FINAL REPORT

Kinjal Gada

NUID: 001643864

INTRODUCTION

For any data mining / data analysis project, choice of the dataset plays an important role in achieving final result. It is very crucial that the we know beforehand, what information can be obtained from the dataset and that it aligns with the goal of the project. We should be able to question the dataset and the dataset should have an answer for it. In this report, we will talk about how we can explore the chosen dataset - Yelp, and in what ways can the information obtained from the dataset be used. We will discuss about performing predictive and reportive analysis.

The end result of this analysis will be beneficial for customers and businesses both.

PROBLEM

Given a dataset, one major question we have to face is- what do we do with it? Just having dataset is not necessary. Being able to get something useful from it is of utmost importance. One of the ways to explore the data and extract meaningful pattern from a dataset is by performing predictive and reportive analysis. How do we implement it is another question.

GOALS

Goal is to implement following predictive and reportive analysis:

1. Analyze reviewer's ratings from its text.
2. Predict when a business could be most busy.
3. Predict when a business would be open.
4. If the business has wifi.
5. If the business has parking.
6. If the business is good for kids.
7. What are the difference in the business between the cities.
8. How much is business's success due to own its location or popularity.
9. Find the businesses with highest star ratings.
10. Determine the time during the day with maximum check-ins which will indirectly complement goal 2.
11. Find the average rating given by a particular user to a particular type of

business to understand what a user expects.
12. Find the businesses which have most number of reviews.
13. Find type of cuisine a customer prefers most of the time.
14. Implement visual outputs for better understanding of the analysis.
15. Implement different algorithms like K-means, Random Forest, Neural Network, etc. to find what type of business customer will search for next.

MATERIALS

Yelp Dataset for

1. User
2. Tip
3. Review
4. Checkin
5. Business

PROCEDURE

Using map reduce chaining and with various design and filtering patterns, partitioner, and combiners, different types of analysis can be inferred.

Steps:

1. Read JSON object.
2. Convert it to an array to extract the data.
3. Clean the dataset.
4. Filter the dataset for the type of outcome required.
5. Implement partitioners and/or combiners to make the mapreduce more efficient.
6. Implement various design patterns and joins wherever required.
7. Implement algorithms to get accuracy in the model
8. Divide the data set in train and test dataset, in order to first train and then test the model.
9. Visualize the output for better understanding.

EXPECTED RESULTS

With the help of historical Yelp dataset, Yelp could be more popular with the correct and useful analysis and predictive results by improving the businesses and next best business could be customers next choice. Whereas for customers it will help them try new places which could turn out to be a safe option too.

REFERENCES

1. https://www.yelp.com/dataset_challenge