## 

**Yelp Dataset**

(Recommendation Engine & Sentiment Analysis)

Devesh Kandpal

Soham Das Neogi

Final Project: INFO 7390 Advances In Data Science

# **Overview**

Yelp.com is the go-to website for anybody looking for recommendations about local businesses. Over the period, it has become quite popular for restaurant recommendations as users find it easier to judge a local business by reading reviews about that business posted by other users. In addition to writing reviews and rating local businesses, user can react to reviews, plan events or start up a discussion forum. Yelp has about 135 million monthly customers and 95 million reviews.

# **Goals**

The Yelp dataset is open sourced, and it has enough information about each local business, reviews about those businesses, tips written by the user about a local business and photos uploaded by the user about the local business. This dataset can help us summarize useful information about most popular and unpopular restaurants classified by food genre and user reviews. For each user we can see his social connections and understand how much of an influencer can he be on other users. Lastly, we have an entire dataset comprising of images uploaded by different users with manually class tagging done on each image.

Our goal is handle the following use-cases:

1. Using user profile and his/her taste, we would recommend restaurants using his/her geospatial information in conjunction with collaborative filtering. Additionally, we would perform sentimental analysis on the recommended restaurants reviews.
2. Future scope:

Using image dataset comprising of labeled data, we would like to create a Deep Learning Model that is capable of recognizing objects and giving them an appropriate label.

# **Use Cases**

1. **Everyday User**

* Using geospatial information in conjunction with results from collaborative filtering we would recommend restaurants to users.   
  When the user clicks on a given recommended restaurant, the results from sentiment analysis on reviews would be showcased using appropriate graphical representation.

Methodology: We would be considering the following techniques while developing the recommendation system

1. **Content Based Recommendation**
2. **Collaborative Filtering**
3. **Matrix Factorization Techniques**

* For Sentiment Analysis, we would be using the textual context from the yelp dataset to find positive and negative words using SVM model. We will be calculating overall polarity score (which would tell us how essential a word is among all the reviews) using word frequency and score calculated from SVM model.

Reference: <https://arxiv.org/ftp/arxiv/papers/1709/1709.08698.pdf>

1. **Local Businesses (Future Scope)**

* Our predictive model can help us automatically tag images into specific categories, thus avoiding human intervention.  
  Example: Food, Outside, Inside, Drink, Menu  
  Methodology: We would be developing a **CNN based Deep learning model**

# **Data**

1. **Yelp Dataset**:<http://www.yelp.com/dataset>

We will work with the Yelp Dataset available online.

# 

# **Process Outline**

1. Data Preprocessing

● Data Cleaning, handling missing values

● Joining Business.json, Reviews.json, User.json, Tips.json and Image.json

2. Exploratory Data Analysis

3. Understanding how to build a recommendation engine, we would possibly consider exploring Apache Spark (Spark MLib) while developing.

4. Work on implementing Sentiment Analysis on text reviews

5. Build an automation pipeline using Luigi for data ingestion and cleaning.

6. Deploy the Model on AWS or Google Cloud Computing Platform

7. Build a web application to demonstrate the image classification and recommendation system.

# **Architectural Diagram**

**Google Cloud Platform**

Data Engineering/ML pipeline (Luigi/Python/Spark)

Spring MVC + Tomcat

# 

Rest API

MongoDB

Flask Application (Image Classification/Recommendation)

File System (saved models + Dataset)

# **Milestones**

|  |  |
| --- | --- |
| **Timeframe** | **Delivery** |
| Day 1-3 | Data Preprocessing and Exploratory Data Analysis |
| Day 4-9 | Model Building, Training (Recommendation Engine and Sentiment Analysis on recviews) |
| Day 10 - 12 | Deployment of models on cloud and build web application |
| Day 13-14 | System integration and documentation |

## 

# **Deployment / Infrastructure Details**

1) Language: Python, Java

2) Pipeline: Luigi

3) Cloud Tools/Platforms: AWS / Google Cloud Platform

# **Reference Sources**

<https://www.yelp.com/dataset/documentation/json>

<https://arxiv.org/ftp/arxiv/papers/1709/1709.08698.pdf>