Two 100k yelp review samples were taken for yelp data exploration. **First** sample of approximately 100k restaurant reviews from 15 categories was considered for basic exploration tasks of ratings distribution, topic mining and two subset topic mining. The **second** sample of 100k from 10 categories was considered for cuisine reviews distribution and cuisine topic mining.

Following criteria was considered for selection of the sample Yelp reviews:

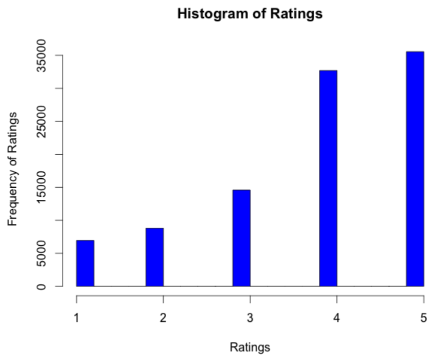
- One of categories of the business should be ‘Restaurant’
- category considered for sample should have at least 30 reviews in it

183 out of 240 categories were found to have at least 30 reviews associated with it. Out of 183, only 15 sample categories were considered for first sample and 10 for second sample because of the CPU hardware configuration constraints.

<u>First Sample Data Exploration</u>

Ratings:

It was observed that around 68% of the reviews in the sample had high ratings (i.e. a rating of 4 or 5). Below is the histogram of the ratings from Yelp reviews sample: (using R for histogram)



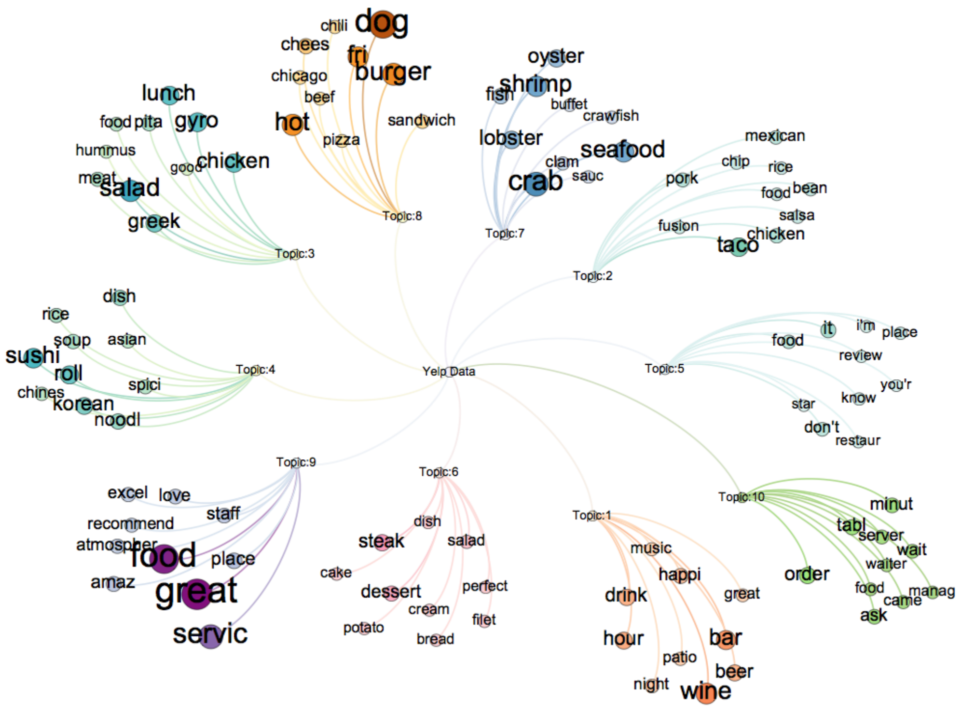
1.1 Topic Mining

LDA algorithm was used to discover topics. MeTA tool was used for applying LDA algorithm on sample reviews.

Below are the **parameters** used for applying **LDA** in **MeTA**:-

Inference=”gibbs”; max-iters=1000; alpha=1.0; beta=1.0; topics=10; model-prefix=”lda-model”

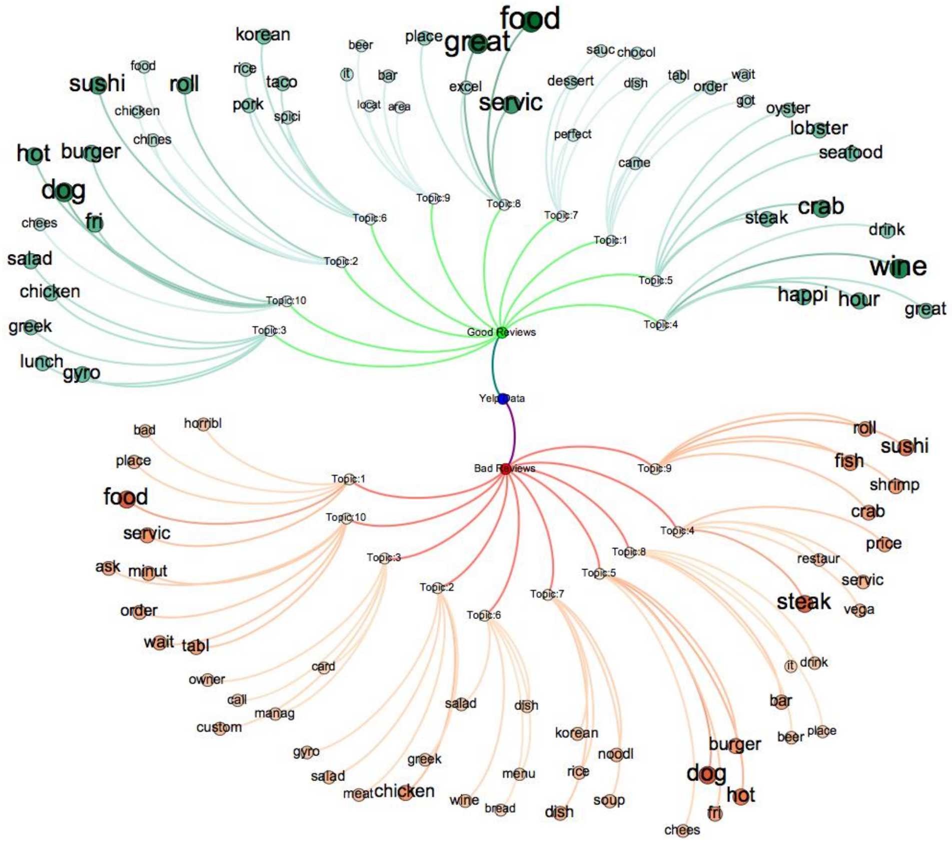
The number topics were kept as 10 so that high level topics can be mined first. Also it was easier to understand and visualize the output for small set of topics. **Gephi** tool was used to create this visualization. The output from LDA algorithm was written to text file and **Python** parser was written to convert the topics output into graph node/edges csv files that Gephi can import. Here is the visual output of topics that were discovered by LDA algorithm:



The opacity of each node corresponds to its weight in each topic. The node label size and node size also indicates the weight of that word in the topic. Colors are used to distinguish separate topics. It can be observed that major topics are based on **restaurant service, ambience or food**. From above we can observe topics such as **American sea food (Topic 7), Japanese/Korean sea food (Topic 4), Mexican (Topic 2), Restaurant Service (wait time, server, manager etc. – Topic 10), Drinks (Topic 1), Ambience (Topic 9), and Mediterranean (Topic 3)**.

1.2 Subset Topics Mining

For this task, two subsets of sample reviews were taken for analyzing difference in the topics. The first subset had reviews associated with high ratings (4 or 5). The second subset had reviews associated with low ratings (1 or 2). LDA was run on the two subsets (good and bad reviews) separately with the **same configuration as 1.1.** A similar process was followed to convert the topics output into nodes/edges CSV that Gephi can import. Gephi was used for visualization in this task as well. Following visualization gives an idea of different topics associated with good and bad reviews.



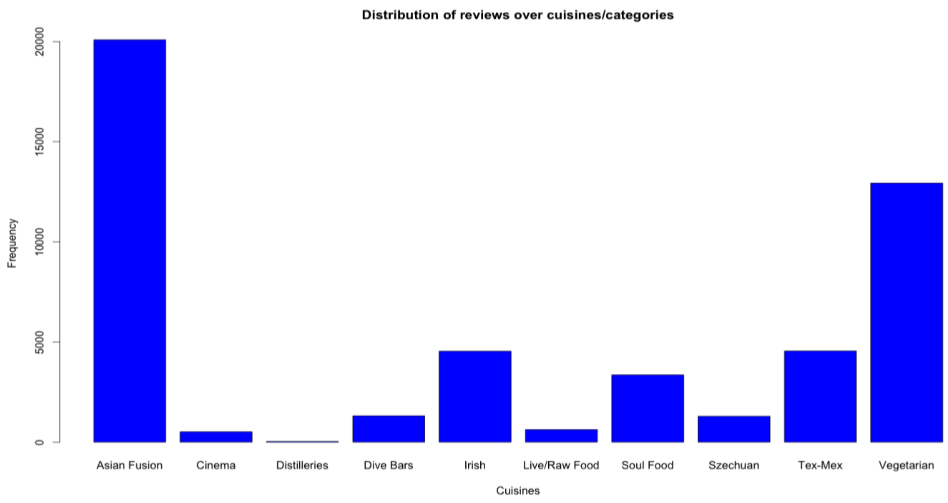
Here again the opacity of each node corresponds to its weight in each topic. The node label and node size also indicates the weight of that word in the topic. **Green** indicating topics for good reviews and **red** indicating topics for bad reviews. We can clearly see some important contrasting topics in good vs bad reviews. Under good reviews we can see **appreciation of great food (Topic 8), ambience and service (Topic 1 and 8)** while in bad reviews we can see topics **like horrible/bad service (Topic 1), *possible high* prices (Topic 4) and *high* wait times (Topic 10)**. But there is partial overlap as well because both have topics regarding different types of dishes (both good/bad review topics contain discussion of various cuisines like Mediterranean, sea food and American fast food).

Second Sample Data Exploration

The second sample was taken from 10 categories and the main objective of this sample was to explore topics specific to certain cuisines.

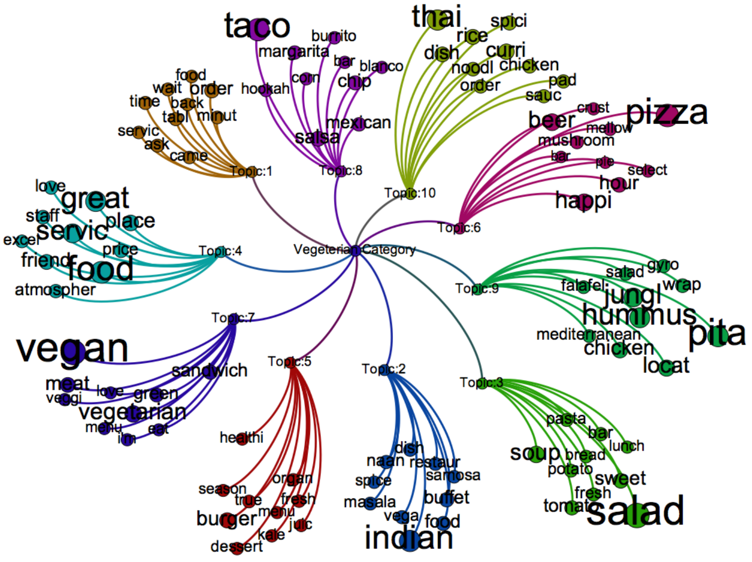
Cuisine Exploration:

Here is the distribution of reviews for different cuisines based on second sample of 100k reviews. (Using R for histogram)



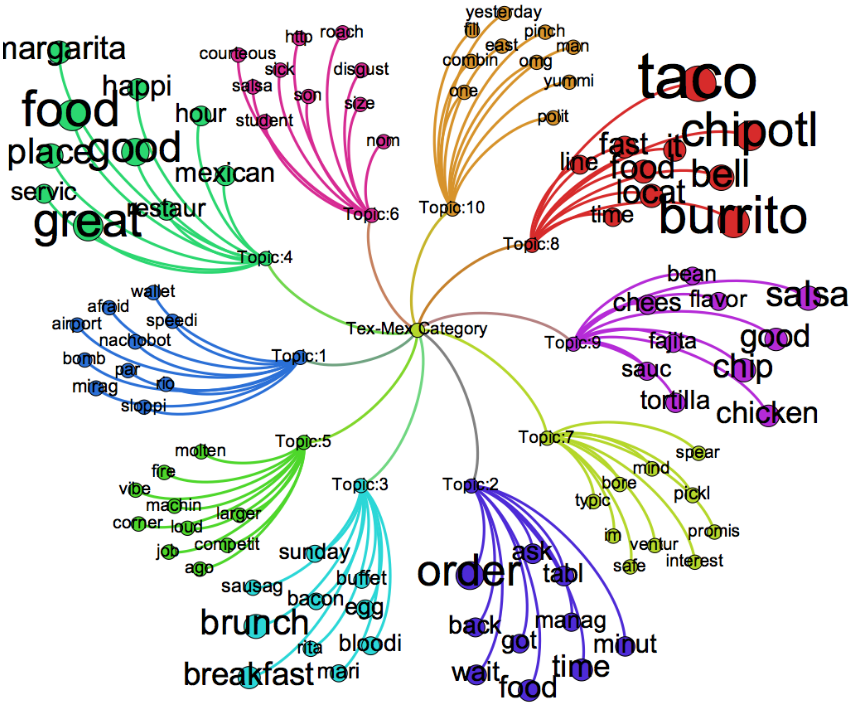
From the above distribution, we can see that Asian Fusion and Vegetarian category have high number of reviews in our sample. Vegetarian was an interesting category to further explore what topics could contain in that. Vegetarian and Tex-Mex categories were considered further for topic mining.

Again, **LDA** algorithm with **same configuration as 1.1** was applied on vegetarian and Tex-Mex categories reviews using **MeTA**. **Gephi** was used to produce the visualizations. The visual elements are similar to 1.1 and 1.2 (Color indicating different topics, label size/node size indicating the weight of the word in that topic)



From the visualization we can see vegetarian have a lot of overlapping cuisines depending on whether that cuisine has vegetarian options. Some good vegetarian overlapping cuisines that we can conclude from above are **Indian (Topic 2), Mediterranean (Topic 9), Mexican (Topic 8), Italian (Topic 3) and Pizzas/Drinks (Topic 6). T**he words that contribute to these topics also suggest some good vegetarian dishes like hummus, samosa, salsa, salad, tomato soup, potato etc. There is one other topic (**Topic 5**) which contains words like **juice, organic, kale, fresh** suggesting some good **healthy** options.

Here is Tex-Mex category topics visualization:



Tex-Mex topics seem to give idea on various options available under Tex-Mex cuisine such as **Taco/Burrito (Topic 8), tortilla/chicken/beans (Topic 9), Breakfast dishes like sausage, bacon (Topic 3)** – **overlapping** with **American brunch**. There are some other topics which discuss about **food service (Topic 4)** and **food quality/hygiene (Topic 6)**.