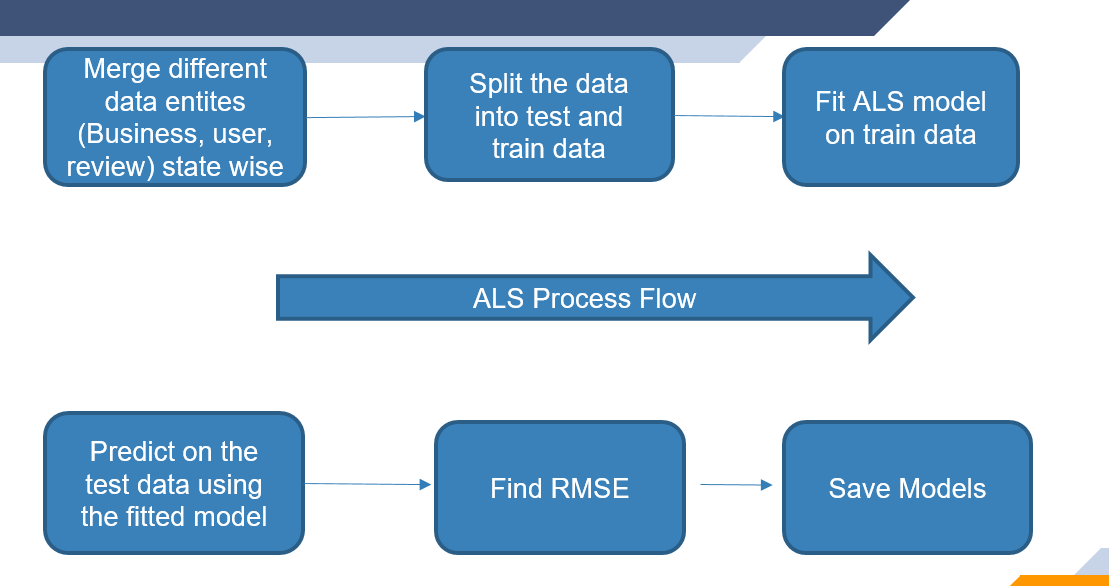
COLLABORATIVE FILTERING

Collaborative filtering (CF) is commonly used for recommender systems. These techniques aim to predict user interests by collecting preferences or taste information from many users. In other words, CF fills in the missing entries of a user-item association matrix. The underlying assumption is that if person A agrees with person B on one issue, A is more likely to have B's opinion on another issue than that of a randomly chosen person.

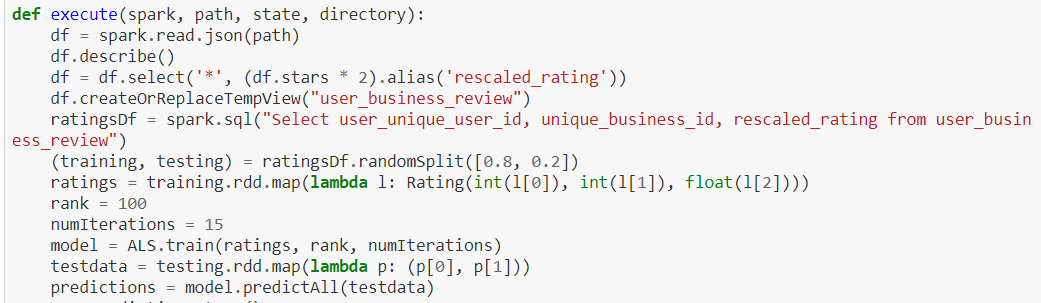
Mathematically, this is done by low-rank matrix factorization, combined with a minimization problem (see picture below). The often-sparse user-item rating matrix R is approximated as a product of user matrix U and item matrix PT, which are built of latent factors. We then form the cost function J, and try to minimize it. Currently in the spark.ml library, the alternating least squares (ALS) algorithm has been implemented to learn these latent factors. Additionally, since we directly rely on the user rating itself, our approach is often referred as "explicit."

[](https://nycdatascience.com/blog/wp-content/uploads/2017/06/CF2.png)

For our project, we pre-train the model and save it. When the recommendation engine boots up, it will load the model and use it for prediction. This architecture is designed so that we can keep training multiple models offline as new data comes in. Once a new model is ready, the recommender engine will make the switch by editing one line of code.



We used Spark ML to build the ALS Matrix Factorization Model for restaurant recommendation. We have createed for different ALS models (one per state) which we pickle and store in the file system after creation. This model is capable of helping us get restaurant recommendations for an existing user as well recommendation for users when given a restaurant.



For each state, we load the json file containing a combined view of user reviews and restaurants. We select user\_id, business\_id, and rating and do a 80:20 random split on the data. The training data is used to train the ALS model and the testing data is used to compute the RMSE. The models are pickled and saved so that it can be consumed by the flask application.

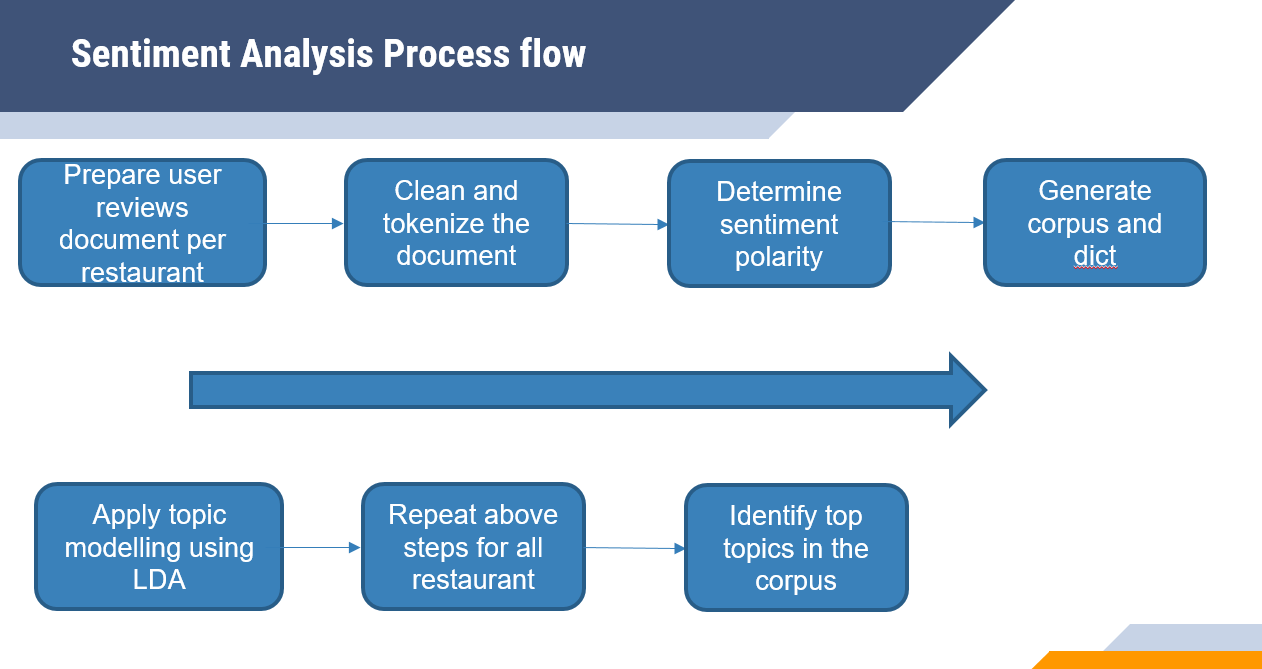
SENTIMENT ANALYSIS

**TOPIC MODELLING**

Sentiment Analysis, also known as opinion mining, is the process of determining whether a text unit is positive or negative. It can have a wide range of applications such as automatically detecting feedback towards products, news and characters or improving customers’ relation model.

To automate the extraction or classification of sentiment from sentiment reviews, sentiment analysis (Basant et al., 2015) uses the natural language processing (NLP), text analysis and computational techniques. Becoming one of the hot area in decision making, sentiment analysis is widely used in many fields such as Consumer information, Marketing, books, application, websites, and Social Media.

Various approaches have been used to evaluate the sentiment underneath the words and expressions or documents. Some of the most common machine learning algorithms used in NLP fields include Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM) (Joachims, 1998), and unsupervised learning (Turney, 2002).

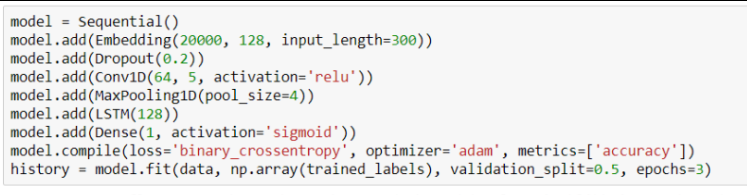


The overall sentiment polarity showed a preference on service in the reviews, which might allude that customers ‘self-select’ the food they like. In the other hand, we could mine many valuable insights that cannot be directly revealed on the dashboard of websites like Yelp. Yelp’s dashboard merely shows an overall rating towards a business rather than several ratings for various aspects of businesses, while we considered it at a more granular word-level.

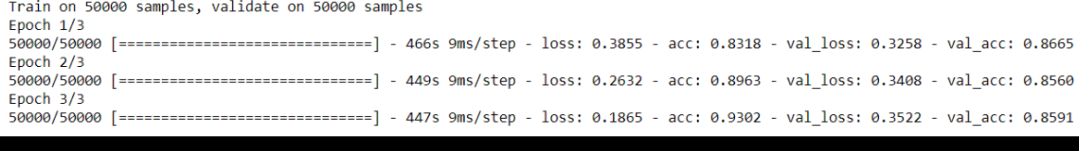
**DEEP LEARNING**

The merged review-business data were randomly separated into training, and testing set according to ratio 5:5. Specifically, we assumed and labeled reviews with ratings greater or equal to 4 as positive while the rest as “negative”. This decision was made based on our observation of the distribution of ratings.

Long Short-Term Memory networks, usually called “LSTMs”, were introduced by Hochreiter and Schmiduber. These have widely been used for speech recognition, language modeling, sentiment analysis and text prediction. RNNs are designed to learn from sequences of data, where there is some kind of time dependency.



The network starts with an embedding layer. The layer lets the system expand each token to a more massive vector, allowing the network to represent a word in a meaningful way. We used Tokenizer to vectorize the text and convert it into sequence of integers after restricting the tokenizer to use only top most common 20000 words. I used pad sequences to convert the sequences into 2-D array.

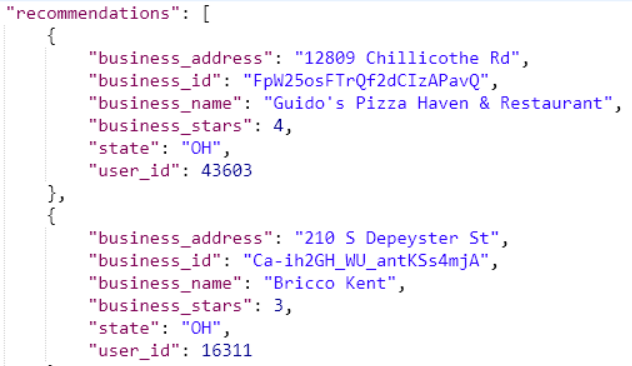


Accuracy and validation accuracy are similar for the epochs. LSTM outperforms the other models when we want our model to learn from long term dependencies. LSTM’s ability to forget, remember and update the information pushes it one step ahead of RNNs.

**RESULT**

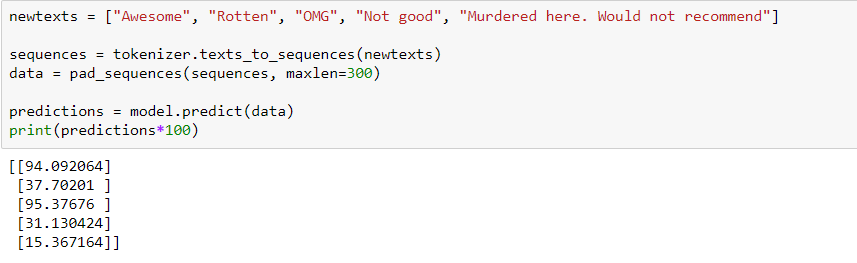
1. **Restaurant recommendation using Collaborative filtering**

Based on the Collaborative filtering ALS model, we have trained our data against various user preferences and taste choices. If the user is already registered we recommend restaurants to user based on their past preferences and if the user is new to the system, we determine the restaurants based on their restaurant categories and business choice.



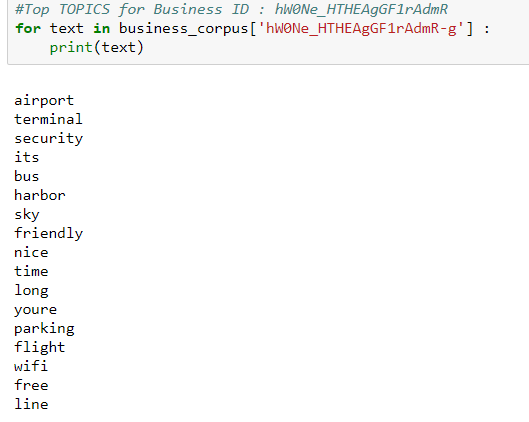
1. **Sentiment Analysis using Topic Modelling**

We have trained our review dataset to predict the polarity of the new data phrases or the test data. It displays the polarity in terms of percentage. 100 being the most positive word and 0 being the most negative word. We have trained our model with enough sarcastic inputs in order to identify the sarcastic review as well and predict the accurate polarity.



1. **Sentiment Analysis using Deep Learning**

Our dataset is modelled to determine the top topics for a particular business id. LDA model works with the DT matrix along with the genism module to infer top topics from the data.



**DISCUSSION**

We tried out various recommendation models for recommending restaurants to user. Prior to ALS, we tried SVD Model. The RMSE score of this model was not as per recommended standards. Also, it took more time to train the model. As per our observation, SGD is not practical if the dataset size is huge, instead ALS is better.

LDA model and train it on Document-Term matrix. The training also requires few parameters as input. The gensim module allows both LDA model estimation from a training corpus and inference of topic distribution on new, unseen documents.

LSTM outperforms the other models when we want our model to learn from long term dependencies. LSTM’s ability to forget, remember and update the information pushes it one step ahead of RNNs.

**REFERNCES**

1. dos Santos, C. N., & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts
2. Huang, J., Rogers, S., & Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. iConference 2014 (Social Media Expo).
3. M. Hoffman and D. Blei. ”Online Learning for Latent Dirichlet Allocation.” Neural Information Processing Systems, 2010.
4. Victor Köhler, https://medium.com/radon-dev/als-implicit-collaborative-filtering-5ed653ba39fe, 2017.
5. Kaushik, Chetan, and Atul Mishra, "A scalable, lexicon based technique for sentiment analysis." arXiv preprint arXiv:1410.2265 (2014).
6. Sahayak, Varsha, Vijaya Shete, and Apashabi Pathan,"Sentiment Analysis on Twitter Data." International Journal of Innovative Research in Advanced Engineering (IJIRAE) Issue 1 (2015).