ESTIMATING PREFERENCES BY CITY USING YELP DATA

Batuhan Yaman, Mehmet Alperen Derin, Mehmet Mustafa Yilmaz
Department of Computer Science
Hacettepe University
06800, Beytepe, Ankara
{batuhan.yaman,mehmet.derin}@hacettepe.edu.tr
b21127787@cs.hacettepe.edu.tr

ABSTRACT

In this project, we tried to determine what a city values most in a restaurant. We've analyzed Yelp Data[1] through the project. We did pre-processing and used *latent Dirichlet allocation* on it. After getting topics, we did some basic post-processing on results to get our final results. Our final results are satisfactory.

1 INTRODUCTION

What makes a restaurant to get higher star ratings from customers in Yelp is a valuable business information and this information can be inferred from Yelp user reviews. This information can vary across cities. For example in a city where people are wealthy, they will care less about prices but more about the overall quality of food and services. It applies to cousins too. For example, if one ethnicity is a major ethnicity in a city, maybe we can say that people would want to find that cousin in restaurants. This applies to more topics like beverages, services and time of delivery etc. To predict this information we've followed some well-defined steps.

2 RELATED WORK

2.1 LATENT DIRICHLECT ALLOCATION, LDA

LDA is a generative probabilistic, three-level hierarchical model for collections of discrete data such as text corpora in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. [2]

2.2 IMPROVING RESTAURANTS BY EXTRACTING SUBTOPICS FROM YELP REVIEWS

This work aims to generate the most important subtopics for the restaurants in Yelp Dataset. It also predicts hidden star rating for each subtopic, for example, a restaurant that has an overall rating of 4.0 might have a predicted 'service' star rating of 4.5 and 'healthiness' rating of 3.0. It uses "Online Learning Algorithm for LDA" which is a variation of "Expectation Maximization Algorithm" to get the topic distribution from the LDA model. To discover the latent topics for restaurant reviews, this work uses this approach by processing the reviews in 'batches' and updating the topic model incrementally after each batch is processed. [3]

2.3 PERSONALIZING YELP STAR RATINGS: A SEMANTIC TOPIC MODELING APPROACH

This research is aimed to extract a more personalized restaurant rating based on a semantic topic modeling approach. In Yelp, the star rating for a particular business is the mean of all star ratings given to that business. However, a user may only care about the quality of the food, but a 1-star rating given to a restaurant because of a late service ruining a delicious food, the user may pass

on the restaurant and continue searching. In this paper, the author propose an approximation of a modified LDA which conditions topics term distributions not only on the Dirichlet parameter but also on star ratings. The approximation is based on the assumption that star ratings are an approximate function of adjectives of positive and negative connotations within review text. [4]

3 The Approach

Our approach is straightforward and simple thanks to our clear problem definition. To separate what a particular city likes or dislikes about restaurants in that city we've classified the reviews that have 4 or 5 stars as 'positive' and the reviews that have 1 star or 2 stars as 'negative'.

3.1 Storing the Data and Pre-processing

Yelp Data[1] is a well-structured data, but we needed to find a way to access it efficiently. We have done some denormalization on the data and we have stored it in MongoDB, with this way we could access the parts of the data that we needed most. We keep every review as a 'document' in MongoDB with the following structure:

"business_id" : The ID of the business that reviews belong to.

"city" : The city that the business in.

"name" : Name of the business (Not necessary for our approach.)

"stars" : How many stars that review have.

"state" : In which state the city in (Not used in our approach.)

"text" : The review text written by a customer.

We stored only 500.000 of the reviews to MongoDB to test our approach.

Since our input data is natural language, we needed to do some pre-processing on it. We did lemmatisation and removed stop words from reviews, just after we've fetched it from MongoDB. We've used Natural Language Toolkit [5] for this.

3.2 Applying LDA

After pre-processing the data, we have created a corpus from the reviews we stored on MongoDB. We've used scikit-learn's TfidfVectorizer[6] for this. After creating the corpus, we have applied LDA on businesses that have more than 20 reviews and for 'positive' prediction 5 of these reviews should be positive, else it doesn't get evaluated, same applies to 'negative' prediction too. The reason behind this is a business that has so few review could mislead our results since we will count the topics later. The counting step will be explained in the following subsection.

3.3 Post-processing On LDA Results

After getting topics related to documents, if we did LDA with a bigram corpus, we divided the topics into two words. Then we counted the topics for the whole city and we sorted them by count. Finally, we are removing adjectives and verbs and some other vague words between these topics and we obtained the final results. We repeated these steps for both 'positive' and 'negative' comments separately.

4 Experimental Results

We chose three cities of USA to analyze: Las Vegas, Phoenix, and Pittsburgh. The reason for selecting these cities is to compare the difference and similarities due to geographical location, culture, and population. Las Vegas and Phoenix are closer to the southwest of USA and have higher population while Pittsburgh is in the northeast and have less population. After applying removal filter for businesses that are below the threshold, we analyzed approximately 90k reviews for Las Vegas, 110k for Pittsburgh and 170k for Phoenix.

4.1 General Analysis

Table 1: Las Vegas 1

Positive (1,1)	**Negative (1,1)**	**Positive (2,2)**	**Negative (2,2)**
Time	Time	Food	Food
Place	Place	Place	Place
Service	Service	Service	Service
Food	Food	Time	Time
Staff	Order	Staff	Room
Price	Customer	Store	Hour
Restaurant	Day	Price	Start
Location	Restaurant	Experience	Customer
Year	Minute	Pizza	Location
Experience	People	Room	Experience
Store	Location	Chicken	Car
Day	Star	Car	Restaurant
Order	Hour	Restaurant	Staff
Customer	Store	Shop	Pizza
Work	Drink	Bar	Store

Table 2: Phoenix 1

Positive (1,1)	**Negative (1,1)**	**Positive (2,2)**	**Negative (2,2)**
Place	Time	Food	Food
Time	Place	Place	Place
Service	Service	Service	Service
Food	Food	Time	Time
Staff	Order	Staff	Minute
Price	Restaurant	Pizza	Order
Year	Table	Price	Star
Restaurant	Chicken	Year	Location
Chicken	Bar	Chicken	Chicken
Experience	Minute	Experience	Experience
Lunch	People	Bar	Car
Location	Sauce	Store	Customer
Fresh	Cheese	Car	Store
Order	Drink	Location	Pizza
Work	Customer	Sandwich	People

Table 3: Pittsburgh 1

Positive (1,1)	**Negative (1,1)**	**Positive (2,2)**	**Negative (2,2)**
Place	Place	Food	Food
Time	Food	Place	Place
Food	Time	Pizza	Service
Service	Service	Service	Time
Restaurant	Order	Time	Pizza
Price	Table	Restaurant	Minute
Staff	Minute	Bar	Order
Bar	Chicken	Cheese	Cheese
Night	Bar	Price	Restaurant
Menu	Minute	Beer	Experience
Chicken	People	Chicken	Bar
Cheese	Sauce	Coffee	Chicken
Drink	Cheese	Sandwich	Drink
Lunch	Drink	Staff	Star
Order	Customer	Fry	Sauce

From tables Las Vegas 1, Phoenix 1 and Pittsburgh 1 we can infer these:

- Crowded cities like Las Vegas and Phoenix complains about time rather than food. Less crowded cities like Pittsburgh doesn't complain about time. Instead, they value food more. It's a well known common sense that in crowded cities life goes faster and people have less time for eating. This shows our data is not incoherent.
- We can see specific dishes in Pittsburgh and Phoenix but there is no specific dish in Las Vegas. Since Las Vegas is visited by so many tourist absences of a preference about dishes could be normal.
- We can see the word 'Room' in Las Vegas, but not in Phoenix or Pittsburgh. This could be related to the high density of hotels in Las Vegas.

Table 4: Las Vegas 2

Positive (2,4)	**Positive (2,4) Collapsed**	**Negative (2,4)**
Customer Service	Service	Customer Service
Love Place	Food	Food Good
Staff Friendly	Place	Tasted Like
Food Good	Happy Hour	Credit Card
Great Service	Fried Rice	Fried Rice
Great Food	Mexican Food	Horrible Service
Food Great	Price Reasonable	Looked Like
Good Food	Room Clean	Terrible Service
Great Place	Ice Cream	Oil Change
Service Good	Recommend Place	Parking Lot
Happy Hour	Lunch Special	Resort Fee
Friendly Staff	Chinese Food	Service Good
Fried Rice	Prime Rib	Chinese Food
Mexican Food	Fast Food	Bad Service
Price Reasonable	Old school	Good Food

Table 5: Phoenix 2

Positive (2,4)	**Positive (2,4) Collapsed**	**Negative (2,4)**
Love Place	Service	Customer Service
Customer Service	Food	Food Good
Great Service	Place	Tasted Like
Food Good	Happy Hour	Happy Hour
Great Food	Mexican Food	Mexican Food
Staff Friendly	Chip Salsa	Minute Later
Great Place	Chinese Food	Service Good
Food Great	Great Price	Chinese Food
Good Food	Price Reasonable	Recommend Place
Service Great	Carne Asada	Fried Rice
Happy Hour	Lunch Special	Horrible Service
Mexican Food	Fried Rice	Waste Time
Service Good	Fast Food	Chip Salsa
Friendly Staff	Ice Cream	Egg Roll
Chip Salsa	Egg Roll	Credit Card

Table 6: Pittsburgh 2

Positive (2,4)	**Negative (2,4)**
Food Good	Customer Service
Love Place	Food Good
Great Food	Tasted Like
Staff Friendly	Service Terrible
Great Place	Egg Roll
Beer Selection	Food Mediocre
Food Great	Chinese Food
Happy Hour	Looked Like
Best Pizza	Giant Eagle
Bar Food	Ice Cream
Great Service	Minute Later
Food Delicious	Service Slow
Service Great	Half Hour
Friendly Staff	Fried Rice
Chinese Food	Food Average

In this second part, we used ngram(2,4) for corpus and did not parse down the final to one word to generalize. Instead, we wanted to analyze more specific details. This caused some redundancy in our results. Such as "Food Good" and "Good Food" which basically tells us the same thing. That is why in some tables, there is a collapsed version of the results. Also, positive words occur in negative results. This is because of our cleaning algorithm. Algorithm simply removes the words that are irrelevant to subject such as *"the", "was", "not"*. So in other words, *"The food was not good"* becomes *"food good"*. In the beginning, this may seem wrong but we are looking for the subject, for the topic. So this tells us if we are considering a negative review, it's about food and its quality. It's about *"what"*, rather than *"how"*. From tables Las Vegas 2, Phoenix 2 and Pittsburgh 2 we can infer these:

- We can see that in every city, people care about "*Happy Hours*".
- Since Las Vegas and Phoenix are closer to the Mexico, They have more Mexican food culture than Pittsburgh. Though "Chinese Food" can be seen in all of the charts. Which is normal because it has no effect with the location of the cities.
- In the Collapsed part (Table 4 and 5) and Positive part for Pittsburgh (Table 6), most common foods can be seen. We can observe the change in the food culture and compare the similarities to generalize.
- All of the cities complain mostly about food and customer service rather than cleanness or price.

5 CONCLUSIONS

We've used Yelp's data to analyze the preferences and estimate them. While doing it, we used mainly LDA and NLTK. We've seen that we can develop a profile of cities by their preferences about related to food and restaurants. We have presented a way to determine these preferences from comments about restaurants. We can uncover known and unknown subjects without prior knowledge about the city. Our work categorized as *Unsupervised Learning* so that we could only evaluate our results with common sense in a theoretical way.

We had some weaknesses during the process. After classification of the review with starts as positive or negative, we assume that every sentence in that review is classified correctly. In other words, if a user gave a 4-star review and it includes 3 positive and 1 negative review, we take as 4 positive reviews. Basically, our main defect comes from NLP problems.

For future work NLP processing can be done more efficiently. Prior knowledge about the cities would benefit during analyze process so working with such people would improve results. Also applying this algorithm to more cities and providing a visual map might provide a deeper understanding of preferences.

REFERENCES

[1] Yelp Dataset - $https://www.yelp.com/dataset\_challenge$

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[3] Huang, James, Stephanie Rogers, and Eunkwang Joo. "Improving restaurants by extracting subtopics from yelp reviews." iConference 2014 (Social Media Expo) (2014).

[4] Linshi, Jack. "Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach." Yale University (2014).

[5] Natural Language Tool Kit - $http://www.nltk.org/$

[6] TfidfVectorizer - $http://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.TfidfVe$