CS-598|Data Mining Capstone

Task 1: Yelp Data Exploratory Analysis

Net ID: shaukat2

Description

I have used mainly business and reviews json files for this analysis. I performed join on these two datasets based on business id. Then I dropped irrelevant columns to keep the analysis clean and tidy. First analysis is performed on all open businesses with review counts more than 9. Second analysis is particularly restricted to restaurants business type.

Topic Modelling

Topic modelling is done for samples selected randomly. I applied topic modelling on four tasks.

1. Random 25000 sample from whole dataset for business type(category) as restaurant. **(Task 1.1)**
2. Random 25000 samples from ratings >=4 (positive reviews) for business type(category) as restaurant **(Task 1.2)**
3. Random 25000 samples from ratings <=2 (negative reviews) for business type(category) as restaurant **(Task 1.2)**
4. Random 10000 samples from most positively rated restaurants for location i.e. Las Vegas, NV

Data Pre-processing:

Data for topic modelling is pre-processed and cleaned by apply Gensim Bigram model to review text and then stop words removal, stemming and lemmatization was applied to make data coherence. This data is then used to build dictionary and calculate term frequencies.

Model Specifications

LDA model from Gensim library was applied for topic modelling. Following parameters were set.

num_topics = 10	Eval_every = 5	Random_state = 12345
Iterations = 100	Passes = 10	Per_word_topics=True

Visualization Specifications

For visualizations, I used matplotlib, seaborn and WorldCloud libraries.

Other Libraries:

Nltk	Spacy	Pandas
Numpy	Collections	Pickle

Environment Specifications:

I used Anaconda with Python 3.9 on Windows 10[16 GB RAM, NVIDIA].

Task 1: Yelp Data Exploratory Analysis

Star Rating Distribution Analysis:

- First analysis is done on all business types which shows that positive ratings are dominant which implies that most businesses are performing good or above average. This is shown in Chart 1.
- Second Analysis is performed on specifically restaurants category type which shows that good and above average reviews are still dominant.

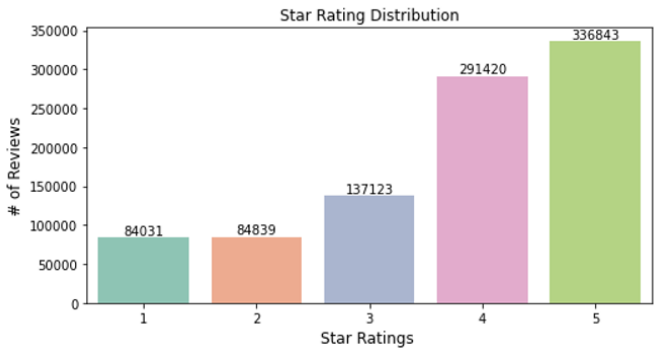

Chart 1: Star Rating Distribution

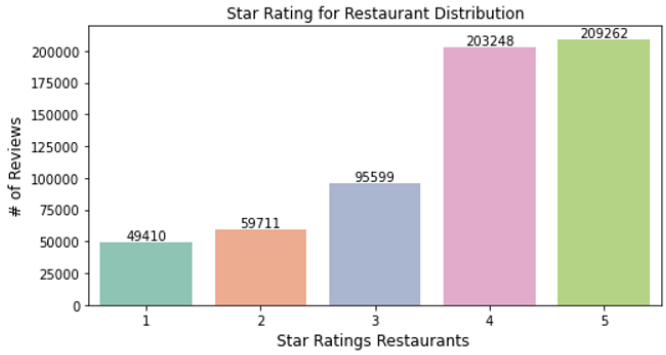

Chart 2: Star Rating Distribution for Restaurants

Top 10 Rated Cities and Cuisine Analysis:

- In Chart 3, top 10 cuisines are identified based on reviews. American Cuisine is at first place, while 'Misc' includes generic categories like Pubs, bed and breakfast etc.

- Chart 4 depicts 10 highly rated cities distribution which shows that most highly rated restaurants are in Las Vegas, NV.

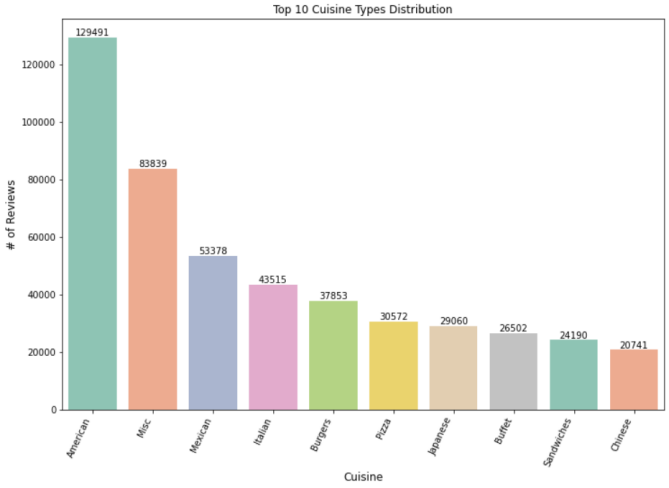

Chart 3: Top 10 Cuisines Distribution

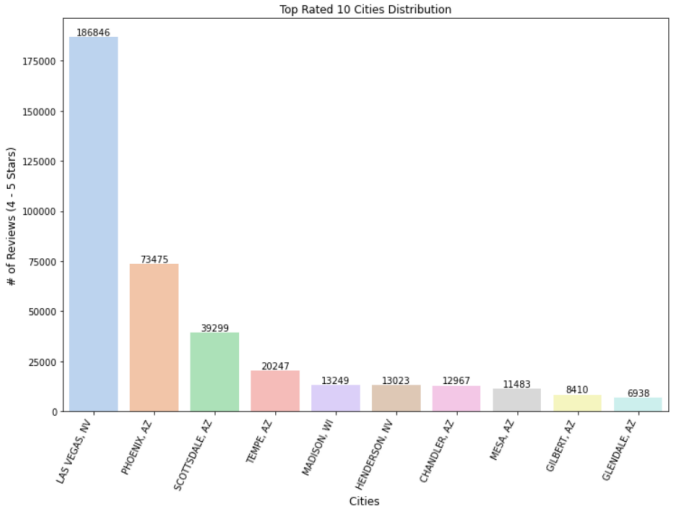

Chart 4: Top Rated 10 Cities Distribution

Top Rated Restaurants:

Following are the top 10 rated restaurants with mean ratings and number of review counts.

	name	Cuisine	mean	count
0	Mon Ami Gabi	French	4.515957	3008
1	Earl of Sandwich	Sandwiches	4.593308	2899
2	Hash House A Go Go	American	4.494086	2621
3	Wicked Spoon	Buffet	4.473161	2012
4	In-N-Out Burger	Burgers	4.616066	1805
5	The Buffet	Buffet	4.448860	1711
6	Bouchon	French	4.541212	1650
7	Bachi Burger	Burgers	4.584740	1599
8	Bacchanal Buffet	Buffet	4.582888	1496
9	Grand Lux Cafe	Desserts	4.353599	1431

Task 1.1 – Topic Modelling on Review Text

LDA model was applied on 25000 randomly selected samples. This was done using Gensim library with parameters described previously. I used word cloud to generate the distinct topics and the most important words which are proportional to weight.

- Chart 4 displays word clouds for top 10 topics.

Chart 5: Word Clouds for Top 10 Topics

- Chart 5 shows word counts vs word weights for each topic. It seems like that topic 1 might be a good choice for topic selection.

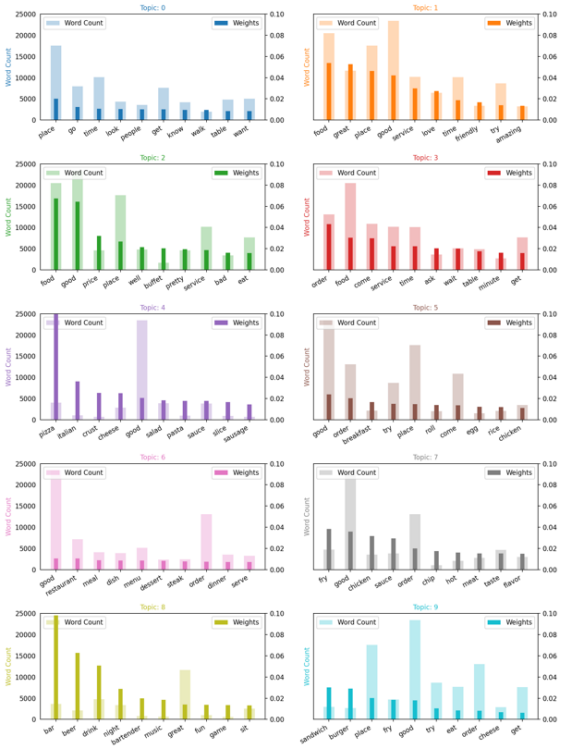

Chart 6: Word Count vs Word Weight Distribution for Top 10 Topics

Task 1.2 – Topic Modelling on Review Text Subsets

For this task, I performed topic modelling on positive and negative reviews data for restaurant category type.

- Chart 7 shows word cloud for positive topics.
- Chart 8 shows word cloud for negative topics.

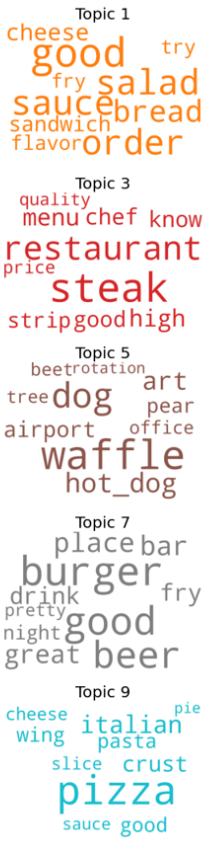

Chart 7: Word cloud for Top 10 Positive topics

Chart 8: Word cloud for Top 10 Negative topics

I also performed topic modelling for Top Rated location for restaurants which is Las Vegas, NV.

- Chart 9 shows top 10 topics for reviews in Las Vegas, NV restaurants

Chart 9: Top 10 Topics in Top Rated Location (Las Vegas, NV)

Discussion:

Overall, comparing various subsets of review data allowed for interesting analysis with LDA topic modelling. We can see that LDA has done a pretty good job of creating topical clusters for positive reviews, negative reviews, and reviews for restaurants in Las Vegas, NV. For instance, the word clouds for the positive reviews highlights words such as: "great", "good", "high", "great", etc. The word clouds for the negative reviews highlights words such as: "dirty", "bad service", "loud", etc.

The word clouds associated with the Las Vegas reviews show words such as: "great", "friendly", "good", "love" etc. These prominent words within the topics make intuitive sense, but some of the topics themselves are reasonable.