****

**DATA REQUIREMENTS AND ASSUMPTIONS**

**AIDI 1003 – CAPSTONE TERM 1**

**AMIT MARAJ**

**AHNCH BALA 100424062**

**SONAKSHI KARKERA 100720763**

**SURBHI THAKUR 100732335**

**ARUN KALAESWARAN 100771700**

**October 11, 2019**

# Overview

Our solution will be using a data set that has been provided by Yelp that they created for the Yelp Dataset Challenge. The dataset contains 6,685,900 reviews, 192,609 businesses, 200,000 pictures and 10 metropolitan areas. It is a large JSON file that is located on their website at this link; <https://www.yelp.com/dataset/download>.

## Data Requirements

In order for the model to be successful, the data has requirements that it needs to meet. The following are the requirements:

* Our data set should have the following features: *Business Name, City, Stars, Review Count, Is Open, Categories, Attributes, Hours of Operation, Review Date, Review Text, Review Star*. If any of the aforementioned features are missing we will discard the record.
* The “star” ratings must be segregated from 0-3 and then 4-5. This is done so that we have better control over the data to achieve more accurate results.
* We need to ensure that our data does not have any null or missing values. Records that include such values will be imputed or removed.
* The data must be preprocessed in order to extract meaningful content from each of the records. All uppercase letters will be converted to lowercase and punctuation marks will be removed.
* Our solution requires data from 2017, which has information regarding when a closed restaurant was last open. This can be deduced by locating the last known reviews of such restaurants as they would indicate a date of when the closed restaurant was last open.
* In order to have more restaurant attributes we’ll do feature extraction. Chain status will be determined by whether the number of restaurants with the same name was greater than five/ten in our data.
* We’ll extract restaurant type based on the category labels provided by Yelp (e.g., fast food, bars, pizza, Mexican etc.)
* The longevity of the restaurant will be determined by the latest and oldest date of the review and will be calculated by subtracting the two.

## Assumptions

* We are assuming Yelp’s reviews provided aren’t spam reviews and are actual reviews by individuals. It is an ongoing problem of reviews being created by bots, however, we will make the assumption all the reviews are intact.
* We also make the assumption that the restaurants that are open or closed as of January 19, 2019, are still open or closed as of today.
* We may have to make assumptions on the Categories feature, as some restaurants may be Fusion but can still be labelled as Thai food or Indian food.
* As the dataset provided by Yelp contains 6 JSON files- business, check-in, tip, photo, the user and review; For the scope of this project we will only be using the business and review JSON files. File sizes are quite high (8GB each), we may need to switch to cloud option for training our model.
* The businesses described in the Yelp dataset belong to different categories, such as restaurants, shopping, hotels, travel etc. We’ll just be focusing on restaurants, therefore, we do not use the rest of the data. Reviews of only those restaurants will be trimmed from the dataset.
* Restaurants that are open are financially stable since we do not have the bookkeeping to truly determine the financial state of a restaurant

## Data Requirements Not Met

There are requirements that our solution will not meet as well. The following is a list of constraints our solution faces:

* The dataset provided by Yelp contains data of 10 metropolitan areas, however, we will be considering only 3-5 metropolitan areas. In order to get an accurate prediction for places across the globe, we would need a more comprehensive data set.
* In order to accurately predict the closure of restaurants, we will be prioritizing reviews from January 2017 until January 2019. We have limited ourself to 2 years of data due to the immense size of the dataset provided by Yelp.
* The financial information of the restaurant in question is unattainable. We are unable to make our predictions more accurate due to this constraint.
* When we compare open and closed restaurants there is a class imbalance. The data for closed restaurants is less in comparison to the ones that are open. We will either reduce the amount of data for open restaurants or synthesize data for closed restaurants.