****

**Yelp Rating and Recommendations**

**Team 3:**

**Aldred Lau**

**Ng Hui Ying**

**Michelle Teo**

**Tan Yi Hao**

**1. Background**

Yelp is a ratings and recommendation website for businesses to connect with their users. Their main source of revenue comes from paid advertisements from businesses. To expand their business, they wanted to monetize the analytics that comes with the vast amounts data on Yelp. With that, they opened a new department for recommending solutions to businesses using analytics. Our group aims to make use of the given dataset to derive possible actionable solutions for businesses.

**2. Problem statements**

**At Yelp, we want our users to be matched with the best business, providing satisfied customers. Today, we are only able to promote the best matched business of choice to the users. However, this might not match the user’s expectations but only be the best of all existing choices. Therefore, we will use analysis in the area of text and sentiment to identify areas of improvement for businesses to be a better match to customer expectations. As customer expectations would be better met, we are hopeful that they would influence their friends to be a supporter of Yelp as well.**

**Primary Objective: Text and sentiment analysis of Yelp business reviews**

With the text and sentiment analysis, we want to provide recommendations to the existing and new businesses, specifically:

1. How can existing businesses leverage on the reviews and tips provided by the Yelp community to improve existing attributes of their business (e.g. good service, but can improve in terms of pricing)
2. Based on comparison to reviews and tips of competitors within the same business category, how can businesses better suit the preferences and tastes of users?
   1. What is the distinction between popular businesses versus less popular businesses? What are the attributes value (i.e. Happy Hour: True) that cause a business to become popular? What are the attributes value (i.e. Alcohol: None) that cause a business to become unpopular? (Generally, popular businesses have many check-ins and minimum business rating of 4 while unpopular businesses have few check-ins and max business rating of 2)
   2. What are the differences between the preferred attributes for different business categories / locations?
3. How can new businesses understand the trends, tastes and preferences of the country/state/city/ neighborhood to tailor their business propositions to best fit the preferences of users?
   1. What are the number of businesses under a unique category within a specific location? Is there a saturation of specific businesses? Is there a preference for specific operations within the area?

**Secondary Objective: Visualization of business attributes and ratings**

1. Visualization of attributes / services that people look out for in businesses for specific country, state and city and business category. The suitable visualizations will be word cloud and bubble chart.
2. Distribution and Trend analysis of business ratings over time. The suitable visualization will be time series line chart.

**Tertiary Objective: Clustering of users based on the similar Yelp behavior**

1. How can Yelp leverage on the reviews, check-ins and tips provided by users to provide recommendations for users on places to visit?
   1. How can we cluster similar users together? What are the important data attributes to consider?

**3. Exploratory Data Analysis (EDA)**

Our Exploratory Data Analysis (EDA) first started with the cleaning of the data followed by the univariate analysis of the data attributes. Below are some of our findings after the initial EDA process:

1. There are 61184 unique business with 789 different business categories
2. The business can belong to either zero, one or multiple categories
3. The business location has the following granularity: (Country - State - City - Neighborhood - Address)
4. The businesses are located in 4 countries across 10 cities.
5. The business can belong to either zero, one or multiple neighborhoods
6. There are 2 types of Yelp users namely the elite user and the normal user
7. There is a timestamp associated with each business review given by a user
8. The business review has a time range between the Year 2004 and the Year 2015. (Yelp was founded in Oct 2004)
9. The user can give optional tip by choosing the relevant checkboxes after giving a business review
10. The users can check in using Yelp when they visit the business
11. There are 2 types of ratings namely the rating for user and rating for business
12. The Yelp community can rate the user’s business review by giving cool, useful and funny votes

**4. Approach**

Our main approach would be using text and sentiment analysis on the business reviews and tips by customers. These would be analyzed at according to:

1. User ratings for businesses - High (4-5 stars) and Low raters (1-2 stars)
2. Business Geography - Country, States, City and Neighborhood
3. Business Categories - Restaurants, Retail, Beer etc.

We will be using the Named Entity Approach (NER) while leveraging on existing standard Neuro-Linguistic Programming (NLP) tools to build a system which can provide more tailored and specific recommendations to businesses.

**Business Reviews given by users**

This is the main field on which we would be using text and sentiment analysis. The aim of which is to derive certain attributes of businesses that customers find important for a specific business.

**Business Ratings given by users**

We understand that each user is unique and has a separate mentality when providing reviews. Some people tend to focus on the positive aspects of their experience with the business while others may tend to pick out the negative aspects. This would then translate into good or bad ratings. Therefore, we would like to separate the users in high and low raters based on the average or median or mode rating which users provide to **compare the relations** between the top 3 desired attributes. Ideally, there should be a normal distribution of ratings provided by users in order to obtain an objective response from the users. Hence, we would want to eliminate any overly ‘optimistic’ or ‘pessimistic’ users who may affect the accuracy of ratings. We would then explore if there are correlations so that a ranking system can be designed to provide recommendations.

**Business Tips (Optional comment) given by users**

After exploring the review system provided by Yelp, we hypothesize that users who provide tips have generally had a good experience with the business and hence would provide a higher rating. This is because the tip is an optional comment after providing a review. Hence, we would like to explore if there is a positive correlation between the number of tips and the ratings. Following on, we would try to identify key attributes which would lead to tips which might provide a higher weightage when clustering certain attributes. They would then influence the recommender system provided to businesses.

**Business Geography**

We would also separate businesses into **countries** to conduct a text and sentiment analysis on the reviews to identify important attributes most users are concerned about across all businesses (e.g. good service) which would tend translate into ratings, to make the the insight relevant to the country of the business. Also, we would like to understand if there are unique abbreviations or colloquial specific to the states in order to conduct clustering to group similar words together (e.g. tomorrow, tmrw,2mr). This is because within different countries, users tend to display different forms of speech and based on the large distribution of user profiles on Yelp, there might be significant differences in the figurative speech in reviews. This would help to improve the accuracy of the NER tool.

We would also attempt to conduct more in-depth analysis with a similar approach based on the different **states**, **cities** and **neighborhoods** to further improve the accuracy of our recommendations.

**Business Categories**

After our EDA, we discovered that some businesses fall under multiple categories. (e.g. food, beer, restaurant) Hence, we will attempt to re-classify these businesses under an umbrella category in order to refine our analyses and recommendations. Similarly, we will then split the businesses based on their geographic locations to understand the specific attributes which are desired by users for a particular business category to provide a ranking and preference system based on the top 3 attributes. (e.g. 1. carpark 2. location 3. service) This would help existing and potential businesses understand the profile of users within the states and cities and determine

the best fit for their business model.

**5. Possible Tools:**

1. Data Cleaning:
   1. Excel VBA
   2. R
   3. Python
   4. SAS EG
   5. RapidMiner
2. General Analysis:
   1. R
   2. Python
   3. Graphlab
3. Text Analysis:
   1. Semantria
   2. TAPoR
   3. RapidMiner - Text miner function
   4. Python - Natural language toolkit (NLTK) library
4. Visualisation:
   1. Tableau
   2. D3.js
   3. Python - Wordcloud and Matplotlib libraries

**6. Schedule of project**

|  |  |
| --- | --- |
| Week 3 | Data Cleaning  EDA  Understand and Evaluate the Literature Reviews  Proposal submission |
| Week 4 | Data Cleaning  EDA  Understand and Evaluate the Literature Reviews  Exploration of software tools i.e. SAS, Tableau, D3, Python, Text Miner for text and sentiment analysis and visualization |
| Week 5 | Selection of Literature Reviews  Initial Text and Sentiment Analysis  Initial Visualization for Text and Sentiment Analysis  Initial Key Findings and Insights for Text and Sentiment Analysis  Initial Business Rating over Time Analysis  Initial Visualization for Business Rating over Time Analysis  Initial Key Findings and Insight for Business Rating over Time Analysis |
| Week 6 | Revised the selection of Literature Reviews if necessary  Revised Text and Sentiment Analysis  Revised Visualization for Text and Sentiment Analysis  Revised Key Findings and Insights for Text and Sentiment Analysis  Mid-Term Preparation  Wiki Updates |
| Week 7 | Mid-Term Presentation |
| Week 8 | Mid-Term Review for Text and Sentiment Analysis  Mid-Term Review for Text and Sentiment Visualization  Mid-Term Review for Text and Sentiment Key Findings and Insights  Mid-Term Review for Business Rating over Time Analysis  Mid-Term Review for Business Rating over Time Visualization  Mid-Term Review for Business Rating over Time Key Findings and Insights  Revised Text and Sentiment Analysis  Revised Visualization for Text and Sentiment Analysis  Revised Key Findings and Insights for Text and Sentiment Analysis  Revised Business Rating over Time Analysis  Revised Visualization for Business over Time Rating  Revised Key Findings and Insights for Business over Time Rating |
| Week 9 | Initial Clustering Analysis  Initial Visualization for Clustering Analysis  Initial Key Findings and Insights for Clustering Analysis |
| Week 10 | Revised Clustering Analysis  Revised Visualization for Clustering Analysis  Revised Key Findings and Insights for Clustering Analysis |
| Week 11 | Final Revision of all the Analysis  Final Revision of all the Data Visualization  Final Revision of all the Key Findings and Insights |
| Week 12 | Final Presentation Preparation  Wiki Updates |
| Week 13 | Final Presentation  Final Report Preparation  Poster Preparation  Wiki Updates |
| Week 14 | Final Report Submission  Poster Submission  Wiki Submission |

**7. Roles and Responsibilities**

The tasks of data cleaning, EDA, exploration of tools, literature reviews, reports and presentations are to be shared responsibility amongst team.

The main roles would be:

Ng Hui Ying - Project Manager: in charge of the project schedule, poster and wiki updates

Aldred Lau - Data Analyst: in charge of ensuring the accurate generation of text and sentiment analysis / business rating over time analysis and coming up with suitable visualizations

Michelle Teo - Data Analyst: in charge of ensuring the accurate generation of text and sentiment analysis / user clustering analysis and coming up with suitable visualizations

Tan Yi Hao - Business Analyst: in charge of understanding the key business problems faced by Yelp and providing feasible business solutions based on the key insights provided by the data analyst

Despite the roles to ensure the driving of the important parts of the project, the tasks would be evenly split out among the members of the group to ensure a fair learning environment.

**8. Tentative final deliverable**

1. Report
2. PowerPoint
3. Poster
4. Dashboard

**9. Data dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| Businesses | | | |
| Data Field | Components | Description |  |
| type |  | business type |  |
| business\_id |  | encrypted id of business |  |
| name |  | name of business |  |
| neighborhoods |  | name of neighborhood |  |
| full\_addres |  | localized address |  |
| city |  | city name |  |
| state |  | state name |  |
| latitude |  | latitude |  |
| longitude |  | longitude |  |
| stars |  | star rating (rounded to half-stars) |  |
| review\_count |  | review count |  |
| categories |  | localized category name(s) |  |
| open | true/false (boolean) | if shop is opened or closed |  |
| hours | (day\_of\_week): {  'open': (HH:MM),  'close': (HH:MM)  } | opening hours |  |
| attributes | (attribute\_name): (attribute\_value), |  |  |