**Total = 80 points**

* **Abstract [-5 if missing]**
* Add an updated abstract from the proposal version. In addition to a brief summary of the project, add a summary of the dataset description and analysis, and baseline results.
* Should not be more than 1 column.
* **Dataset description [35 Points]**

We use the dataset provided by Yelp on [Kaggle](https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp_academic_dataset_business.json) for training and testing the models. The dataset includes reviews of many different businesses of many cities across US and contains information about 209393 businesses. Out of which, we are using reviews only for the restaurants of city Avondale in Arizona. Original dataset contains 5 json files – business, review, user, check-in and tip. Our project related data present in business.json and review.json files. Hence, we use only those json files and filter the data thereafter.

* **Data preparation: [10 Points]**
* Source [2 points]
* Describe how you got the dataset (e.g., crawling, API, from a website, etc.), give proper references wherever applicable (-2 for missing references of papers, websites, or APIs)
* Data preprocessing steps and explanations [4 + 4 Points]
* Explain how the data preprocessing, cleaning, imputation, and other processing was done.
* Explain why this dataset is necessary and sufficient to achieve the goals of the project.
* Please note: the dataset should be fixed and finalized [-10 if not]
* Exception: if you are crawling the dataset yourself, then you can continue to crawl the data beyond the midterm, as long as a big fraction of the dataset has been collected already.

**Raw Data Statistics: Explain the dataset [10 points]**

After cleaning the data, Dataset includes five-star ratings as well as text reviews. There are total 12662 reviews (data points) available with 8031 unique users and 163 unique restaurants. These data were created between 2005 to 2018 for the restaurants in city Avondale.

Following are the attributes of the dataset: Review ID, User ID, Business ID, Restaurant Name, Latitude, Longitude, Stars, Text Reviews.

**Important and Features and their basic Statistics:**

Following is the plot of 20 most reviewed restaurants and their counts. ‘*Flavours of Louisiana’* is the most reviewed restaurant with more than 700 reviews.

A screenshot of a cell phone

Description automatically generated

Following is the analysis of the most reviewed restaurant through years. Interestingly, the restaurant is rated more than 4 stars all the years.

A picture containing drawing

Description automatically generated

Following is the plot of 20 top rated restaurants and their average ratings. ‘Colados Coffee & Crepes’ is the top-rated restaurant with around 4.7 average stars.

A screenshot of a social media post

Description automatically generated

Following figure shows the Wordcloud for the top rated restaurants’ reviews:

A picture containing food

Description automatically generated

From following graph, it is inferred that users are giving 5 ratings more than other ratings:

A picture containing device, umbrella

Description automatically generated

Rating Distribution satisfies the long tail property. Very small fraction of items is rated frequently. That’s why, those few items are popular and most other items are rated very rarely. We can see that roughly 40 restaurants are rated more than 100 times.

A screenshot of a cell phone

Description automatically generated

Distribution of Users giving ratings is similar to the distribution of ratings among restaurants and follows long-tail property. Very small fraction of Users is very actively giving ratings to the restaurants and majority of the users are not interested in rating restaurants.

A screenshot of a cell phone

Description automatically generated

**Data Analysis: Explore your data and talk about findings [15 points]**

Discuss relevant features, correlations, cluster visualizations, sentiment statistics, network statistics, centrality distributions from the data.

Add at least 5 insights or interpretations [5 \* 2 = 10 Points]

Add at least one visualization (figure or plot) [5 points]

All findings and figures should be quantifiable (-3 per non-quantifiable instance). Example: a word cloud is non-quantifiable.

-3 points for every non-relevant analysis below 5.

Please make sure to add only relevant visualizations and insights. Inserting vague plots and figures, for example, unrelated word cloud or generic network visualization will incur penalty (-5 points penalty).

* **Describe the Experimental Settings [10 Points]**
* Explain the parameters of the experiment: what is the split used, cross-validation setting, evaluation metrics (for example, accuracy, AUC, or precision etc.), system settings (RAM, GPU, or CPU statistics).
* -2 per missing relevant detail.
* **Baseline results and discussions [30 points]**
* The report should have results from at least 2 baselines. At least one baseline should be from a published paper or preprint. Creating one reasonable baseline yourself (e.g., using feature engineering and standard ML classifier) is allowed. No additional points will be awarded for having more than 2 baselines.
* **Baseline description [7.5\*2 = 15]**
* Describe the baseline. Give a short technical description of the baseline, along with its reference, provide details of the kernel or hyper-parameters used, provide links to code repository used [5 points]
* -2 per missing detail, reference, link
* Why is this baseline suitable for the problem? [2.5 Points]
* **Baseline result [10 points]**
* Results of the baseline on your dataset, presented in a table or figure (e.g., a bar chart) [5\*2 = 10]
* The baselines should be compared on the same metric [-5 if not]
* **Result discussion [5 points]**
* Compare the results of both the baselines. Why does one perform better than the other? If applicable, compare the result to the state-of-the-art reported in literature.
* **Next steps [5 Points]**
* Explain in detail your own proposed approach and what novelty or improvement you are adding over the baselines. [2.5 + 2.5 = 5 Points]
* -2 for unclear explanation
* For development projects, clearly describe what will be done by the final report and how exactly this will be achieved. Example, if you are creating an app, where do you plan to host it.
* **Contribution [-5 points if missing]**
* Distribution of team member effort. Can be as simple as "all team members have contributed a similar amount of effort" or this can be a more detailed description of effort breakdown. If effort distribution is too uneven, I may assign higher scores to members who have contributed more.