Team 6: Final Project Report

Patrick Kelly, Christoph Sieberer-Kefer, Somesh Yadav, Yash Wali, and Patrick Foli

28 April 2022

Yelp! Project Report

Project Summary

Data Exploration:

One of the findings through our thorough data exploration was that many check-ins happen at or after midnight. This coupled with the fact that most businesses in our Yelp! datasets are located in the Western US, we concluded that the hours recorded for check-ins may not be adjusted for time zones.

Further visualizations of our data analysis can be viewed in the Appendix Section of this report.

GraphFrames findings:

Our goal was to quantify the influence a user has based on their connections (direct friendships and downstream connections) using two algorithms supported by GraphFrames. For either we created vertices and edges and we expected the pageRank and triangleCount algorithms to quantify said influence, but the results did not reflect any meaning (as can be seen in **Figure 18 & 20**).

Sentiment Analysis findings:

We implemented two models to determine sentiment, one unsupervised and one supervised. The unsupervised model uses TextBlob, which is a python library for NLP, and returns primarily positive sentiment learned from reviews (**Figure 17**). The resulting sentiment labels (positive, neutral, and negative) showed a decent correlation to the star feedback given by the users. Please refer to **Figure 19** to see a table showing the produced sentiment labels and the actual star ratings associated with each review. Our second model, a logistic regression, makes use of Spark's pipeline concept. We defined transformers (including HashingTF and IDF) in our pipeline and an estimator (which is defined by the pipeline). The estimator transformed our data as intended and with that we fit a logistic regression (LR) model (also an estimator). While the performance in terms of execution time and accuracy was promising, the one disadvantage was having to classify the reviews as either positive or negative. Neutral was due to the binary classification nature of a LR not a possible outcome.

Apart from the user influence objective, we were able to complete our data exploration and analysis as well as our model implementations for sentiment analysis.

Challenges

- Reading the CSV file into spark DataFrame, the ‘text’ column, was being read as multiple columns and rows instead of one creating incorrect alignment of the rows with their corresponding columns. Our solution was to add `multiline = True’ and `escape = “ ” ” while reading the file.
- The review file was too large to convert to Pandas DataFrame hindering some Word Cloud and other visualizations on the Review file.
- Running Graphframes, MLlib for sentiment analysis, and Java Heap Space on the local computer and Google Colab took hours even though we were running on 14GB of RAM. We did configure it for better driver memory but it still was a cumbersome process.
- Graphframes Pagerank algorithm did not give favorable results due to memory constraints, the number of users in the Dataframe were not able to circle back thus all pageranked similar as no node reached back they all.

Updated Member Contributions

	Description of Work
Christoph Sieberer-Kefer	Project coordinator, Coding(Graphframes & Sentiment Analysis), Project Report, PPT
Patrick Kelly	Coding(Spark Dataframe and Visualizations), Project Report, PPT
Somesh Yadav	Coding(Graphframes & Sentiment Analysis), Project Report, PPT
Yash Wali	Coding(Spark Dataframe and Visualizations), Project Report, PPT
Patrick Foli	Coding(Support for both teams), Project Report, PPT(lead)

Appendix

Figure 1.

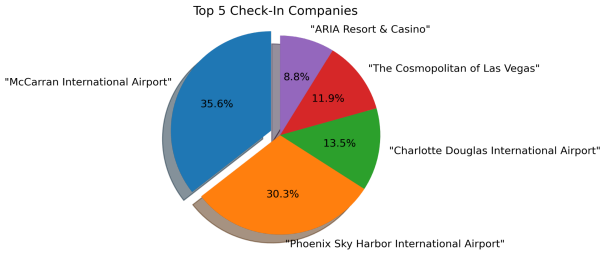

Figure 2.

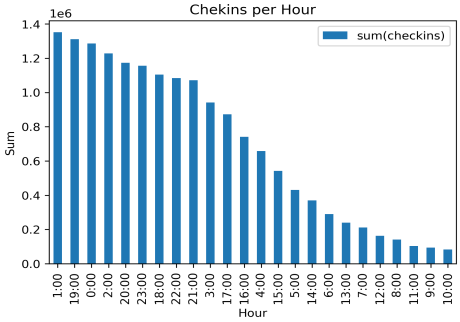

Figure 3.

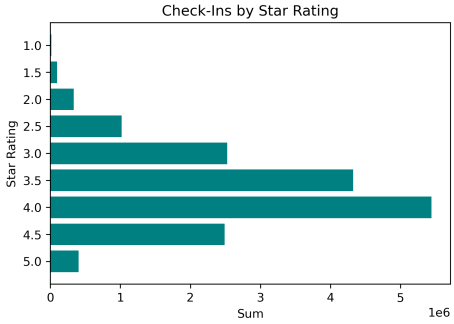

Figure 4.

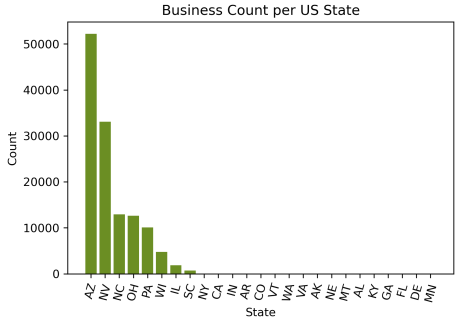

Figure 5.

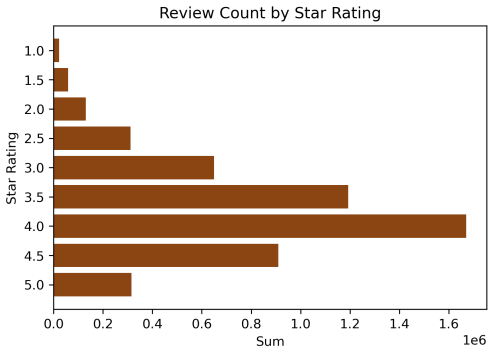

Figure 6.

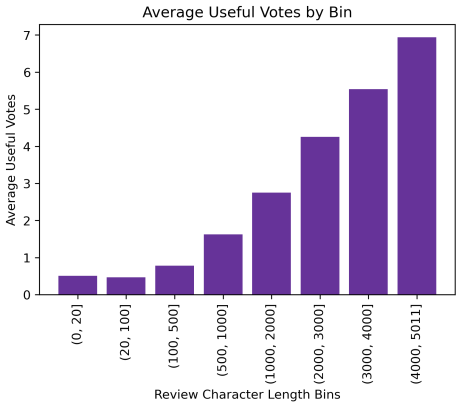

Figure 7.

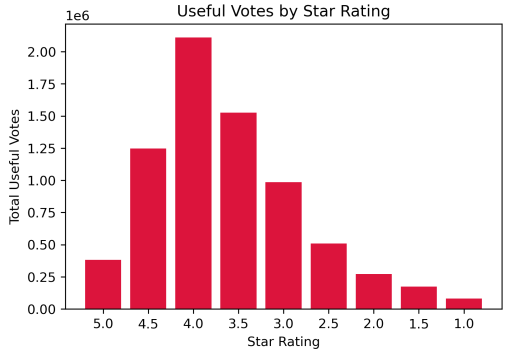

Figure 8.

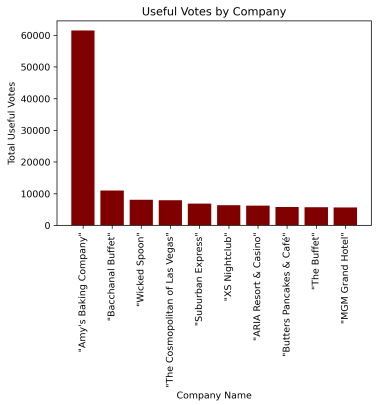

Figure 9.

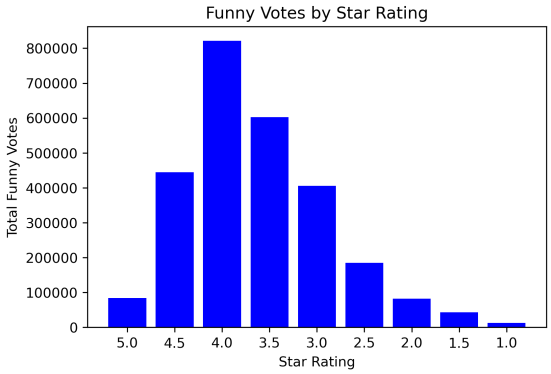

Figure 10.

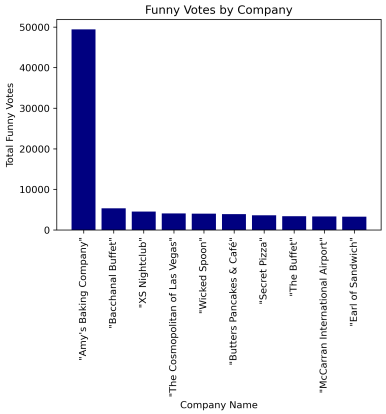

Figure 11.

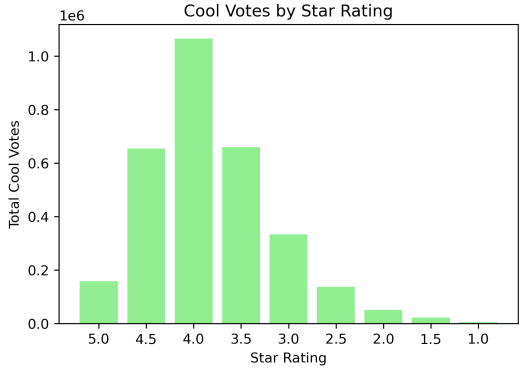

Figure 12.

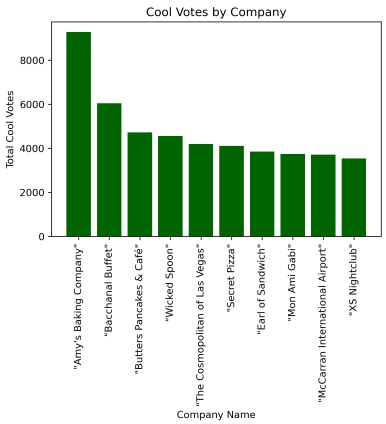

Figure 13.

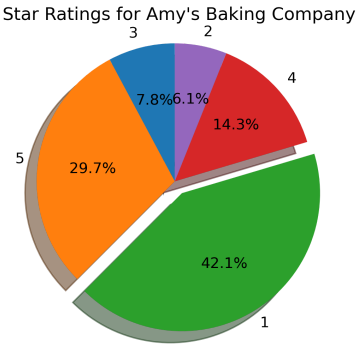

Figure 14. Full WordCloud

Figure 15. Bad Reviews (1-3 stars)

Figure 16. Good reviews (4-5 stars)

Figure 17.

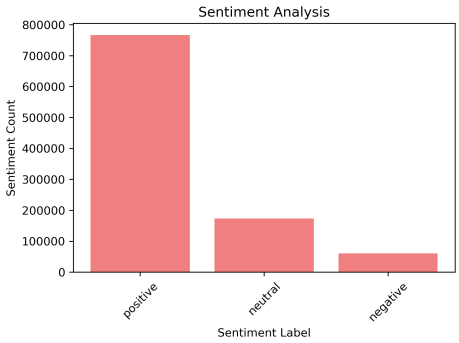

Figure 18.

```
+--------------------+------------------+----+
|                  id|          pagerank|name|
+--------------------+------------------+----+
|4E8--zUZO1Rr1IBK4...|0.9999999999999987|Lisa|
|H54pA7YHfjl8IjhHA...|0.9999999999999987|Chad|
|Ob-2oGBQ7rwwYwUvh...|0.9999999999999987|   B|
|WRae-wZkpRoxMrgJd...|0.9999999999999987|Mike|
|kmyEPfKnHQJdTceCd...|0.9999999999999987|   A|
+--------------------+------------------+----+
```

Figure 19.

```
+---------------+-----+
|sentiment_label|stars|
+---------------+-----+
|       positive|    5|
|       positive|    5|
|       positive|    5|
|       positive|    4|
|       positive|    4|
|       positive|    5|
|        neutral|    4|
|       positive|    4|
|       positive|    3|
|       positive|    5|
+---------------+-----+
```

Figure 20.

```
+--------------------+-----+
|                  id|count|
+--------------------+-----+
|W5mJGs-dcDWRGEhAz...|    0|
|JJ-aSuM4pCFPdkfoZ...|    0|
|mBneaEEH5EMyxaVyq...|    0|
|4E8--zUZO1Rr1IBK4...|    0|
|uUzsFQn_6cXDh6rPN...|    0|
+--------------------+-----+
```