# Analysing Customer Restaurant Reviews on Yelp Website by using Aspect based Topic Modelling in Python

## Overview

Nowadays, people rely on reviews and ratings given by other customers before going to a particular restaurant to dine. For this reason, it is important that the businesses improve on various aspects of their quality by taking the customer reviews into account. This will not only help them improve but also expand their businesses over time. Today, many factors contribute towards customer's choice; especially reviews and rating of a particular business. In a review, customers write about various factors of the restaurant like ambience, service, overall experience, hygiene and so on. It is important for the businesses to improve on the aspects that the customers give negative reviews about so that they can propose a better customer experience and ultimately grow their businesses.

Since large amount of data is generated every day, finding information from them has become more challenging. We need new tools and techniques to find useful information from a seemingly never-ending amount of data. Effectively leveraging the text data can be of exponential importance when aiming to improve any type of businesses. Efficiently managing and exploring the text data is the main task in any text data analysis project. To extract relevant factors that contribute to the success of a business, advanced opinion mining techniques and topic modelling concepts are needed. Topic modelling technique allows for organising and understanding unstructured text data by discovering hidden topic patterns. Topic Modelling works by finding a collection of words that best represent a particular topic. For example, in restaurant reviews, the words like *flavour, tender, crispy soft, perfect, dry, salty* could refer to a topic *'taste'* in restaurant reviews. It would be useful to understand the relationship between the topics and the associated words to improve a business. For example, for restaurant reviews, if for a topic *'service'*, If there are many negative words like *loud, crowd, wait, slow*, then it can be inferred that that restaurant needs to improve their quality of service for better customer satisfaction.

Yelp is currently the most popular merchant information platform. It allows for customers to write reviews about different types of businesses and give the businesses a star rating. There is a great potential in the yelp dataset to provide valuable insights into different businesses. Our project aims to exploit the Yelp's reviews information to find different aspects of the businesses that the customers talk about and do a sentiment analysis on those aspects to identify the factors which can be improved upon by the restaurant businesses. Based on the results, we would also build a restaurant recommender system for customers depending upon the positive reviews and the location of the restaurant. We would be using the concepts of Topic Modelling, Word embedding and Natural Language Processing techniques for our analysis.

# Analysing Customer Restaurant Reviews on Yelp Website by using Aspect based Topic Modelling in Python

## **Problem Statement**

Identification of various aspects/factors that the restaurants have to improve in order to grow their business by finding business improvement and process improvement opportunities.

## **Datasets**

The Yelp Open Dataset is a subset of all Yelp businesses, reviews and user data made publicly available by Yelp. We would be working on this Yelp publicly available dataset which consists of the customer reviews along with other relevant details. The dataset is 5.79 gigabytes and in the form of json files. This dataset has information on reviews, check-in, users, businesses and tips. Our project will mainly focus on users, businesses, tips and reviews data. The link for the dataset is in the references section.

The business.json file contains information like business id, name, address, city, state, postal code, latitude, longitude, stars, review_count and is_open.

The user.json file contains information on user id, number of reviews that the user has written, time since the user has joined yelp and the average ratings that the user has given.

The reviews.json file contains the information on review id, the id of the user who has written the review, the id of the business for which the review is written, the review content and the star rating for that business given by the user.

The tip.json file contains user id, business id, text, date and compliment count.

Since, we are only interested in restaurants for the purpose of our project, we will filter out all the other type of businesses and since closed businesses would not contribute much to the analysis, we will filter out the closed restaurants too by using the is_open column in business.json file. For our analysis, we will convert the json files into csv files so that we can do nlp processing on it.