B565 Spring 2018: Assignment 03

Due on Saturday, March 03, 11:59 PM

Dr. Predrag Radivojac

Arnav(aarnav), Dheeraj(dhsingh), Pulkit(maloop)

Saturday, March 03

Contents

1: Project Title

Yelp Dataset Challenge: Analyzing Restaurant, Item Trends and Impact of Influencers

2: Team members

(a) Arnav Arnav (aarnav@iu.edu)

(b) Pulkit Maloo (maloop@iu.edu)

(c) Dheeraj Singh (dhsingh@iu.edu)

3: Objective and Significance

Describe what the goal of the project is, why is it important, and your motivation for doing it.

Yelp is a company which provides crowd-sourced reviews about local businesses. The company also trains small businesses on how to respond to reviews and provides data about businesses. Yelp has open-sourced its data to discover hidden insights that lie in the data. This big dataset consists of information about business, users, reviews, pictures, check-ins and other information across various metropolitan areas.

Objective: The objective of this project is to find insightful trends in the Yelp dataset that shows us how the popularity of certain business changes over time and what factors mainly affects those trends. In particular, our focus is on identifying how an influencer's positive or negative review on a restaurant affects its popularity. In addition, we also aim to build a model to find the top recommended items of restaurants using user reviews.

Importance: Analyzing and predicting an influencers impact on a business can guide the business owner manage their business more effectively and lead to a higher revenue. Further, the current recommendation system of Yelp does not show each restaurant's highly rated food items by users and helping businesses find these items as well as recommending these items to users can help drive their business to great success.

Motivation: *Influencer Marketing is not Advertising.* Our motivation for the project comes from the fact that people engage with and react better to people they know, trust or admire due to their expertise or passion for a product, service or topic. There has also been criticism that businesses pay influencers to highly rate their restaurant to cause a boost in their popularity. Thus, influencers play an important role in the success of a business and through this project, we plan to capture if a business' popularity is in fact impacted by these so-called influencers by performing exploratory analysis of Yelp dataset and building models to capture this behavior.

4: Background

(a) Introduce all important concepts and background information.

Yelp Data Challenge: The project uses data from the Yelp Data Challenge [1]. The challenge encourages students to use Yelp's open source data and find important insights that help the company improve its services. The Data challenge is a semester-long challenge and has completed 10 iterations before January 2018. Every iteration of the challenge results in various interesting research papers that are published by the participants. We are participating in the 11^{th} Yelp Data Challenge.

Influencer: There has been active research in identifying Influential users on social networks who have the capability of influencing users' choices. An Influencer can be identified by centrality measures computed on networks and other properties such as review counts, and user fan base [2].

Sentiment Analysis: The project relies heavily on *natural language processing*, and text-mining techniques, such as keyword identification and sentiment analysis, to analyze reviews and score them based on the sentiment in the text. Sentiment analysis is the task of finding different types of emotions associated with a given text or speech document. This can be done in traditional ways that look at words associated with positive and negative sentiments, and score a document based on these words [3]. This technique, however, can be fooled to result in a negative document having a positive sentiment score since it looks at individual words. Deep learning has been used to train deep neural networks, that build on top of grammatical structures and can capture the context in a document [4].

Time-Series Analysis: To explore trends in data, we need to extract data from the database and sort it by time. *Time series analysis* aims to find non-random patterns in data, where successive entries in the data are taken one after another at a uniform time step. It also aims to forecast how these trends are likely to change in the future with time [5].

Hypothesis Testing: The results obtained need to be evaluated, to determine whether there is an impact of influencers on popularity or the observations in the data may have occurred due to randomness. Hypothesis Testing is a technique that helps in determining whether a relationship between two datasets, defined by a hypothesis is statistically significant [6].

(b) Search the literature and describe previous work on this problem.

- Michael Luca [7] studies online consumer reviews effect on markets for goods. He presented findings of how increase in Yelp rating leads to rise in revenue of individual restaurants by studying the change in restaurant's demand affected by consumer reviews
- Peng et al. [8] identifies key influencer in a social media environment for business marketing utilizing topic modeling and social diffusion analysis
- A post by Influencive [9] studies the impact of Influential Marketing and what makes user a great influencer. Also, there is discussion about how brands exploits Influential Marketing in [10]
- There has been an extensive research in the field of text-based sentiment analysis. Yun Xu et al. [11] provided numeric score (star rating) using text reviews on Yelp using multiple supervised learning algorithms, compared their performances and concluded that in their context, binarized Naive Bayes combined with feature selection with removed stop words and stemming gives best performance. In an another admiring work, Lei Zhang et al. [12] lists different deep learning techniques implemented for sentiment analysis

(c) If there exists previous work on the problem, describe what makes your work particularly interesting.

We see that there has been research and projects on predicting restaurant ratings based on sentiment analysis of user reviews on Yelp. A lot of recent research has gone into clustering users, finding communities and finding influencers in various topics on social media platforms using graph analysis techniques, and detecting fake reviewers. Interesting research has gone into finding how events and offers on different social media platforms (such as Groupon) are related to Yelp reviews about a business. Research has been done to understand the impact of reviews and ratings on business revenue on Yelp.

This project aims to find whether the reviews of influencers on Yelp impact the popularity of a restaurant based defined based on the overall sentiment about the business, which has not been studied directly in the existing works.

5: Proposed Approach

(a) **Data Description:**

Yelp has an ongoing challenge[1] and has made its complete dataset public. This dataset consists of **6.52 GB** JSON or equivalently **7.55 GB** SQL and **7.47 GB** of photos. It comprises of 5,200,000 reviews, 174,000 businesses, 200,000 pictures across 11 metropolitan cities. Our work is focused primarily on the restaurant data as major portion of data comprises restaurant business. The complete schema of the data set can be summarized as below:

- *business*: Information about multiple business units on Yelp platform: restaurants, dentist, salons, etc. This information consists of ratings, location, attributes, open hours, review counts, etc.
- *review*: Text reviews, star-ratings, and other votes along with time stamp of a business unit given by a user
- *user*: User information such as name, review count, number of fans, friends, and other meta-data
- *checkin*: Check-in time on different days of week for a business
- *tip*: This includes information about tips written by users on a business unit and how many likes it received
- *photos*: Information about uploaded photos such as labels, captions, etc.

(b) **Method and Implementation:**

- Data Preparation:
  The data is available as a set of JSON files and also as a dump of a relational database. To setup the database we use PostgreSQL to load the database from the database dump. Postgresql is an open source object relational database which allows writing queries in SQL efficiently. We can then extract relevant data from the database (such as grouping restaurants by cities) by using proper queries and store it in new tables to avoid performing same queries again in the future. If queries on the extracted data still take a long time, we can move to spark, and store the database as a Spark Resilient Distributed Database (RDD) to allow parallel lookups in the data.

- Data Exploration:
  We perform exploratory data analysis on the data to get a better understanding of the dataset resulting in some useful insights. We start by looking at a number of restaurants per city, the average number of reviews for restaurants in every city, finding the type (category) of restaurants per city and more. Further, we need to identify top trending restaurants in a given city and top trending items for a given restaurant.

  - *Top trending restaurants*: We would perform sentiment analysis on users' reviews to provide a sentiment score to restaurants. This can be done in many ways: traditional approaches as well as state-of-the-art deep learning methods[4]. We would implement both techniques and keep the one with better results. Then for a given city with a maximum number of restaurants (Las Vegas), we would identify top 10 trending restaurants using a model which takes sentiment scores and average ratings as features.

  - *Top trending items*: Every restaurant has its own specialty for which it is known across a region. We aim to find restaurant-wise top trending items from customer's point of view. For every restaurant, we would consider their reviews and can find item keywords (Nouns) using

part of speech tagging and then we can find an associated adjective with that noun. Then we would a give score to the associated adjective depending upon its nature, for example, 'good pizza': 'good' (adjective) would have a positive score associated with pizza (noun), similarly, 'bad pizza' would have a negative score for a pizza here. Iteratively performing this across all reviews for a given restaurant, we can find items with high scores and those would be top trending items for that restaurant.

- Impact of influencer's review on a restaurant popularity:
  Yelp is a social media platform where people can explore businesses and more specifically restaurants to learn from experiences of others. Influencers are people who are very active on the platform and keep reviewing restaurants regularly. We aim to explore whether these reviews are impacting the popularity of restaurants in real life. To start with, we would analyze a trend of change in positive sentiment score for our top trending restaurants over a time period. In the next step, we would identify influencers from the user data. Influencers are users who have a high number of useful reviews, lots of friends, regularly post reviews and photos, high likes on tips written, are 'elite' users on the platform. Using these metrics we can score every user and subset those with relatively very high scores as influencers. Then for a given top trending restaurant, we would find out when these influencers reviewed (if any) that restaurant.

  Once we have these informations, we can explore if there is a significant difference in the popularity before and after influencer's review. Using this information, we can perform a hypothesis test to find out if there is a significant change in the popularity of a restaurant before and after an influencer's review.

  If we find that there is an impact of influencers on restaurant's popularity and we find out if the impact is significant, we would try to build a prediction model that would predict how the popularity of a given restaurant would be impacted given an influencer has written a review in future.

(c) **Evaluation Strategy**:

- *Hypothesis Testing*: To conclude with certainty that the reviews of influencers have an impact on the popularity of a restaurant, we set up a hypothesis test in the following way:

  For a given restaurant, we find the influencers who have given positive reviews for a restaurant. Then, we find the average difference in the change in the popularity of the restaurant over a fixed time window, before and after the reviews from different influencers. We then perform a right-tailed test on the null hypothesis that the positive reviews do not have an impact that is, there is no change in average sentiments before and after the reviews on average, against the alternative hypothesis that the restaurant's popularity increases after a positive review from the influencer, that is there is a positive change in the sentiments for a restaurant after the review.
  We do a similar left-tailed tests for negative reviews by the influencers.

- *Future Popularity Predictor*: To evaluate the future popularity predictor, we split the data into training and test sets. We perform cross-validation on the training data to estimate the accuracy and then test this on the test data. To split the data, we need to be careful as it is a time series data, and we should not randomly choose non-consecutive points in the data.

(d) **Expected outcome:**

- We expect to see a significant difference in popularity trends of top rated restaurants from those that are rated lower and have low sentiment score.
- We expect that trending items for a restaurant, from our predictions should make an intuitive sense, for example, a pizza restaurant should have a userbase loved pizza such as Cheese Pizza as one of the top trending items, etc.
- We expect to see a rise in the percentage change of positive reviews if influencers have reviewed positively and drop otherwise, for top trending restaurants, meaning that influencers do have an impact on a restaurant's popularity.

(e) **Fallbacks:**

- If there is little or no impact of influencers reviews on the popularity of the restaurant, we perform further analysis to try to find other factors from the data that impact the popularity of a restaurant.
- In case of failure in prediction of the future trends based on reviews of influencers with acceptable accuracy, we use the earlier trending items found as a result of text mining to build a recommendation system that suggests restaurants to users based on their search and corresponding trending items

6: Individual tasks

In this section we define what tasks each of the team members take on. These tasks are subtasks of those defined in the proposed approach section.

- Database Setup and Data Preparation : Arnav
- Data Preprocessing and Exploration : Arnav and Pulkit
- Sentiment Analysis of Reviews : Dheeraj
- Exploratory Analysis and Prediction of Trending Restaurants : Dheeraj and Pulkit
- Prediction of Trending Items of Restaurants : Arnav and Pulkit
- Identifying Influencers : Dheeraj
- Hypothesis testing for Influencers's Impact : Arnav
- Future Trend Prediction model : Pulkit
- Further Analysis : Arnav, Dheeraj and Pulkit

These assignments are tentative and may change as new subtasks are identified. These will be mentioned in detail in the final report.

References

[1] Yelp. Yelp dataset challenge. Yelp Dataset Challenge Website, January 2018.

[2] Ermelinda Oro, Clara Pizzuti, Nicola Procopio, and Massimo Ruffolo. Detecting topic authoritative social media users: a multilayer network approach. *IEEE Transactions on Multimedia*, 2017.

[3] Stanford NLP. Stanford core nlp - natural language software. stanford nlp website.

[4] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[5] statsoft.com. How to identify patterns in time series data: Time series analysis. statsoft.com website, 2018.

[6] Wikipedia contributors. Statistical hypothesis testing — wikipedia, the free encyclopedia, 2018. [Online; accessed 4-March-2018].

[7] Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. 2016.

[8] Wei Peng and Tong Sun. Method and system for identifying a key influencer in social media utilizing topic modeling and social diffusion analysis, November 13 2012. US Patent 8,312,056.

[9] Pavan Belagatti. The importance and impact of influencer marketing in 2017. Influencive Webcite, 2017.

[10] Romy. What is the impact of social media influencers? Digital Me Up Website, April 2017.

[11] Yun Xu, Xinhui Wu, and Qinxia Wang. Sentiment analysis of yelps ratings based on text reviews, 2015.

[12] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *arXiv preprint arXiv:1801.07883*, 2018.