Recommendation System Optimization for Businesses within the Yelp Crowd-Sourced Reviews Sharing System

Amr Maraqa (261101609)

Kenza Sqalli (260907188)

Tarek Cheaito (260886446)

Xénia Sozonoff (260901022)

Desautels Faculty of Management

McGill University

Montréal, Québec, Canada

November 4th, 2022

A Team Project for INSY662 – Data Mining and Visualization (2022 Fall)

I. Introduction: Yelp

Yelp.com is a crowd-sourced business review web and mobile platform. Its main purpose is to link the consumer to businesses. Yelp users have access to the businesses registered on the platform via the designated business page. There, users can acquire more information about the business, like working hours, available amenities, telephone numbers, business website URLs, etc. and submit reviews about their experiences of the services or products offered by the business on a scale of one to five. Moreover, every Yelp business page contains different interactive features, such as a redirecting link to the business location on Google Maps, pictures and videos of the business services or products, and recommendations of other businesses similar or in the proximity of the one on the current page.

II. Project Conception:

The idea behind the project:

Following analysis of the products offered by Yelp, we imagined a way of improving the recommendation system of companies found on Yelp for users spending time on their website/application. Our team would like to suggest creating a recommendation system based on the activity of the users' friends. People's tastes/desires tend to coincide with those of the people with whom they surround themselves. For instance, one would be less tempted to try a restaurant if their friends reviewed it as being awful. On the other hand, that same person would be interested in trying a hotel/hair salon/boutique that a friend would have found fantastic with excellent service. This would serve users better, as they would start getting advertisements that present higher probabilities of being aligned with their preferences. Furthermore, businesses would benefit from increased conversion rates that result from these friend-based advertisements.

Potential benefits for Yelp:

A recommendation system in this case would benefit all stakeholders, which include Yelp users as well as businesses found on the platform. This would, consequently, lead to benefiting the organization, Yelp, in a broader sense.

In the world we live in, many people are influenced by their friends' personalities and actions. As a user scrolling through Yelp, one would now be able to see which businesses are worth experiencing and which are not, according to their friends' reviews. Instead of wandering on the platform cluelessly, one would be able to access his/her friends' positive reviews (of a business) on the category in which he/she is interested. This would completely narrow down the options to a couple of businesses, rather than hundreds. Not only do users benefit from this by saving time and energy in the decision-making process, but they, generally, start receiving advertisements that are better aligned with their interests.

As a result of our recommendation system, more people across the globe would engage on Yelp's web and mobile platforms. Many would be willing to explore different businesses simply because their friends expressed their content with the service or product. This would benefit the businesses on Yelp, as it increases the probability of users becoming their customers. Due to this rise in user conversion rates, Yelp would gain the favor of potential and existing business partners from which increased retention rates and ad revenue ensues.

III. Project Implementation:

Data Manipulation

Any subsequent text in this report will form an explanation regarding the manipulation of data from Yelp that will take place using Python on Jupiter Notebook.

Data Gathering

The first step towards the implementation of the project is to gather the data necessary for its accomplishment. As described earlier, the factors to be targeted in the given situation are the users and the reviews on the platform. Therefore, we decided to use the following two CSV documents: 'yelp_user.csv' and 'yelp_review.csv'. To facilitate the explanation that will follow, we will call the first table 'Users', and the second 'Reviews'.

- **Step 1: Remove all users without 'friends' on the platform**:

  The objective being to only gather users who have friends on the platform, we start by creating a copy of the 'Users' table from which is deleted any user that has the value 'None' in the 'friends' column. The code formulating this constraint is as follows:

```
ufrds = user[user.friends != 'None']
```

- **Step 2: Merge the previous query with the 'Reviews' table**

  Now that we know which users we are interested in for our project, we only want to keep these in the 'Reviews' table in order to analyze their reviews, and thus recommend (or not) certain institutions to their friends. To do this, we start by importing the Pandas library under the acronym 'pd', then we use it in the following code to merge the two tables:

```
urev = pd.merge(ufrds, reviews,  how = 'inner', on = 'user_id')
```

We now have the table **'urev'** with the desired information. Due to the limitations of the computation power of our devices, we could not apply natural language processing to the entirety of the 'urev' data frame. Therefore, we extracted the first five entries in 'urev' and stored them in a new data frame called 'urev2'.

Data Preprocessing: (see attached code file for more details)

We now want to clean the text of each review. The goal is to keep the minimum number of characters in order to facilitate text analysis. To do so, we break down the sentences in the reviews into a collection of only words that add meaning to the message intended by the user and a uniform format.

- **Step 1**: We began by deleting english stopwords in the reviews. To accomplish this step, we imported the Natural Language Toolkit library (nltk) into Jupyter NoteBook. Once all the works that don't contribute to the meaning of the sentence we removed, the resulting text was stored in a new **'urev2'** column called 'review_without_stopwords'.
- **Step 2**: Thereafter, we removed punctuation. Since punctuation is only linguistically structural, we eliminated it to simplify the processing of the text.
- **Step 3:** Finally, we set the text to the lower-case format to eliminate the difference between the same words. i.e., 'Good' and 'good.'

Data Analysis & Implementation:

Now that our text is ready for use, we can analyze it.

- **Step 1:** Tokenization and Lemmatization

Tokenization is the construction of a list of the words that constitute the sentence. Once the preprocessed text is tokenized, it undergoes the process of lemmatization, where each word is associated with a predefined set of synonyms based on its context.

- **Step 2:** Polarity

We now need to polarize each one of our reviews. To do so, we will attribute to each review a score that will range between -1 and 1. The score will depend on the average amount of 'good' and 'bad' words found in each review. If the score is in [-1, 0], the reviews is classified as negative, and the restaurant/hotel/salon will not be recommended to the friends of the user who left that review. On the other hand, if the score is in [0, 1], the institution reviewed will be recommended to the friends of the reviewer.

IV. Post Implementation Strategy:

This strategy can be improved by upgrading the available data about relationships between users. The need for this additional data came to our attention when we filtered the 'users' data frame to only include users who have friends. The total number of entries dropped from 1,326,100 to 760,007. This means that our strategy could only be applied to about half of the registered users.

Our suggestion is to add data about the relations between users to a higher degree. This means that not only would the direct friends of a user be known, but bigger relational networks that extend to friends-of-friends (second degree) and higher would be available for analysis as well. Through the presence of this data, users who do not have friends on Yelp's platform can still be included in the recommendation system's algorithm and benefit from its advanced target marketing techniques. Additionally, such data would aid Yelp in capitalizing on a larger pool of users in this strategy, which would increase its advertisement reach rates. Thus, with the extension of the strategy to include this data, Yelp becomes more auspicious for personalized user recommendations, which in turn increases its user base, and becomes more attractive to potential and existing business partners who are looking to sponsor their listings.