Data Mining Capstone Task 4 & 5

March 20, 2020

Asad Bin Imtiaz CS598
Data Mining Capstone
MCS-DS UIUC
Email: aimtiaz2@illinois.edu

1 Introduction:

In this task, popular dishes in the Yelp Review Academic Challenge Dataset for specific cuisines were mined and ranked for popularity by considering reviewer's sentiments and other factors about the restaurant facility. Moreover, a ranking for restaurants is also determined for popular dishes based on sentiment and rating analysis of reviews and restaurants. The analysis and findings are part of Task-4 and 5 for Data Mining Capstone course, wherein the Yelp restaurant reviews were explored to identify popular dishes' and restaurants' rankings for a specific cuisine type. The goal of this task was to come up with an appropriate approach to rank and classify popular dishes and recommend the respective restaurant businesses they are offered at.

In the following, the implementation of these tasks sub tasks for this task are explained.

1.1 Tools and Libraries:

The exploration and analysis of Yelp restaurant dataset was performed with **Python 3** using Jupyter notebook. Several python libraries such as **Gensim**, **scipy, numpy** and **Scikit-learn** were used to analyze and cleanse the review texts. For visualization, **D3.js** was used to generate bar charts.

1.2 Application:

The first step was to select a cuisine to explore popular dishes in. There was a dish list for Chinese cuisine already provided by the staff. However, this list did not look complete or belonging to a specific cuisine, therefore the popular dish list extracted in Task-3 using **Seg-Phrase** and **Auto-Phrase** for **Indian** Cuisine was used instead. A total of 160 popular dishes were selected and standardized. Standardization means, since several of the dish names were not in English language and were spelled differently, they were mapped to a standard dish name.

2 Task 4: Ranking popular dishes

In this task all reviews referencing dishes in the selected popular dish list for Indian cuisine were extract and analyzed. There were 2 challenges in doing this analysis.

1. Standardized Dish-name extraction
2. Ranking criteria for popularity

These subs tasks are discussed in chapters below.

2.1 Standardizing dish names

Several dish names have words that are different but are synonyms or have the same meaning. For example, **'vegetable roll'**, **'veg roll'** & **'veggie roll'** all represent the same dish but have different text representations. Moreover, sometimes non-English names are literally translated while other times are written only in English spellings. For example, '**Aloo**' is the literal sound of potato in Hindi/Urdu (Language in India). Sometime '**Potato Samosa**' is written in text as '**Aloo Samosa**' but represent the same dish. The spelling while writing with English character set may be different such as **'Aloo'**, **'Allu'** or **'Aaloo'** all would represent the potato in Hindi. Apart from that, if only dish occurrences are counted, than dish '**Tikka-Masala**' may be counted more times than '**Chicken Tikka Masala**' and '**Vegetable Tikka Masala**' (being a sub-string of later dishes). The generic form of the dish (such as **pizza**) has to be counted separately than the specific form (such as **pizza margherita**) while ranking.

To overcome these issues, following additional processing was done:

- Synonyms or words representing same thing were mapped to a common standard word
- Common spelling variations were also standardized
- Specific and Generic dish names were processed separately

Once the dish name extraction was standardized, selected dishes from all reviews (not limited to category 'Indian') were extracted if the standardized name are referenced in the reviews.

2.2 Ranking popular dishes

In this sub task, a ranking formula was developed to rank the dish name. One naïve assumption could be that the popularity of the dish corresponds to number of reviews it is mentioned in. However, a high-rated review mentioning it once with positive sentiment and a low-rated verbose review mentioning it several time cannot cancel the effect of one another. Therefore, apart from review rating and sentiment, the number of times a dish is mentioning in the review is also important for the ranking.

If a review is endorsed by others, its credibility increases and thus the number of users finding a review useful also influences the popularity ranking. Additionally, the presence of other dishes in the same review verses the presence of only a single dish in the review would also matters.

Thus, a ranking criterion was established based on the following arguments:

- The popularity of dish is directly proportional to the sentiment rating of the review. The higher the rating, the popular the dish.
- The popularity of dish is proportional to the number of times it is mentioned in the review. However, the increase is sublinear and should give diminishing returns with higher count of references.
- The number of users who find review useful have a proportional impact on the ranking of the dish.
- The number of distinct dishes in the review takes away some of the rank rating, because then it is not clear to which dish the positive or negative sentiment has a higher association with.

Based on the above arguments, following ranking formula was developed to rant the dishes:

$$\mathrm{DishRank}_{\,dish} \quad = \quad \frac{\sum_{i=1}^{n} \ln\,(\,1 + ((\,1 + \mathrm{Useful}\,) * \mathrm{Star} * \mathrm{RevCnt}) + \ln\,(1+\,1/\mathrm{DishCnt})}{n} * \ln\,(1+\,\mathrm{TotalCnt})$$

where:

n = Total number of reviews referencing a specific dish

$Useful$ = Useful rating of review

$Star$ = Star rating of review

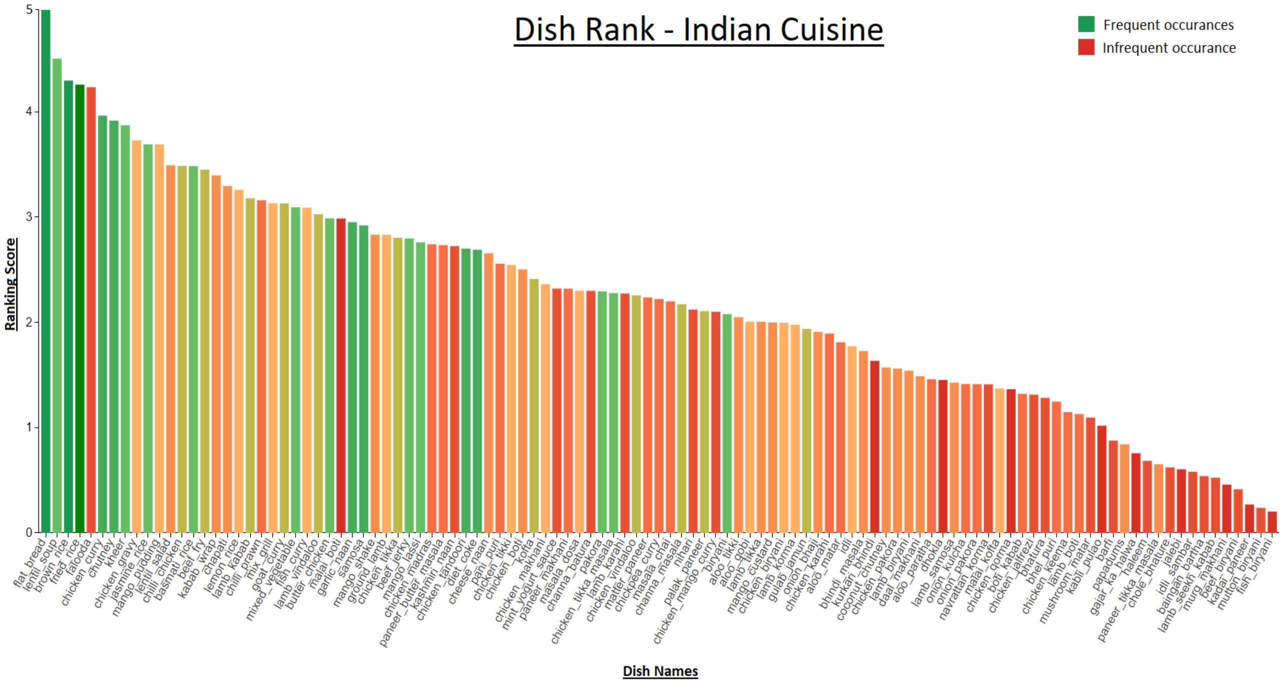
$RevCnt$ = number of references of dish in review

$DishCnt$ = number of unique references of all dishes in review

$TotalCnt$ = total number of references of a dish in all reviews

The first log in summation can be considered as the TF (Term frequency) of the review while the log in second part of the equation is F in whole document space. The second log in summation acts as IDF (Inverse document frequency).

After ranking the dishes with this formula, following visualization was generated (only showing top 100 dishes). There the Hight of the bar represent the popularity rant of the dish while the color represents the unique occurrences of dish in all reviews or scale from red to green.



As can be seen, 'Flat Bread', 'Lentil Soup', 'Brown Rice', 'Fried Rice' & 'Falooda' are the top raked dishes, while mutton and fish biryani are among the least popular. In general, infrequent mentions of a dish carry negative sentiments, while the ones more often mentioned are those that are mostly liked. However, there are some exceptions to it. For example, the dish 'Falooda' and 'Malai boti' both were mentioned rarely, however were ranked high probably die to higher sentiment score and useful scores.

3 Task 5: Restaurant Recommendation

In this subtask the goal was to recommend good restaurants to those who would like to try one or more dishes in a cuisine. This task was done in following two steps :

1. Analyzing review sentiment for a restaurant – *Chapter 3.1*
2. Ranking a restaurant for a dish – *Chapter 3.2*

These subs tasks are discussed in chapters below.

3.1 Analyzing review sentiment for a restaurant

To analyses the sentiment of user for a review, sentiment tagging was performed. List of common negative and positive words and a **Multilayer Perceptron Classifier** (MLPClassifier) was created using **Sklearn** Python library, which gave a rating score between 1 and 5, with less than three being negative, more than three being positive and 3 being neutral. The cumulative review scores gave a sentiment score for the restaurant.

The reason already provided rating score of restaurants was not used was because a phrase-based scoring for sentiment was done, where a phrase such as '**not good**' or '**rarely tasty**' were tagged negative phrases to improve the accuracy of the sentiment rating. This was done by keeping a list of negation words ang generating noun parses from these as well as from adjectives and adverbs preceding nouns.

3.2 Restaurant Recommendation

In this sub task, a ranking formula was developed to rank the restaurants based on sentiment scored for all candidate dish names. Again, simple count of occurrences of dish name in the reviews for restaurants was not taken as popularity criteria, rather the sentiment score restaurant and dish rant as calculated in chapter 2.2 was combined. The reasons behind the ranking criteria was that:

- The sentiment rating of a restaurant is proportional to the rank of the restaurant
- The number of occurrences of dish name in the restaurant reviews increases the restaurant rank, but with diminishing returns.
- The dish ranks is proportional to the restaurant rank for that dish.

Based on the above arguments, following ranking formula was developed to rant the dishes:

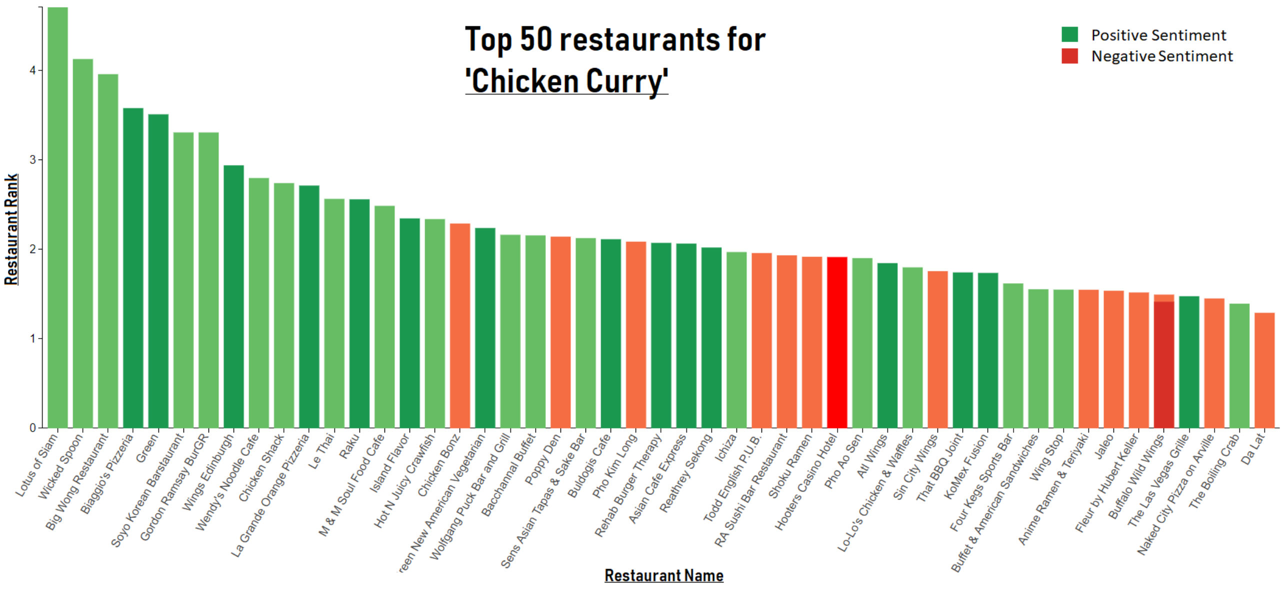
$$\text{RestaurantRank}_{dish} \quad = \quad \ln\,(1+ \text{DishCnt} * \text{SentimentRank} * \text{DishRank} \,/\, \text{TotalCount})$$

where:

$dishCnt$ = Total number of references of dish name in all reviews for a restaurant
$TotalCnt$ = total number of unique references of all dishes in all reviews of restaurant
$SentimentRank$ = Sentiment score of the restaurant
$DishRank$ = Rank of the dish as calculated in 2.2

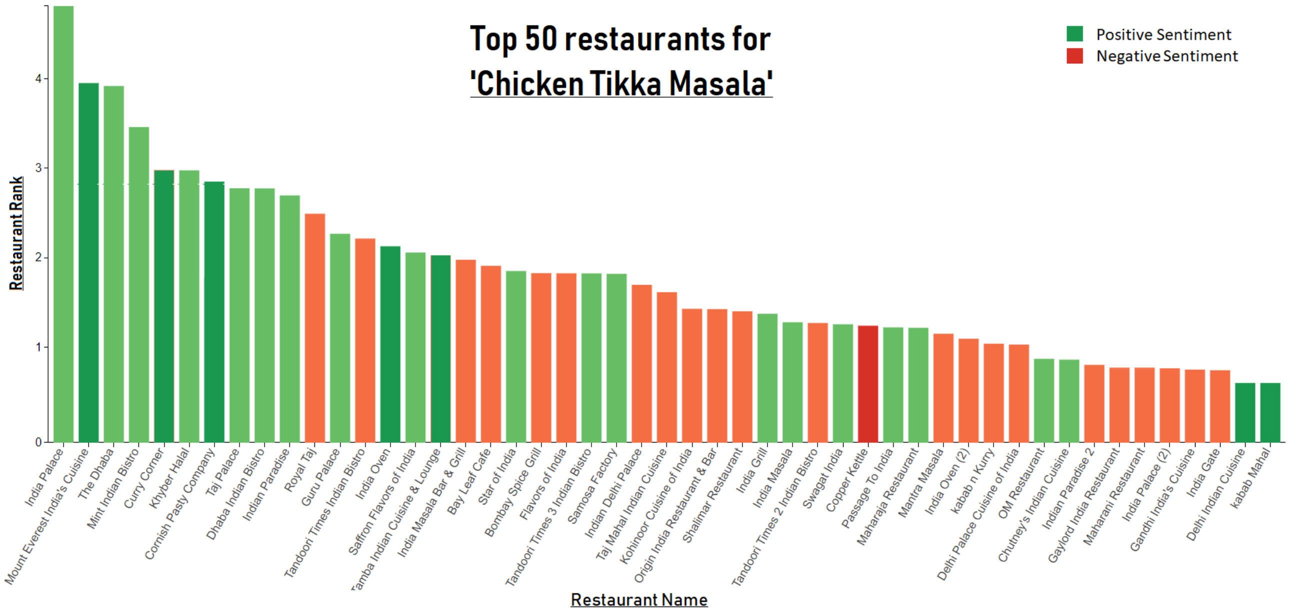
The presence of log gives a sublinear transformation to the number of occurrences of dish name in reviews for the restaurant while acting as the **Term Frequency** (TF). The denominator of unique dishes in would act as **Inverse Document Frequency** (IDF), reducing the score if the dish mention is not unique for the restaurant.

After ranking the dishes with this formula, following visualization was generated for restaurant rankings of top 50 restaurants for dish popular dish '**Chicken Curry**'

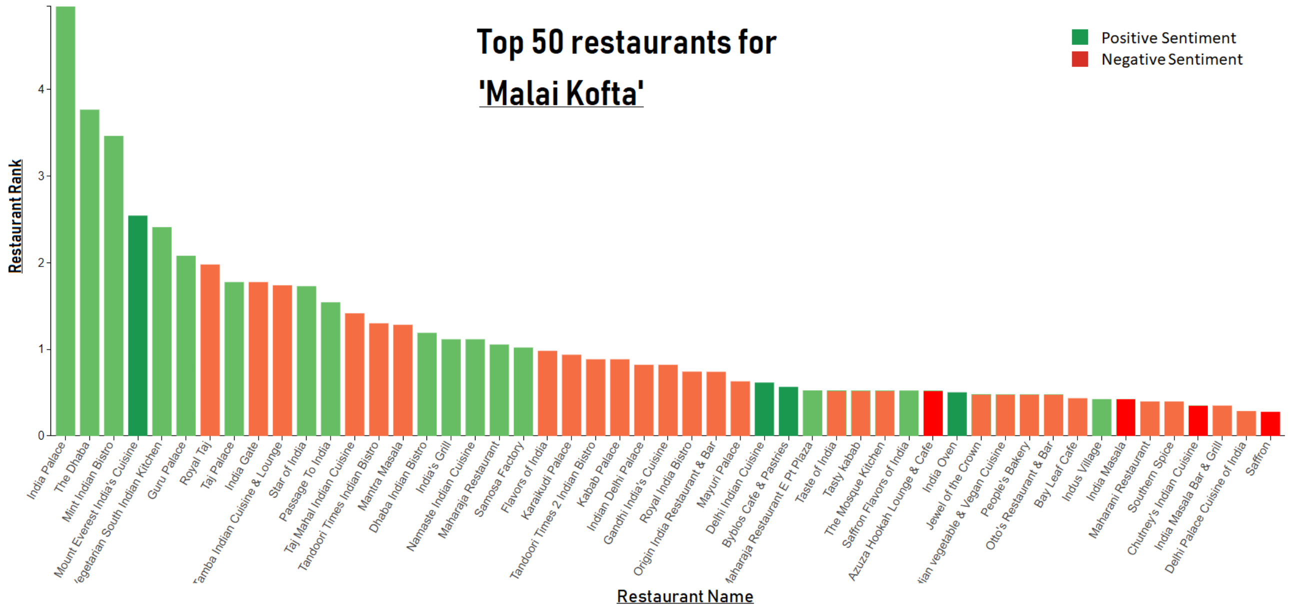


The greener shade represents a more positive sentiment towards the facility, while a darker shades of read shows a more negative overall sentiment of users towards this restaurant. One can see that overall, the top few restaurants for a popular dish is positive.

Another popular Indian dish in Europe and Americas is **'Chicken Tikka Masala'**, which generated the following chart:



The following chart is for dish called **'Malai Kofta'** which is not so popular according to our dish ranking. It's chart, as shown below, shows only a few high-ranking restaurants, while many restaurants have low rating and generally a negative overall sentiment rating, explaining the unpopularity of the dish.



4 Conclusions

It can be concluded that:

1. Standardization of dish names produced higher accuracy.
2. The ranking based only on occurrence of the entity in text reviews will not be high in accuracy.
3. A ranking criterion based in TF-IDF rating, both for dish and restaurant ranking resulted in better results.
4. Mostly positive reviews would mention more than one dish.
5. Generally, a few restaurants will specialize in some specific dishes.