# Progress Report Two

bjk224, bmc78

November 11, 2017

## 1 Research Context

High ratings and awards can drive large tourist crowds into local favorite restaurants, often causing restaurants to change (increase prices, new booking rules, impersonal service) to better accommodate the new customer base. As a result, the most popular and highest-rated restaurants may no longer be true local favorites but instead thrive on their popularity with tourists. By identifying local users and local experts to provide ratings reflective of local opinions, Yelp may become more popular with tourists looking to "travel like a local" and enjoy a more authentic experience.

## 2 Introduction

The objective of the project is to build and combine two models (Local Expert Identifier / Topical Expert Identifier) for the purpose of identifying 'experts' among yelp users. The "Local Expert Identifier" is a Gaussian Mixture Model that identifies clusters in a given user's review locations to predict the user's most probable location. The "Topical Expert Identifier" is currently a supervised learning algorithm that combines different features about the users reviews in a certain category in order to determine if they are an expert in that category. The goal is to see if an unsupervised algorithm would be able to classify users into clusters of expert and non-expert without needing labels. The goal is to combine the models to find local experts in a specific category.

## 3 Methods and Results

### 3.1 Local Authority Model

#### Part One: Gaussian Mixture Model for Locations

The initial step to building the local authority model is to locate a user given their review locations. This is done by fitting a Gaussian Mixture Model starting with 4 clusters and iterating down (3 clusters, 2 clusters...) until the set of clusters are distinct. The mean of the most probable mixture component is

then used as the predicted location of the user.


## Part Two: Time Spans

The second part to the local authority model is the time span of a user's reviews in an area. This can be found by taking the difference between the max and min date after querying for reviews written by a user in their predicted area. The model uses a predetermined constant TIME_SPAN_CUTOFF to filter out "tourists".

## Part Three: Review Ratios

The last part to the local authority model again queries for reviews written by a user in their predicted area and counts the number of reviews of the user in the area. With this information, the review ratio (ratio of user's reviews in the predicted area) can be calculated. The model uses a predetermined constant REVIEW_RATIO_CUTOFF to filter out "tourists" and also a predetermined constant REVIEW_COUNT_CUTOFF to filter out users without enough reviews to determine locality.

## Evaluation of Local Authority Model

Because there are no "Local Expert" labels for the data, evaluation is done by comparing the predictive ability of average ratings of "Locals" and "Tourists". Assuming there will be observable differences in ratings between "Locals" and "Tourists", our Local Authority Model is significant if it divides users into groups that have more similar ratings. To evaluate the predictive ability of the model, we produce mean-squared-errors of average review ratings by "locals" and actual ratings then compare them to the baseline mean-squared-error of average rating of business and actual ratings.

# 4 Individual Work

**Brian**

- Created evaluation framework for Local Authority Model
- Investigated mean-squared error of models for States and Businesses

**Brandon**

- Built a topical identifier model that initially used supervised learning techniques such as Naive Bayes, Random Forests, and Decision Trees, trained on elite users as 'experts'. Evaluation done by using a section of that data for training and for testing.
- Improved upon initial model by using K-means clustering with 2 clusters to try and identify differences in two sets of yelp users.
- Building evaluation framework for unsupervised learning models based on the average categorical review for 'experts' in the category and 'non-experts' in the category. Currently limited to a small number of categories.
- Evaluate by

# 5 Further Studies

To evaluate the performance of the local expert model, we plan to create a baseline model that predicts a user's rating for a business given the following information: ("user_id", "business_id") Then the baseline model will be compared to a similar model that is trained on "user_id", "business_id", AND "is_local" (whether the user is a local from the area of the business as determined by our local expert identifier model).
In order to improve the topical expert model, we are exploring different feature sets that give more robust interpretations of users, and also working on evaluating the effectiveness of unsupervised learning models.

# 6 References

# References

[1] Jindal, Tanvi. Finding Local Experts from Yelp Dataset. University of Illinois at Urbana-Champaign. Published 2015-04-27.

[2] Zang, Yonfeng. Incorporating Phrase-level Sentiment Analysis on Textual Reviews for Personalized Recommendation. State Key Laboratory of Intelligent Technology and Systems Department of Computer Science and Technology Tsinghua University, Beijing, 100084, China. Published February 2015.

[3] Chee Hoon Ha. Yelp Recommendation System Using Advanced Collaborative Filtering. Stanford University.