IMPROVING RESTAURANTS BUSINESS

MEGHNA SRIVASTAV, MONICA MAVOORI
SHIVPRIYA TAMBOSKAR, NIHARIKA SALADY

1 INTRODUCTION

YELP ratings clearly have a profound effect on the success of businesses as "an extra half-star rating causes restaurants to sell out 19 percentage points more frequently" (increase from 30% to 49%). But how can a restaurant point out the demands of its customers from a large amount of reviews? We hope to identify what users care most about when writing their reviews, and ultimately determine what certain restaurants are doing right and wrong in order to receive these ratings.

With a large amount of data, it becomes difficult to extract prominent or relevant features. By breaking the reviews down into latent subtopics using LDA, we are then able to predict a restaurant's star rating per hidden topic. Ultimately these ratings per hidden topic allow us to pinpoint the reasons for a restaurant's Yelp rating, other than food quality. Some latent subtopics that were extracted from Yelp reviews include service, value and family. Additionally, other topics such as meat, veggies and dinner also came up in our findings and proved useful.

Also, an analysis of top 1000 users with maximum fan following helped to narrow down the amount of text to be analyzed. The latent subtopics derived from reviews by these users and the ones derived earlier show a high similarity between them and hence can be helpful.

Another analysis carried out was based on the location of the business. If we were to suggest to a business on where to start, we found out cities with not too many 5 star rated restaurants and also an analysis of what people in that particular area prefer or comment on. We hope to provide a base-line on what to keep in kind in terms of what people like in that area.

1.1 WORK DISTRIBUTION:

Name	Solution Design	Implementation	Data Analysis	Data Visualization	Test Phase	Reports And Presentation
Shivpriya Tamboskar	25 %	15%	25%	15%	30%	20%
Niharika Salady	25%	15%	25%	15%	30%	20%
Meghna Srivastav	25%	35%	25%	25%	20%	35%
Monica Mavoori	25%	35%	25%	45%	20%	25%

2 BACKGROUND

2.1 BACKGROUND BASICS

There are many approaches to factor models for discrete data. We use the Latent Dirichlet Allocation (LDA) factor model to approach the unsupervised learning of factors and topics for the Yelp restaurant review data. This model treats the probability distribution of each document over topics as a K-parameter hidden random variable rather than a very large set of individual parameters (K is the number of sub topics). An alternative to it, Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) is an algorithm which models the document space in a different manner using nearest neighbors.

2.2 RELATED WORK AND WHY OURS IS DIFFERENT

Earlier works in this area have predicted start ratings of reviews using sentiment analysis and predicted business categories using K-mean clustering and rating businesses based on hidden topics.

The Yelp dataset that we work on has information on reviews, users, businesses, and business check-ins. We specifically focus on restaurant data with regards to each type of information. Related work on this dataset include: predicting the category of a restaurant given a text document; Markov Chain review generators that generate reviews automatically; finding the most repeated words of a corpus of reviews. The results are presented in a format which is easily understood by one and all. Also, we try and find alternative to using a smaller dataset with a more select group of users rather than the whole group to base the suggestions forwarded. Suggesting where a new business can be put up using our research is also an easy task.

3 ALGORITHM AND SYSTEM DESCRIPTION

3.1 TOOLS USED

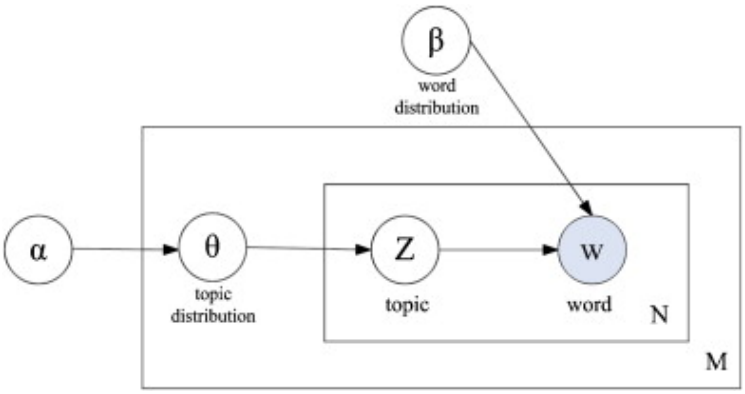
Algorithm and Data Analysis Scripts	**Python**
Python Libraries	**Genism, PyNUM, NLTK**
Database	**MongoDB**
Servers	**AWS Servers**
Visualization	**D3 and Tableau**

3.2 LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) [2] is a Bayesian generative model for text. It is used as a topic model to discover the underlying topics that are covered by a text document. LDA assumes that a corpus of text documents cover a collection of K topics. Each topic is defined as a multinomial distribution over a word dictionary with |V | words drawn from a Dirichlet $\beta k \sim$ Dirichlet(η).

Each document from this corpus is treated as a bag of words of a certain size, and is assumed to be generated by first picking a topic multinomial distribution for the document $\theta d \sim$ Dirichlet(α). Then each word is assigned a topic via the distribution θd, and then from that topic k, a word is sampled from the distribution βk. θd for each document can be thought of as a percentage breakdown of the topics covered by the document.

To discover our latent topics for restaurant reviews we used online learning approach where reviews were processed in "batches" and the topic model was updated incrementally after processing each batch. We find topic models for our text corpus for a range of topic numbers K and for |V | = 10000. After stop-word removal, only the top 10, 000 recurring words by frequency are considered. We found that K = 40 gave us very reasonable results for our restaurant review dataset. For small topic numbers, vocabulary belonging to separate topics would become grouped into single topics, and for large topic numbers vocabulary that we might associate with a single topic (such as service, time) would become separated into several individual topics



W -> Word

Z -> Topic for the word W

θ -> Topic distribution for a document

α -> Per document topic distribution

3.3 FLOW OF IMPLEMENTATION

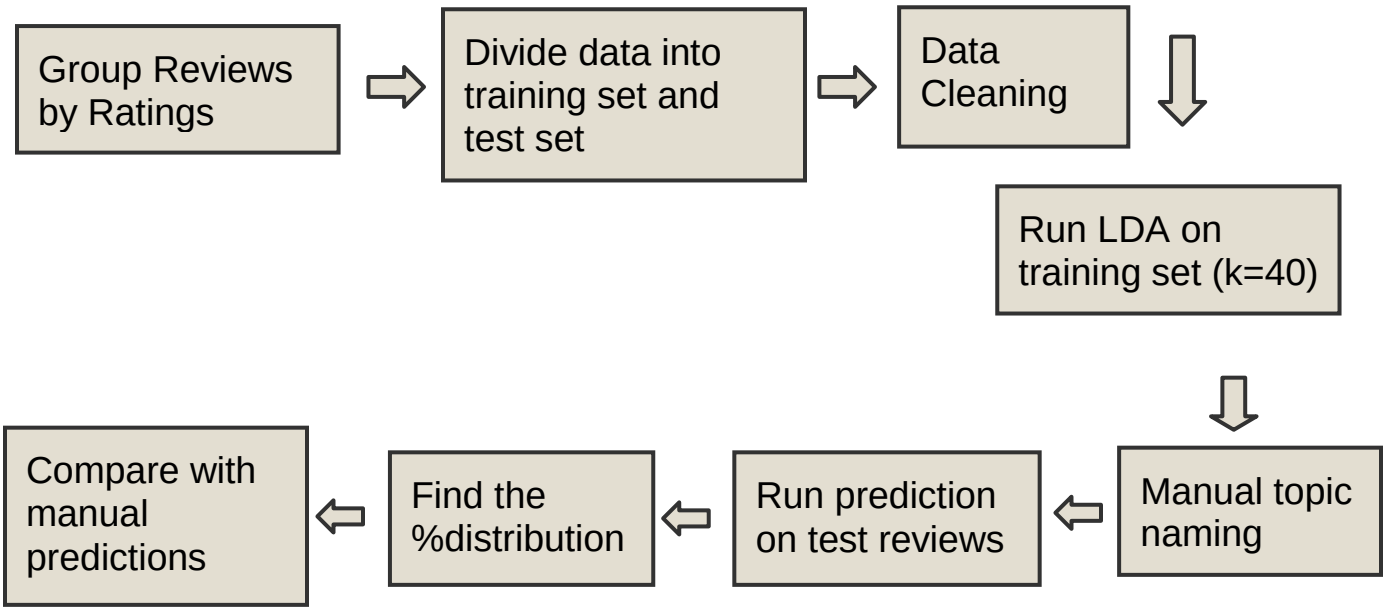
Our Implementation is focused mainly on three sections.

a. To predict Topic Distribution of a NEW review given as input. And suggest the business in improving some topics which users has written in their reviews.

b. Finding few top topics from Topic Distribution of highest Star-Ratings and suggesting those topics for business with low ratings in improving those areas to improve their business.

c. Find busy hours and suggest businesses to maintain and take care of the calculated top topics mainly during those hours to improve their businesses.

Below is the Flow model of our implementation in finding topic distribution from all Review Corpus.



The reviews were first grouped by their rating, then further divided into training set and test set. This data was then stored in form of a collection in database. Next, this data was cleaned by splitting the review into sentences, removing stop words, extracts parts-of-speech tags for all the remaining tokens, stores each review, i.e. reviewId, business name, review text and (word, pos tag) pairs vector. Then, after applying WordNetLemmatizer (nltk) to lookup the lemma of each noun, each review together with noun's lemma is stored in a separate collection. To this collection, we applied genism LDA model to train our model using the training data with k = 40 and V=10,000.

Here, we have chosen k value to be 40 as the outcome for this very much appropriate for our analysis and whereas k with 30 is very generalized without allowing to name topics uniquely and k with 50 is too specialized by repeating the similar topics frequently.

After this, we have manually named each topic in the topic distribution presented as presented below and also give a distribution of the k different topics we come up with.

```
(Breakfast)#0: 0.211*sandwich + 0.037*shop + 0.029*turkey + 0.029*bread + 0.025*lunch + 0.024*salad + 0.020*place + 0.013*cheese + 0.012*meat + 0.
012*order
(Decor)#1: 0.056*ramen + 0.035*dim + 0.020*design + 0.018*twist + 0.016*fall + 0.016*wall + 0.015*mix + 0.015*honey + 0.013*doubt + 0.013*photo
(Cost)#2: 0.040*bland + 0.036*money + 0.032*box + 0.028*girlfriend + 0.026*girl + 0.024*mess + 0.021*soggy + 0.020*bag + 0.019*save + 0.018*worst
(Restaurant Bar)#3: 0.095*strip + 0.081*cocktail + 0.054*lol + 0.034*mall + 0.030*ambience + 0.028*fix + 0.025*drink + 0.024*vegetarian + 0.
024*martini + 0.021*sake
(Service)#4: 0.205*steak + 0.151*star + 0.130*awesome + 0.053*medium + 0.034*rare + 0.023*service + 0.021*sprout + 0.018*movie + 0.018*side + 0.
017*rating
(Burger)#5: 0.208*burger + 0.170*fry + 0.032*onion + 0.031*bun + 0.020*cheese + 0.020*bacon + 0.019*order + 0.018*wrap + 0.015*patty + 0.014*place
(Italian)#6: 0.056*fast + 0.049*caesar + 0.048*hubby + 0.034*finger + 0.032*south + 0.032*eaten + 0.029*lemonade + 0.028*italian + 0.025*online +
0.025*john
(Service+staff)#7: 0.045*customer + 0.037*order + 0.031*owner + 0.024*service + 0.023*guy + 0.021*business + 0.021*staff + 0.020*food + 0.
019*time + 0.016*employee
(Taste+Location)#8: 0.077*chili + 0.066*style + 0.027*chain + 0.024*school + 0.021*diner + 0.020*pig + 0.018*version + 0.017*eatery + 0.
016*gourmet + 0.016*recipe
(Food)#9: 0.240*lunch + 0.053*fish + 0.049*tuna + 0.033*dinner + 0.030*today + 0.029*salad + 0.023*salmon + 0.021*na + 0.021*tempura + 0.020*ayce
```

We also do this with all the reviews, irrespective of rating to get the overall picture and find the topic distribution as below.

Figure: Topic Distribution after naming LDA output

a. Analysis on New Review and Topic Distribution:

The next part is prediction algorithm, where we use the training corpus to predict the topic distribution of the current review given as input.

> *I have never been to this place but I am pretty sure this is aweful. Botto.. What kind of name is that.. badd... Bistro.. who calls their restaurant Bistro? Sounds like a big ball or water or authentic Italian..? what a farse. I am disappointed they do not seve kungPo chicken or chicken vindaloo.. How can they miss those? Also... no sushi.. Califinia's own food? Shame on them.*
>
> *I have never seen their waiters but pretty sure they are bad. I ask for food and they never bring it and will laugh at me., Ohh also they hate colored folks..*
>
> *Altogether, never ever been to this restaurant because it sucks in Yelp reviews!*

Our analysis on topic distribution of the above review with the saved LDA model is as below.

[(7, 0.173979336376367435), (8, 0.09344804518265865), (16, 0.22013217061090186), (14, 0.05372209472083155), (24, 0.01535985308065378), (31, 0.334829314265223476), (33, 0.046977300077683241), (34, 0.094438343698184689)]

The outcome is planned to visualize as below so that a naïve user or business owner can easily realize the topics he has improve in to improve his entire business.

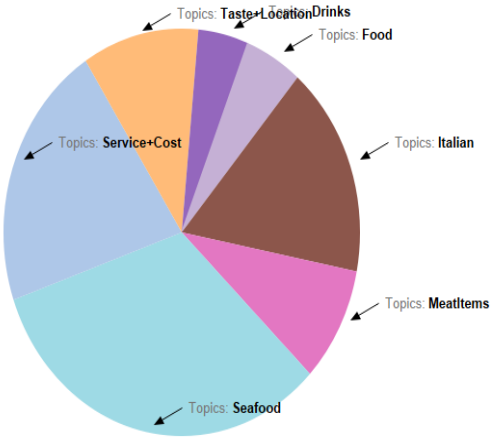
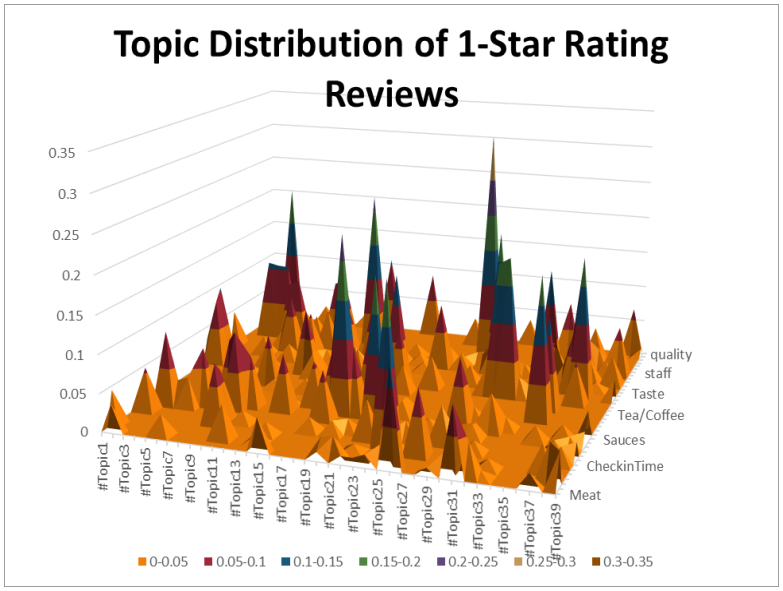


Figure 1 **Figure 2**

Figure 1: Indicated the statistics of the probability of each topic
Figure 2: Highlighted topics indicate the topics from the review based in saved LDA.

We also did Review Analysis on reviews with Star-Rating=1 and Star-Rating=5, and compared the difference of the topic distributions of both the outcomes. As expected a variation has been found among the topics. It indicated that Users mostly talk about few topics in bad reviews inferring those businesses to improve in those sections to improve their business.

Below is the Topic Distribution of the reviews with Rating as 1.



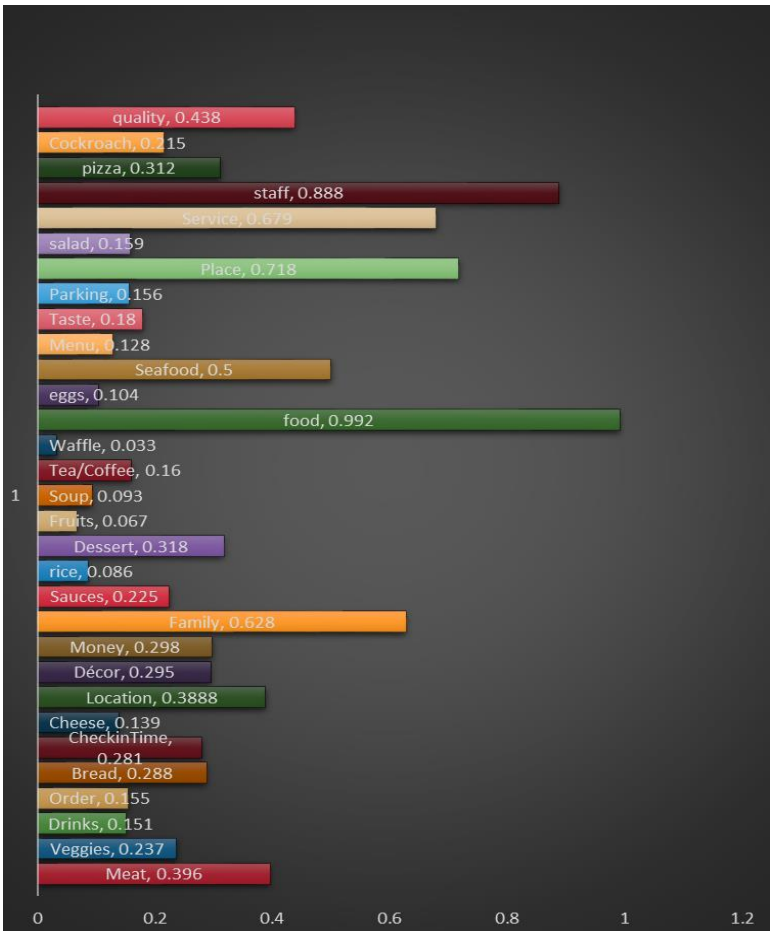
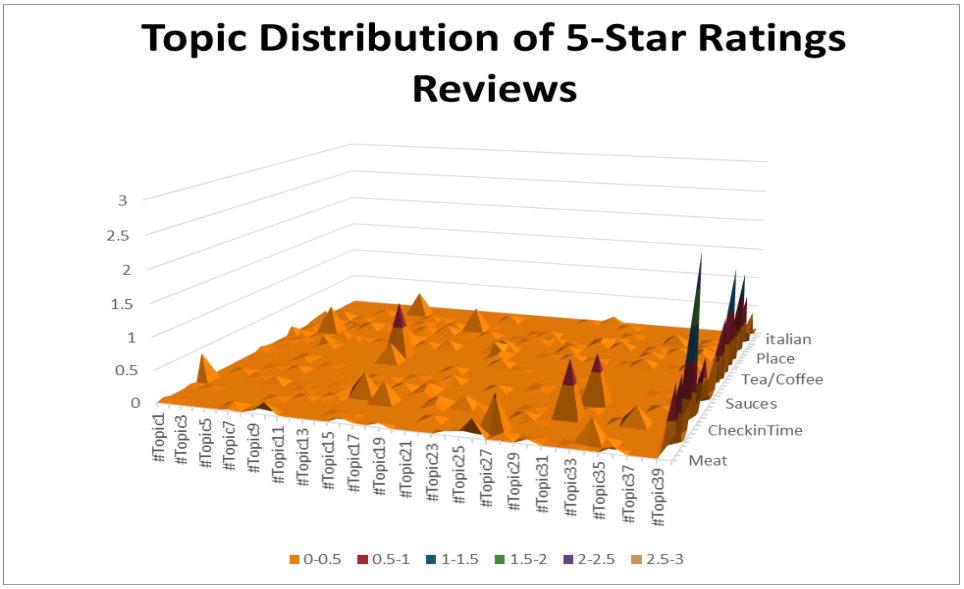


Figure: Statistics of Topic Distribution of the reviews with Rating as 1

Below is the Topic Distribution of the reviews with Rating as 5



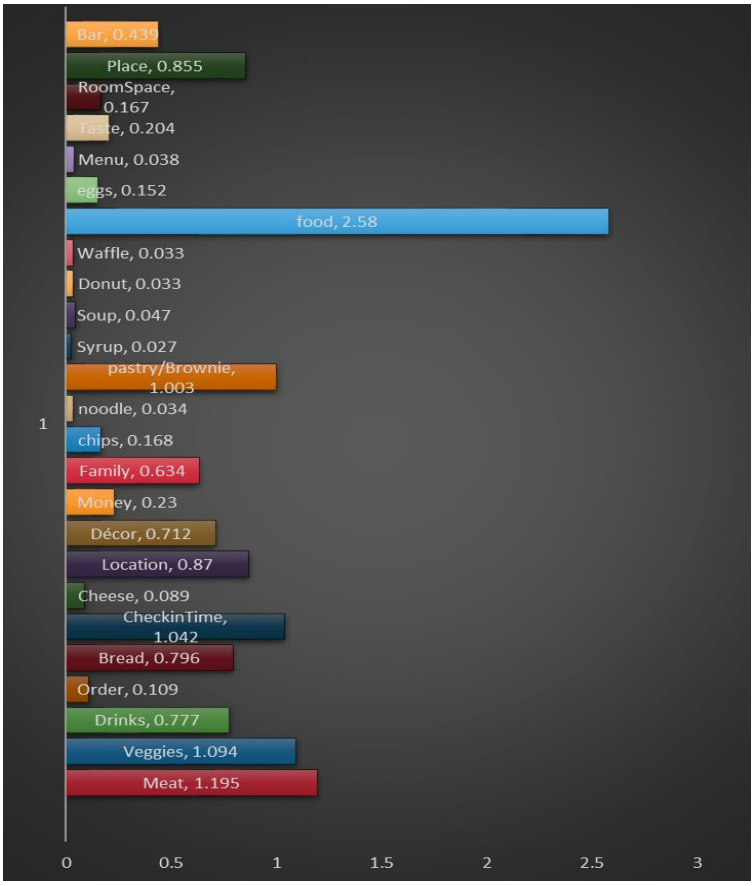


Figure: Statistics of Topic Distribution of the reviews with Rating as 5

This analysis revealed that even though a lot of the times we got the same topics but their inference would be vastly different according to the rating that review had gained.

Topic	**For 1- Rating**	**For 5- Rating**
Food	No more varieties available, Tastes bad	Have good Taste and quality
Staff	Staff is not responding properly	Good compliments has been received by staff
Service	Service is not fast and appropriate	On time service
Family	Not a good place to bring babies along	Good place for whole family

Since the number of reviews to be analyzed remains high, we devised another means to reduce the number. Through the users object, we select the top 10000 users by number of fans and use only their reviews to find a topic distribution. Not only do we get similar topic distribution, we get similar results from our prediction algorithm.

> *I have never been to this place but I am pretty sure this is aweful. Botto.. What kind of name is that.. badd... Bistro.. who calls their restaurant Bistro? Sounds like a big ball or water or authentic Italian..? what a farse. I am disappointed they do not seve kungPo chicken or chicken vindaloo.. How can they miss those? Also... no sushi.. Califinia's own food? Shame on them.*
>
> *I have never seen their waiters but pretty sure they are bad. I ask for food and they never bring it and will laugh at me., Ohh also they hate colored folks..*
>
> *Altogether, never ever been to this restaurant because it sucks in Yelp reviews!*

[(7, 0.09697903472047589), (8, 0.05344804518265865), (16, 0.210087988080808), (14, 0.1598080808999008), (24, 0.0135985308065378), (31, 0.219898914265223476), (33, 0.1897234810375490235), (43, 0.0998927401736489202 12)]

We can safely say that using a handful of users that have most influence on the yelp reviews, can give us a more narrow and accurate dataset to work with.

b. Location Based Analysis:

For a new entrepreneur in Restaurant industry, we hope to find a location ideal for the new setup. We have an idea of where the maximum 5 star ratings are coming from.

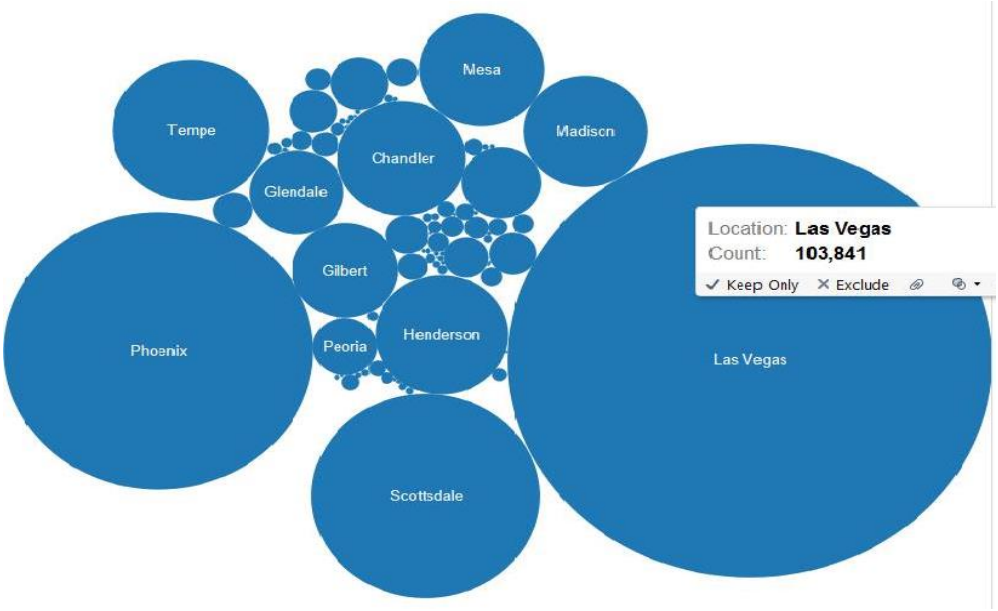


Figure: Number of 5-star Restaurants in different Cities

From the above visualization we can clearly identify the cities with maximum number of high rated restaurant. So for a new business it would be tough to open a new joint in these areas as it will become really challenging to understand customers food trends in that area along with high competition.

Now, for setting up a new joint we can suggest the below idea which helps in improving their businesses

i) Choose a mid-sized location among those
ii) And concentrate on topics found from 5-Star ratings Reviews.

Choosing a mid-sized location signifies that customers in that area to be food-lovers. And concentrating on top topics of 5-star reviews leads customers to love the new place and possibility of rating such restaurant will be high.

c. Check-in Time Analysis:

We also analyzed the check-in time of the customers and compared it with the meals that are talked about.

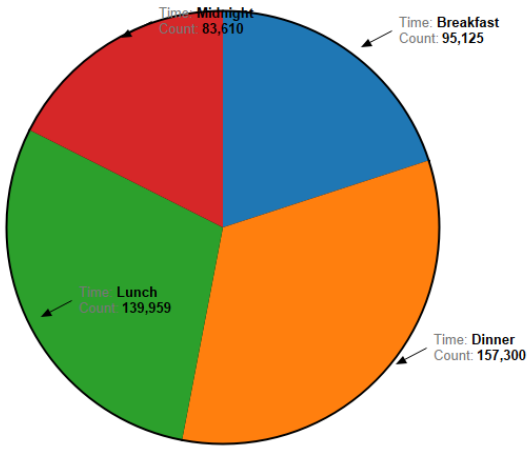


Figure: Check-in Trends for Ratings >3

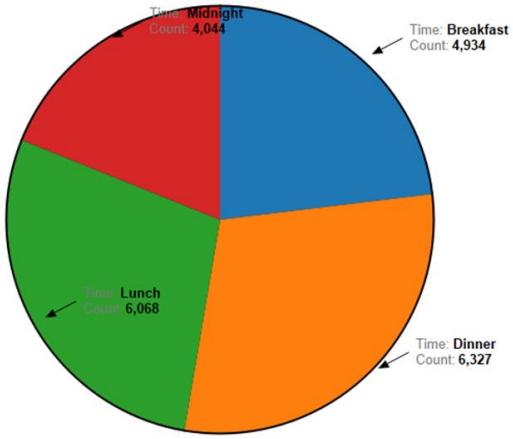


Figure: Check-in Trends for Ratings < = 3

As seen from the topic distribution, it's the lunch and dinner time meals that are most talked about. We can safely say that these are the ones that get rated the harshest so all owners must consider themselves henceforth warned.

4 TESTING PHASE

As mentioned our dataset is divided as

- 90% training set
- 10% testing set

After working with training set and LDA, testing set is applied on saved LDA model. The results found are realistic which means LDA has been modelled perfectly with training data set. This testing data set is used in prediction of its topic distribution as well as run for new LDA model. Both has shown almost similar results which means our prediction on new review is correct.

5 EXPERIMENTS

5.1 DATASET

The data is taken from Yelp Data Challenge. It roughly consists of data of 92 cities and about 706,646 reviews. This dataset includes business, review, user, and checking data in the form of separate JSON objects. A business object includes information about the type of business, location, rating, categories, and business name, as well as contains a unique id. A review object has a rating, review text, and is associated with a specific business id and user id. A check-in object contains check-in time with a specific business id and user id. These are the ones we deal with.

5.2 SUCCESS MEASURE

Our implementation has 2 parts to be measured.
I. Mainly topic distribution
II. Other is prediction of a new review

I. Topic Distribution:

Metric: Topic distribution can be measured on variation of number of topics of the outcome.

Success:

- The topics must be able to label individually. The subtopics grouping manually should be able to group to a unique topic. LDA with input parameter as 40 shows the necessary variation among the topics found
- Difference between topic distributions of various star ratings.

II. Predicting New Review:

Metric: The input review given should be distributed with related probabilities of the topics decided.

Success: Accurate probabilities of the topic distribution shows a successful implementation. It should predict the distribution from the saved LDA model with trained corpus. Topics apart from those shows a failure which we have not come across. Also exact probabilities related to topics has been predicted for a review contain huge amount of text in it. It means data cleaning algorithm training has been done perfectly.

5.3 EXPERIMENTAL HIGHS

Below points are considered as experimental highs

- Dealing with big reviews in-terms of its length while analyzing its Topic Distribution.
- Choosing appropriate number of topics with LDA algorithm, which is suitable to our data analysis.
- Analyzing Topic Distribution of top 1000 users to be very similar to the distribution of whole review corpus, which saves a lot of time to data scientists.
- Clean results indicates data is pre-processed correctly.

6 CONCLUSION

The project was a good learning experience in terms of identifying the various technologies that would be required to most suitably implement the project and then learning those technologies. In the process of development in a team, we have learned to develop the code in modules and to later on integrate them since it was built in stages by different team members. This required a considerable amount of planning if some of these modules were being developed in parallel such as knowing beforehand what would be the format of the output/document that would generated so that if it is the input to the next module then it would be developed accordingly. For example, in the data cleaning process we had to ensure that all the documents had utf-8 encoding and the code for each of the cleaning module was developed such that it accepted and outputted a document that was utf-8 encoded. In terms of systems the first difficulty we faced was to identify what server size in amazon AWS was the most appropriate so that our implementation works effectively without under-utilizing larger sized server or making the implementation run on a server which is smaller and thereby not allowing the system to run smoothly. This was done mostly by trying and testing different sizes (small/mid/large) of servers on amazon AWS. Since we were dealing with large datasets optimization was one of the major difficulties we faced. We had to repeatedly optimize the code so that it would process the data faster such as reducing the number of for loops used and removing irrelevant data. Also, in the initial stages although we could come up with an idea that could be useful to the business owners for analysis we face difficulties in implementation and data availability in the following ways 1) We had to make sure the correct amount and type of data was available 2) The problem was realistic and within the scope. In terms of visualization of the analysis we performed one of our major concerns were to make the visualizations understandable and meaningful to even naïve users and not just to us as programmers. It was very important to find out which visualizations would be the most helpful in improving businesses and which ones were irrelevant. Taking suggestions from people who were not involved in the project helped in this case. We used the following tools in our implementation: 1) Amazon AWS for processing 2) Python NLTK for cleaning the data 3) Python Genism package for topic modelling 3) MongoDB for storing large datasets 4) Tableau for data visualization.

7 REFERENCES

- Improving restaurants by extracting subtopics from yelp reviews – James Huang, Stephanie rogers, Eunkwang Joo
  http://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_ImprovingRestaurants.pdf
- Prediction of yelp restaurant review score – Xi Sun, Xin wang, Zhuoer wang
  http://newport.eecs.uci.edu/~xis2/Yelp/Final-report-pp16.pdf
- Data Set - https://www.yelp.com/dataset_challenge/dataset
- Python gensim - https://pypi.python.org/pypi/gensim
- Topic distribution - http://www.vladsandulescu.com/topic-prediction-lda-user-reviews
- Python NLTK - https://pypi.python.org/pypi/nltk
- MongoDB - http://docs.mongodb.org/manual/tutorial/install-mongodb-on-ubuntu