Determine Business Effectiveness Based on User Reviews And Ratings

Chirag Salian, Latish Khubnani, Shivkumar Dudhani

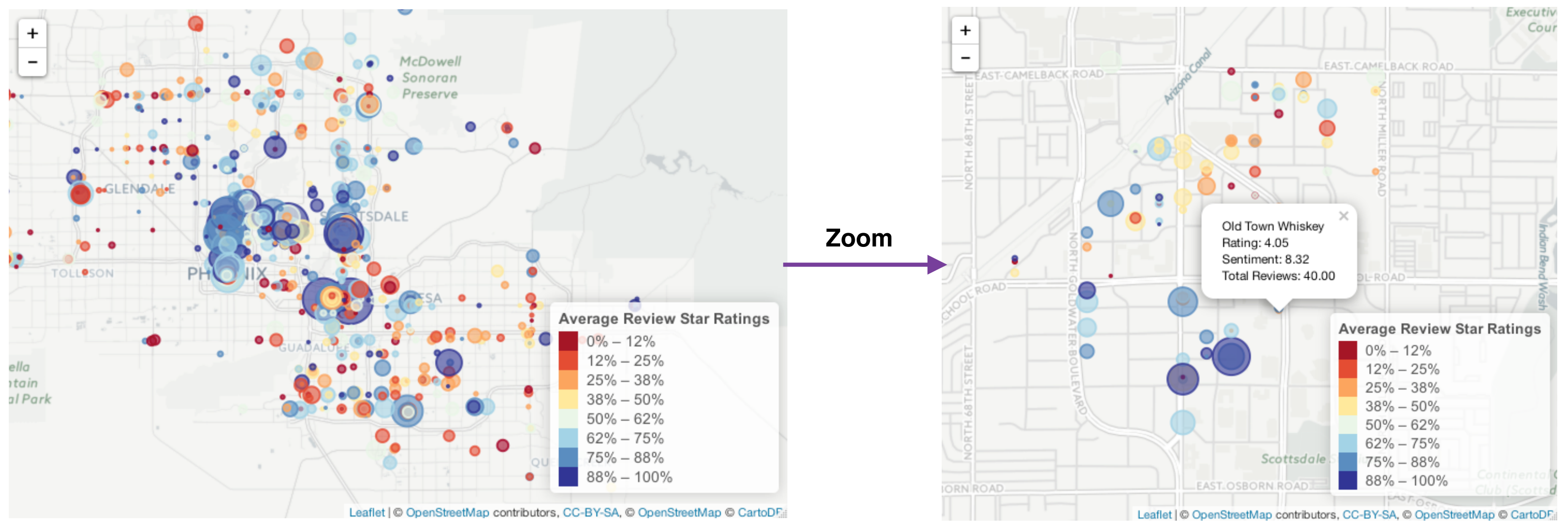
How can a business determine whether it will be profitable in an area? Who are it's biggest competitors? What are other similar business doing wrong in the area? What are the things consumers like? All this can be determined with the help of Yelp's dataset. It is the largest database for consumer reviews, this can help not only existing consumers but also new businesses. We used yelp's academic dataset along with R, Rattle and Shinyapps to do our analysis. We present the information in the form of an intuitive application. With the help of rattle we also made a model to predict the ratings of a business based on the sentiment score that we achieved by doing sentiment analysis of the review text provided by users.

Motivation & Goals

The motivation of the project came in from the Yelp academic dataset which was provided to students for research and project. The curiosity about how yelp uses it's huge information for revenue and also to help the consumers at the same time. It provides consumers with a platform where they can convey their experiences in form of reviews and rate a certain business. On the other hand it also helps businesses to improve themselves by looking at the reviews posted by the consumers. A new consumer business venture has big risks. It involves lots of investments. The location is not only important for the business but also for the consumers. The business should prove useful to maximum number of consumers. It is not feasible for a business to operate in an area where the products it offers are irrelevant to the consumers. Once the location of the business has been finalized it needs a plan on how it will implement it. For a business it needs to know about their competition in a given area. The do's and don't need to be identified. To determine all these factors one needs information about the area, the surrounding business and consumer demand. Consumers are the best source of this type of information. Consumer forums provide data about a consumer's view for a certain business which help us determine the quality of the business. Yelp is one of the biggest companies in the world which has data about business based on location and it also has information provided by the consumer. Our goal is to help the business with the information collected from the consumers to determine if doing business in certain area would be feasible.

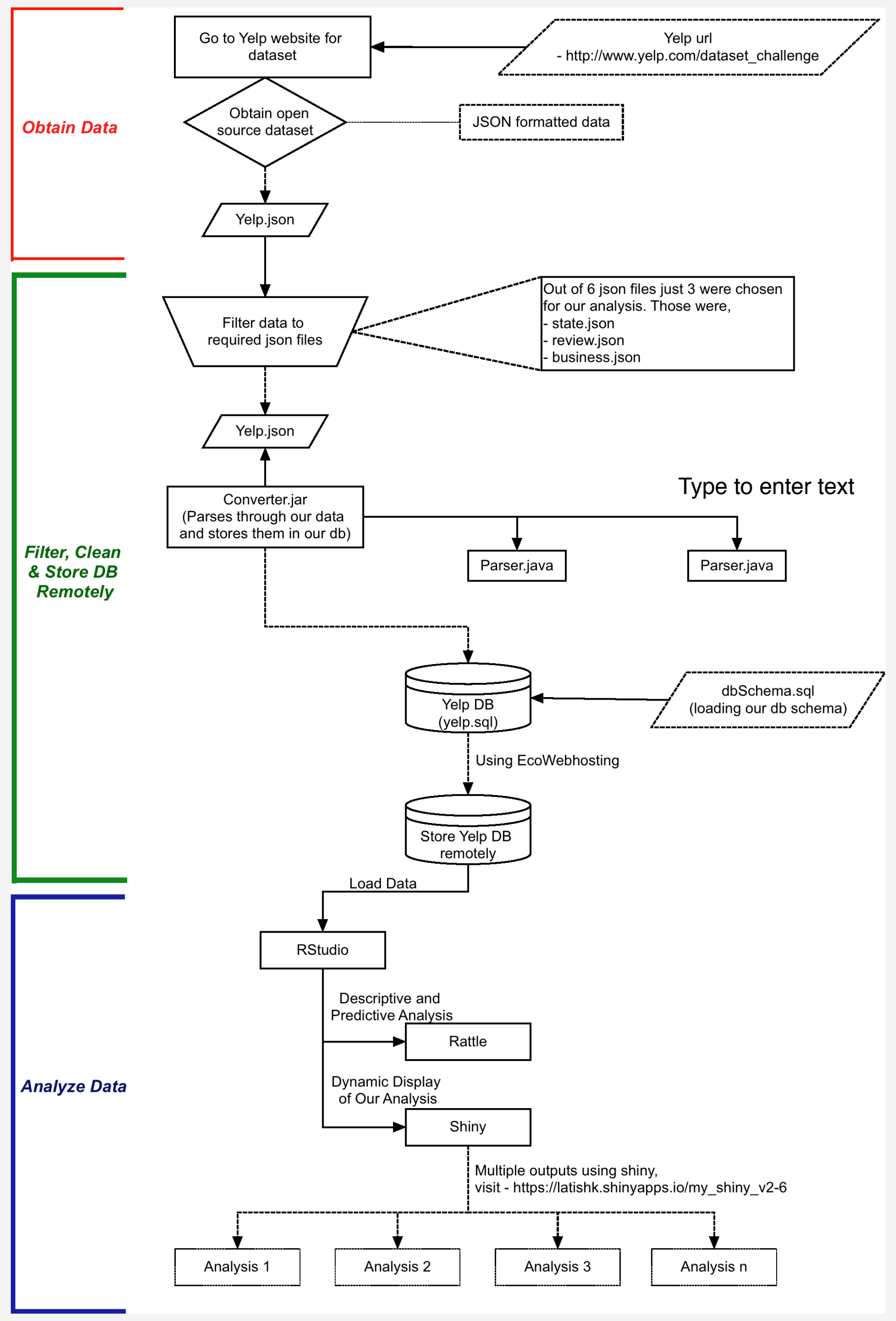
Design Considerations

Analysis needs to be presented and communicated well. We use tools such as Java language, MySQL, RStudio, R, Rattle and Shinyapp. Java and JDBC helped transform the data from JSON to MySQL tables. RStudio provided and interface to build and deploy the Shiny WebApplication. It was used to do all the tasks in R and Rattle. Rattle was primarily used to develop the models and perform classification.

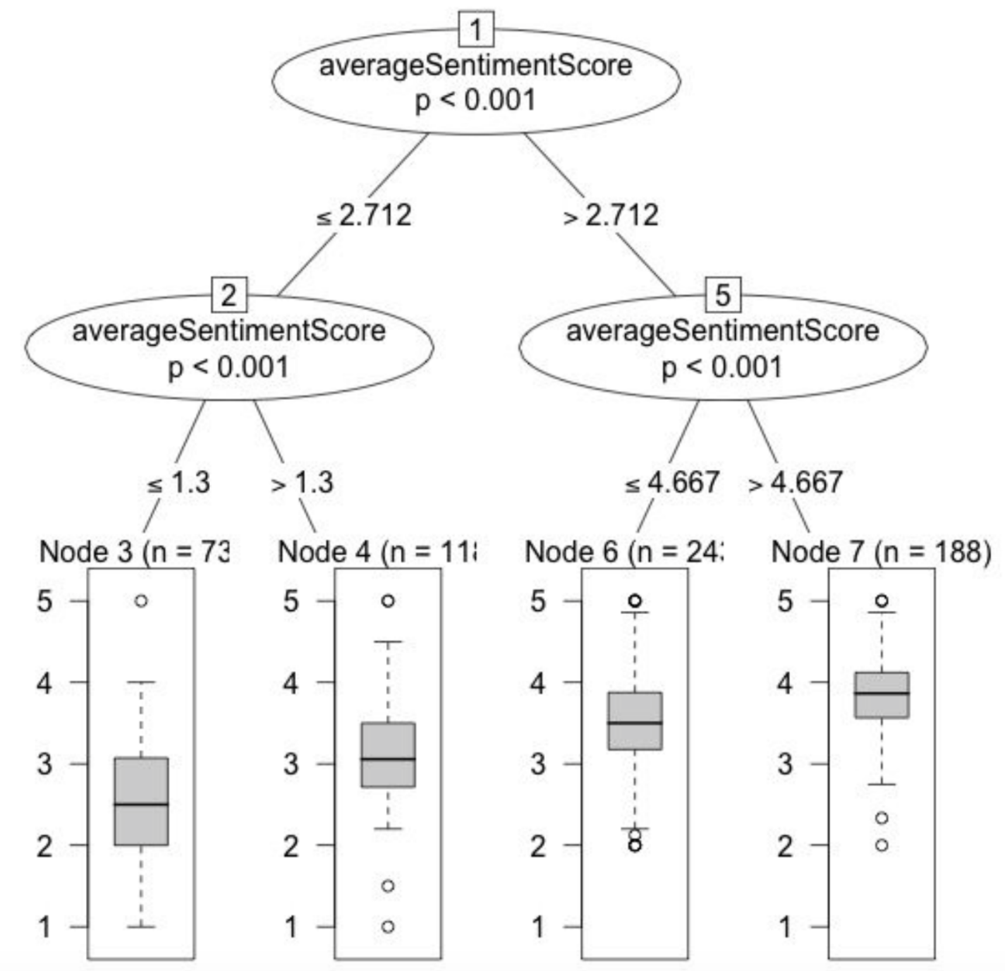


R Leafleat Visualization For Arizona State

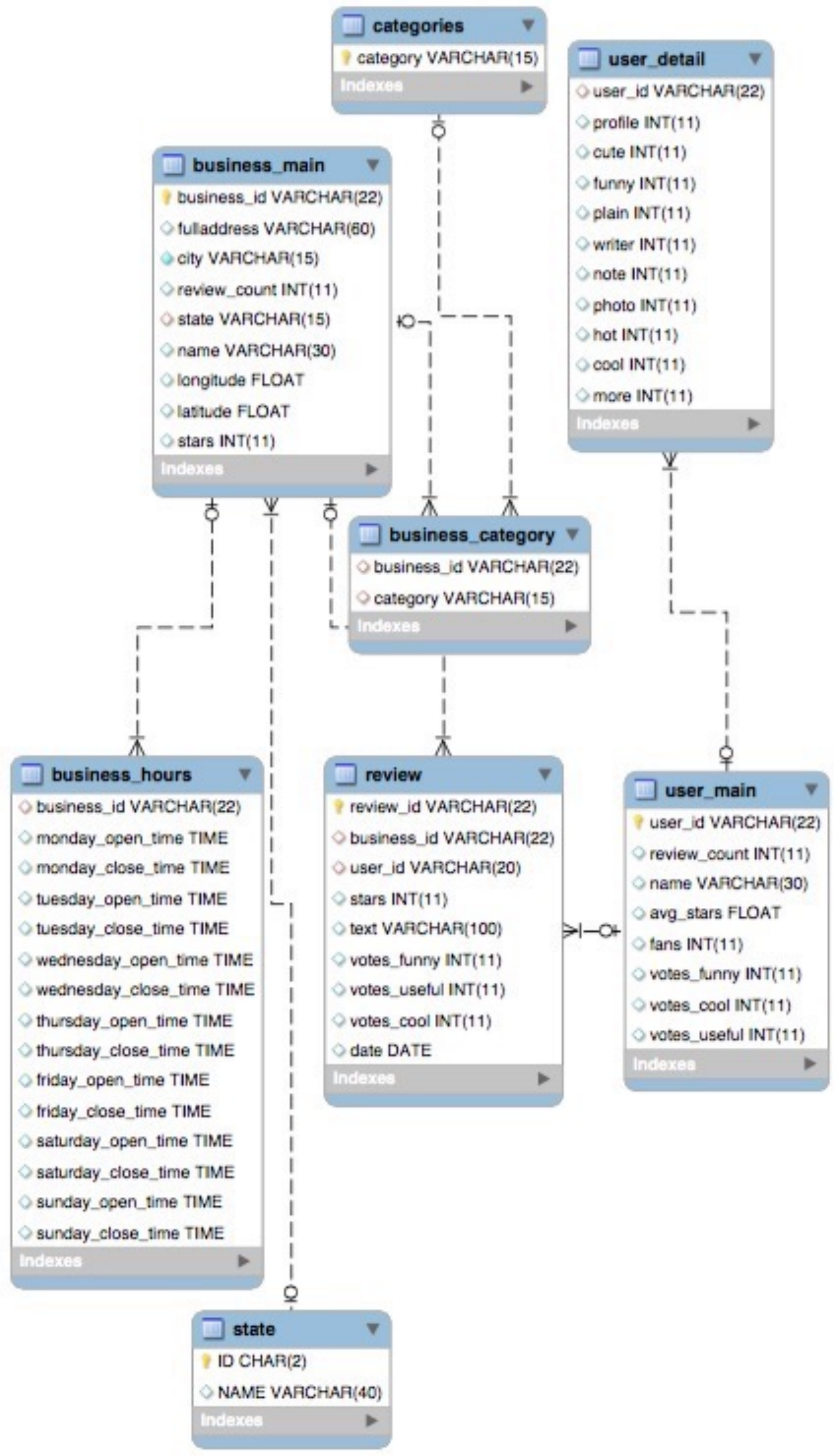
Implementation



Obtain Data
Go to Yelp website for dataset
Yelp url - http://www.yelp.com/dataset_challenge
Obtain open source dataset
JSON formatted data
Yelp.json
Filter, Clean & Store DB Remotely
Filter data to required json files
Out of 6 json files just 3 were chosen for our analysis. Those were,
- state.json
- review.json
- business.json
Yelp.json
Converter.jar (Parses through our data and stores them in our db)
Type to enter text
Parser.java
Parser.java
Yelp DB (yelp.sql)
dbSchema.sql (loading our db schema)
Using EcoWebhosting
Store Yelp DB remotely
Load Data
Analyze Data
RStudio
Descriptive and Predictive Analysis
Rattle
Dynamic Display of Our Analysis
Shiny
Multiple outputs using shiny, visit - https://latishk.shinyapps.io/my_shiny_v2-6
Analysis 1
Analysis 2
Analysis 3
Analysis n



Predictive Model of Sentiment Score Versus Review Rating



Relational DB Schema

Lessons Learnt

Initially we planned to do analysis on the dataset but as we saw the dataset was huge, it couldn't be handled by Rattle. It needed tools which were made to handle such huge amount of data, like Hadoop. So, we chose to use a smaller dataset. We learnt that what is true for one business in terms of predicting the rating changes drastically when applied to any other business in different location. So it is necessary to classify the businesses based on location and the type of business they do. We planned to use just rattle for data analysis but as we progressed we saw the need for an interface to make the analysis dynamic in nature to present itself with results applicable to multiple instances. A web application is good medium to showcase the data as most of the people are accustomed to a web browser. While dealing with the database we realized for our web application CSV format files were easier to read and process. We found that instead of just using R interface, RStudio is better IDE which has many tools integrated into it, Also, rattle can also be used from RStudio.

Discussions

Current work.

The web application has been completed and can be accessed to check business in Arizona state across eight categories. The data analysis has been been largely complete. We could predict the ratings based on the sentiment analysis score. It predicted a rating of 'The Keg Steakhouse + Bar' as 3.876 based on the sentiment analysis score which is very close to the actual average rating of 3.912.

Future Work

Use of tools which can process large amounts of data such as Hadoop, HDFS and GFS. One can also improve the quality of the sentiment analysis using classification methods which take the structure of the sentence to rate if positive or negative, instead of just the word frequency. It is also possible to improve the ratings based on a user's reliability by accessing the number of useful reviews written and total number of fans.

References

[1] Yelp Dataset Challenge. http://www.yelp.com/dataset_challenge/, 2015.

[2] J. Breen. Twitter text mining R slides. https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/, 2011.

[3] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of the 14th International Conference on World Wide Web, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.