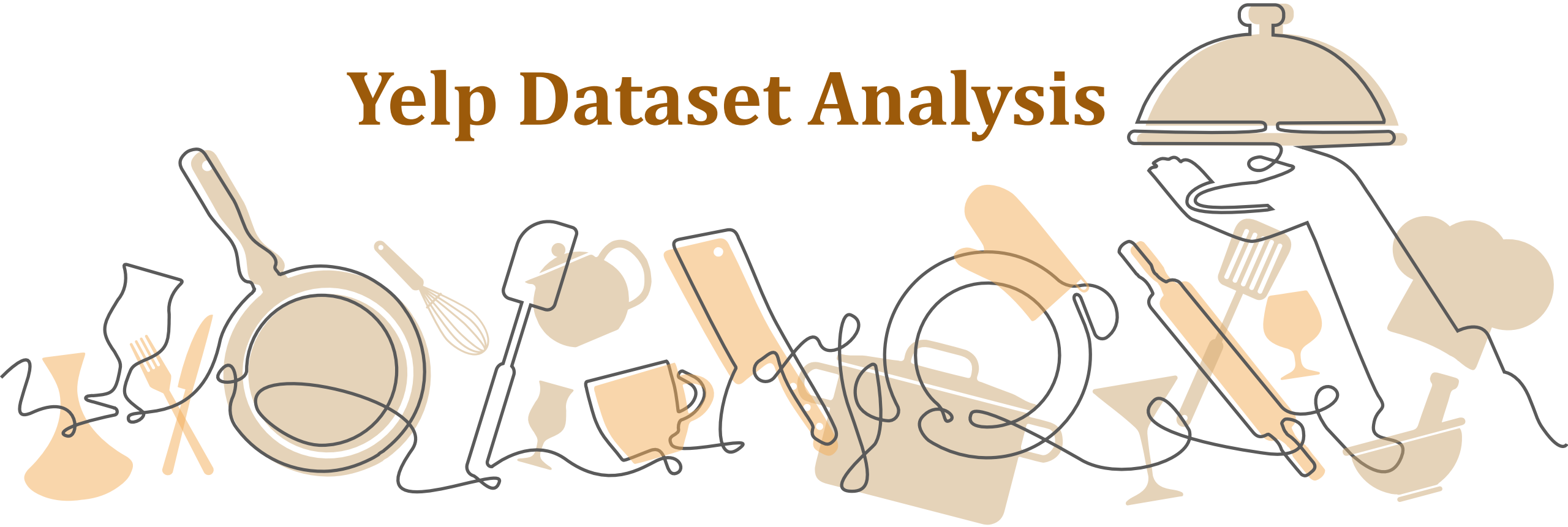
Yelp Dataset Analysis



Omkar Patinge (opp212)
Urmi Jadhav (uj256)

Abstract

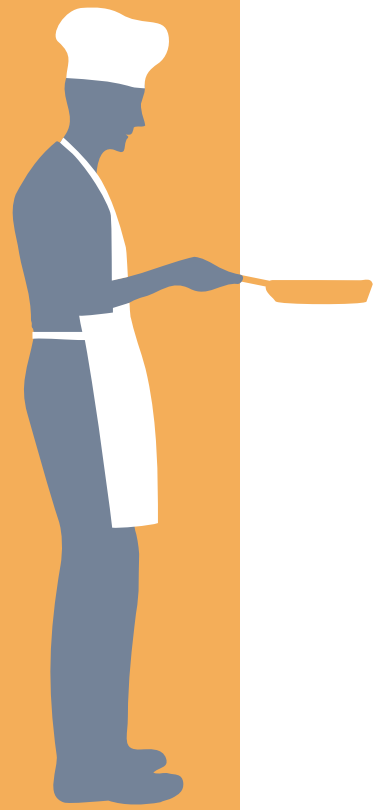
Recommender system is one of the most essential components of any social network. In this case, the Yelp dataset is examined to derive useful information to recommend places based on connections, check the quantified correctness of the reviews, top featured places and busiest times to visit a place.

Opinion mining(NLP) is used on reviews of a particular place to understand the intent of the user and is correlated with the average review for that place. This will help the place owners to decide whether to take a look into the comments for some useful information. Recommending the places to users based on their features and connection will also increase user engagement. Analyzing the check-ins to derive the busiest hours and busiest dates using time series analyzing.

Motivation

- Analyzing reviews will help the place owners to further take a deep dive into the suggestions and user's feedback away from quantified review field.
- Assisting the user to find the right time to visit a particular place by analyzing check-ins is an essential thing.
- Making better use of such social network data for the betterment of the different kinds of users helps everyone grow.

Motivation



1

Who are the users of this application?

The users of the application will be business owners as well as the general public who looks at the ratings and reviews before paying for any service.

2

Who will benefit from this application?

The beneficiaries of the application will be the same as above i.e. the business owners as well as the customers

3

Why is this application important?

In today's time, a lot of importance is given to the opinions posted on social networking sites. However it is important that the users get accurate gist from the plethora of information available online. This application helps in achieving exactly that by processing all the ratings and reviews for a business and displaying the exact sentiment of the reviews

Goodness

- The results obtained from the check-in analytic were verified with the occupancy distribution generated by Google. Also, the restaurants have a trend to be busier during the weekends in comparison to the weekdays while casinos are busiest during the night which coincides with the predictions of the application.

- The words in the positive wordclouds have a positive tone to them like great service, delicious, worth the trip while the ones in the negative wordclouds are pretty bland, worst service, not worth the money. This is generated directly from the review data and is in line with the expected results.

- The food recommendations for a restaurant also coincide with the maximum good reviews for an item as stated by the customers.

Remediation

1. Create alerts for users for the best/worst time to visit a business on any given day

2. Notify business owners when there is a dip in ratings

3. Suggest the most popular/highly rated food items to order at a restaurant

4. Highlight the most positive and negative factors of a business using wordclouds

Data Sources

The data for the application was obtained from Yelp.com that was released as a part of their Yelp Dataset Challenge

REVIEWS

Contains the reviews by different users for a particular business along with other user feedbacks(useful, fuuny, cool etc.)

BUSINESS

Contains the data about businesses throughout the USA along with the characteristics offered and location.

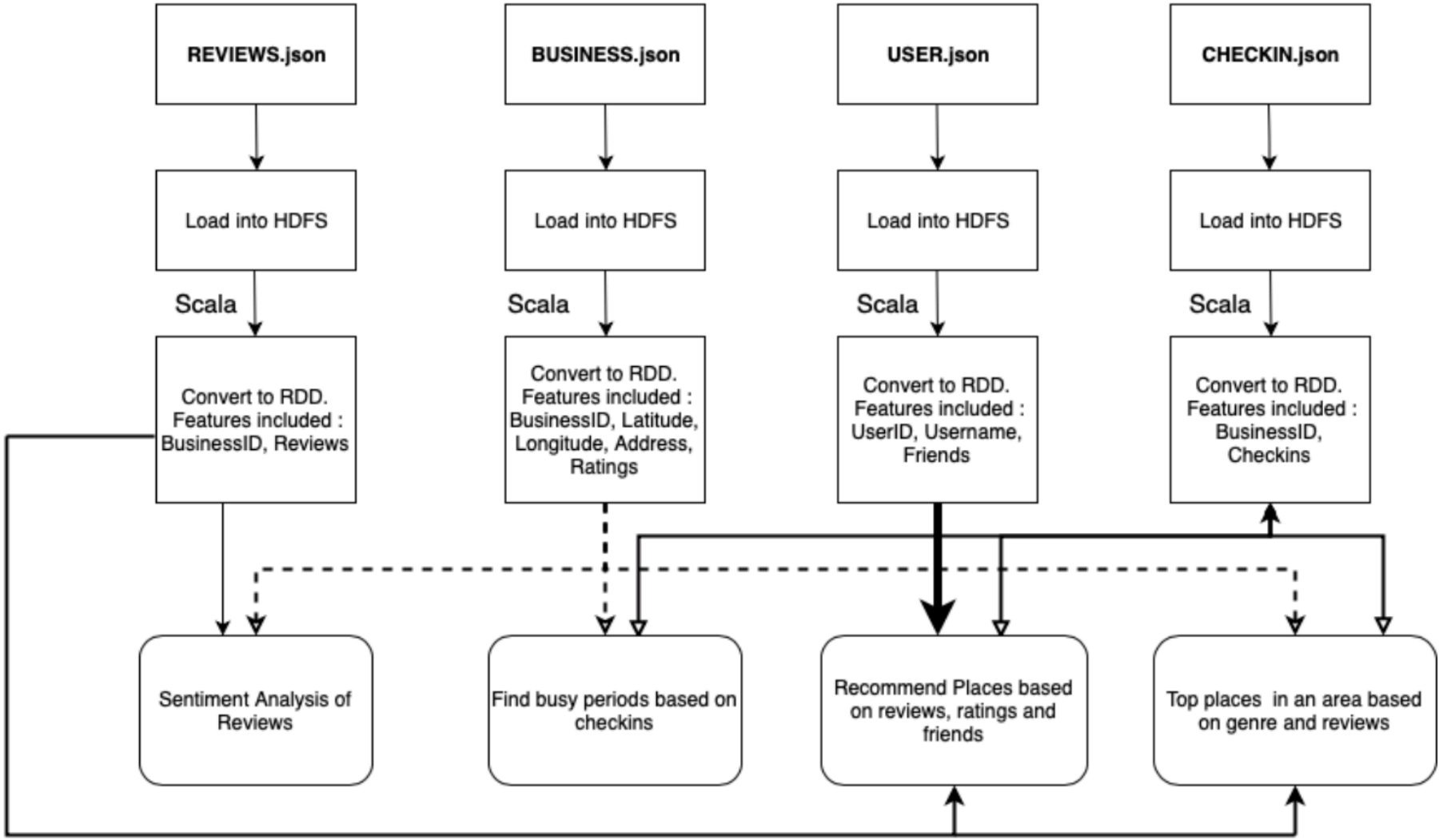
USERS

Contains the user data and connections

CHECK-INS

Contains the check-ins at a particular location with timestamp

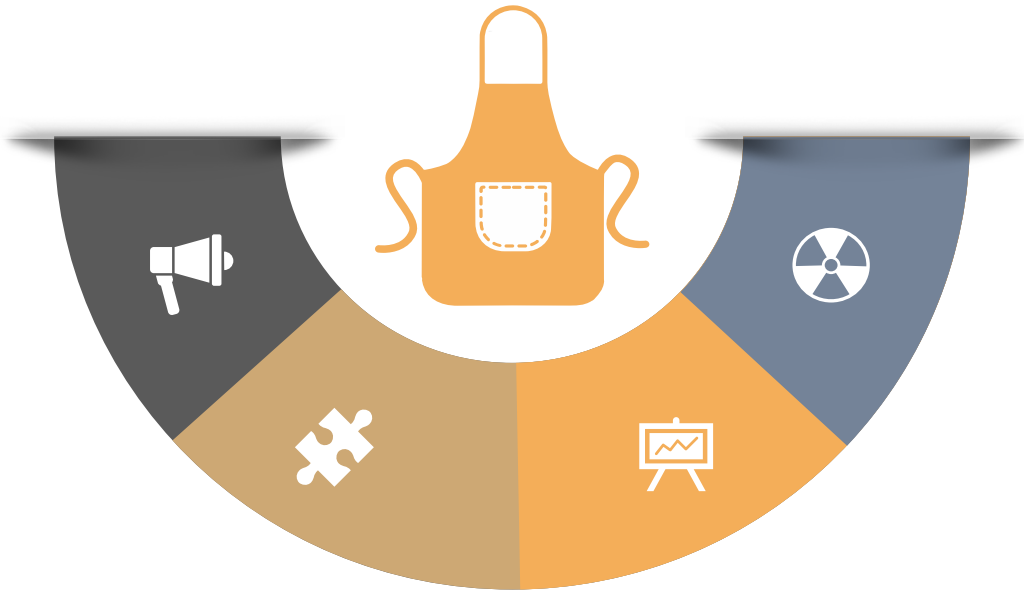
Design Diagram



Platform on which the application runs:

The application runs on the NYU HPC Cluster

Insights



1. The best time to visit a business

Notifies customers about the best time when they should visit a business by analyzing the check-in times logged in by the other users and finding the time slot which guarantees prompt service and lower wait times.
It also tells the user when they should avoid the business due to long wait times.

2. Performance of a business on the basis of user ratings

Informs business owners about how their business is doing in terms of customer appreciation in terms of review ratings. Tells them when they performed their best and also when there is a need for intervention in order to improve their ratings

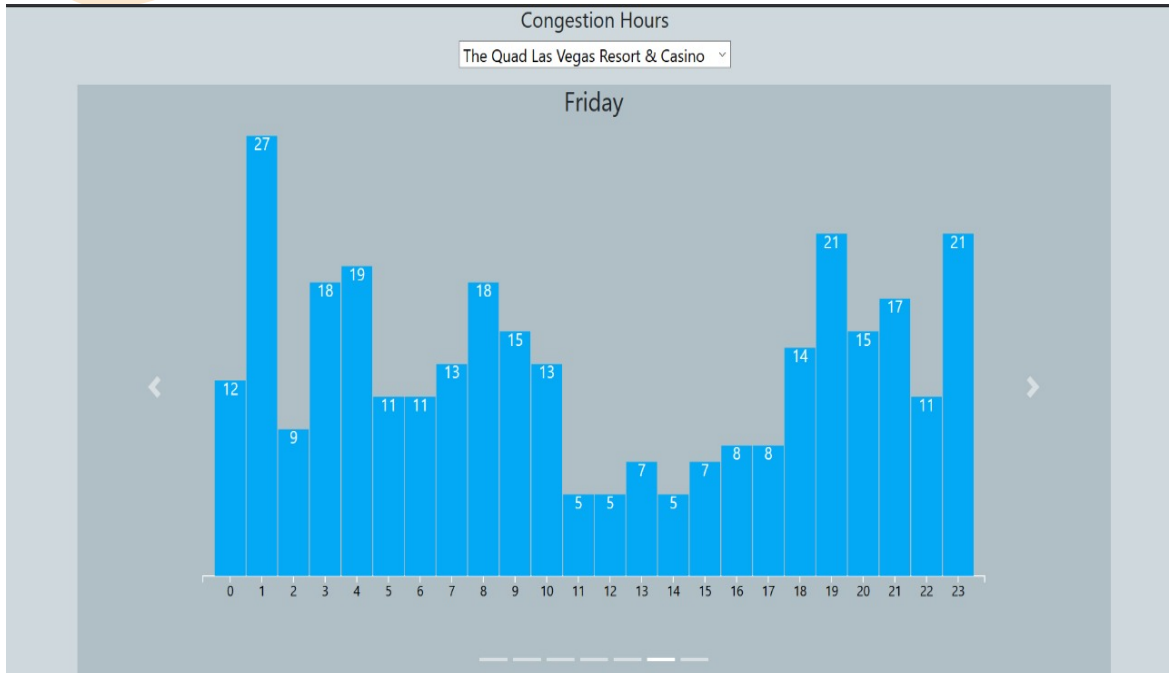
3. WordClouds: positive and negative factors

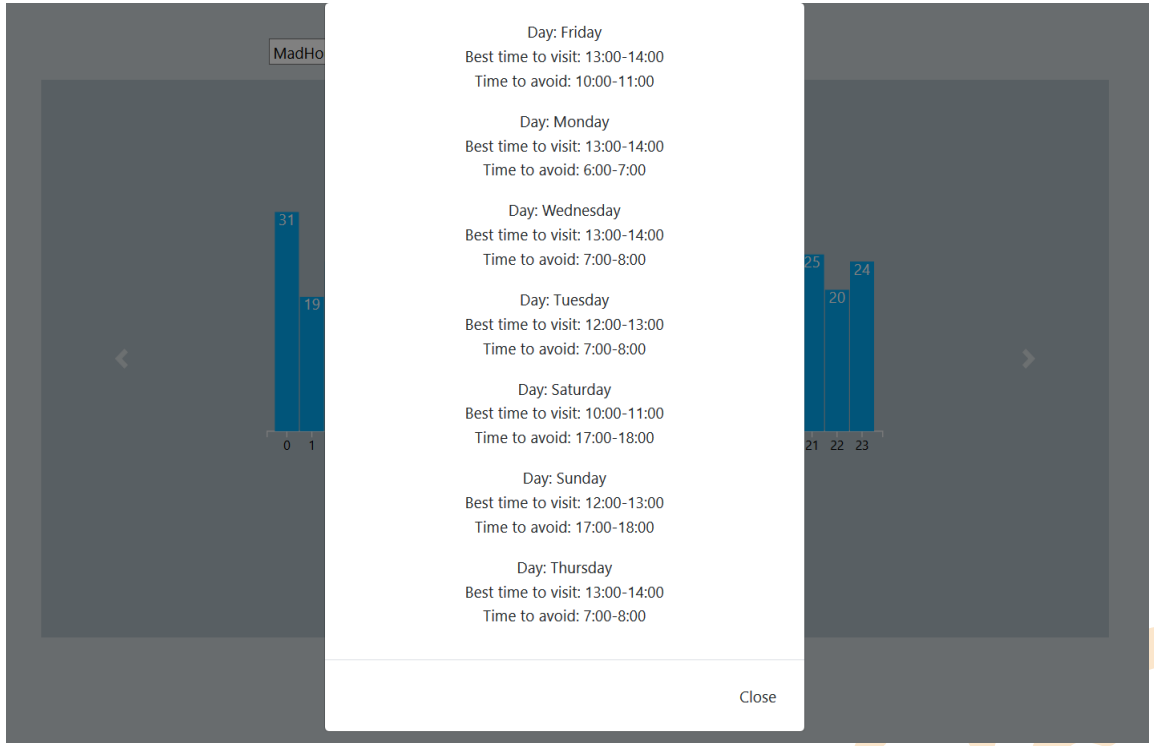
Generate wordclouds that represent the most talked about aspects of a business: both positive and negative. This informs the owners as well as the customers about what is doing well and what needs improvement.

4. Food Recommendations

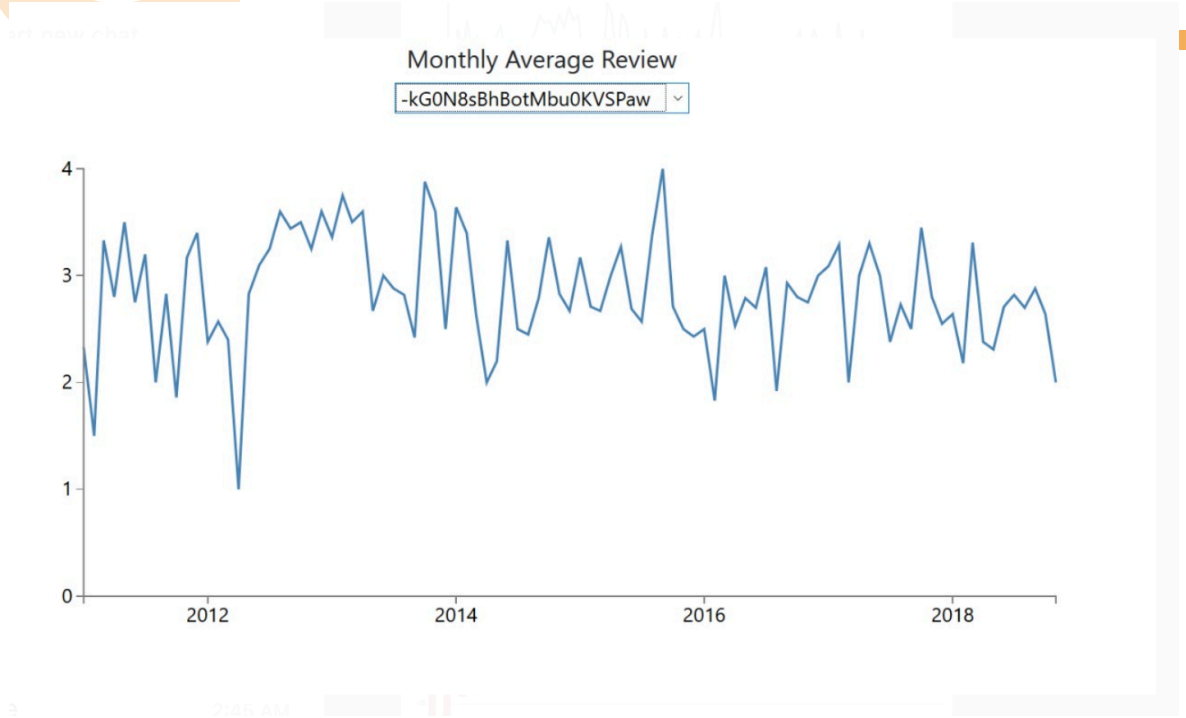
Tells the users about the most popular and highly rated food items at the particular restaurant by using sentiment analysis on the reviews.

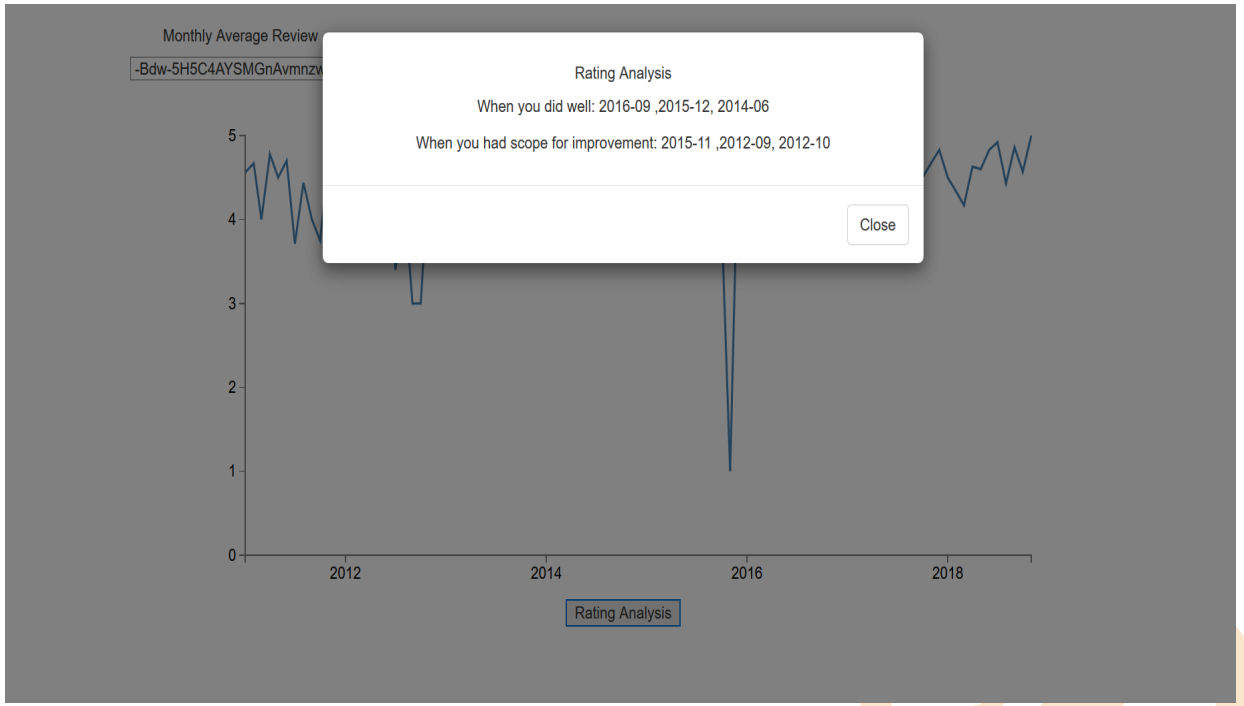
1. The best time to visit a business



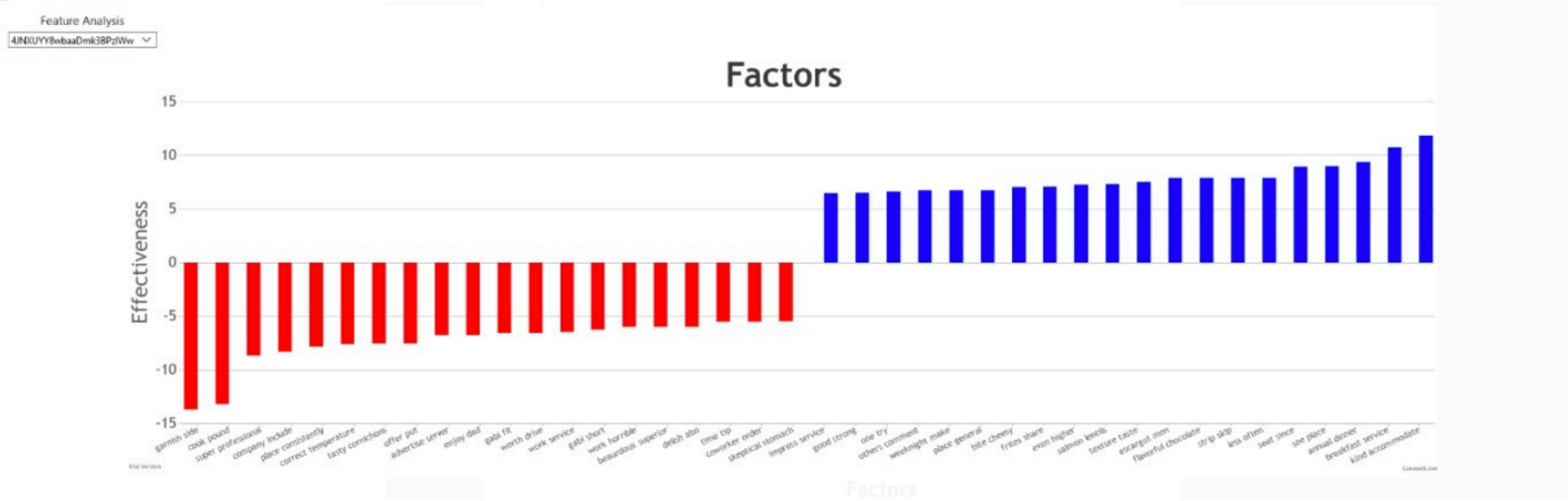


2. Performance of a business on the basis of user ratings

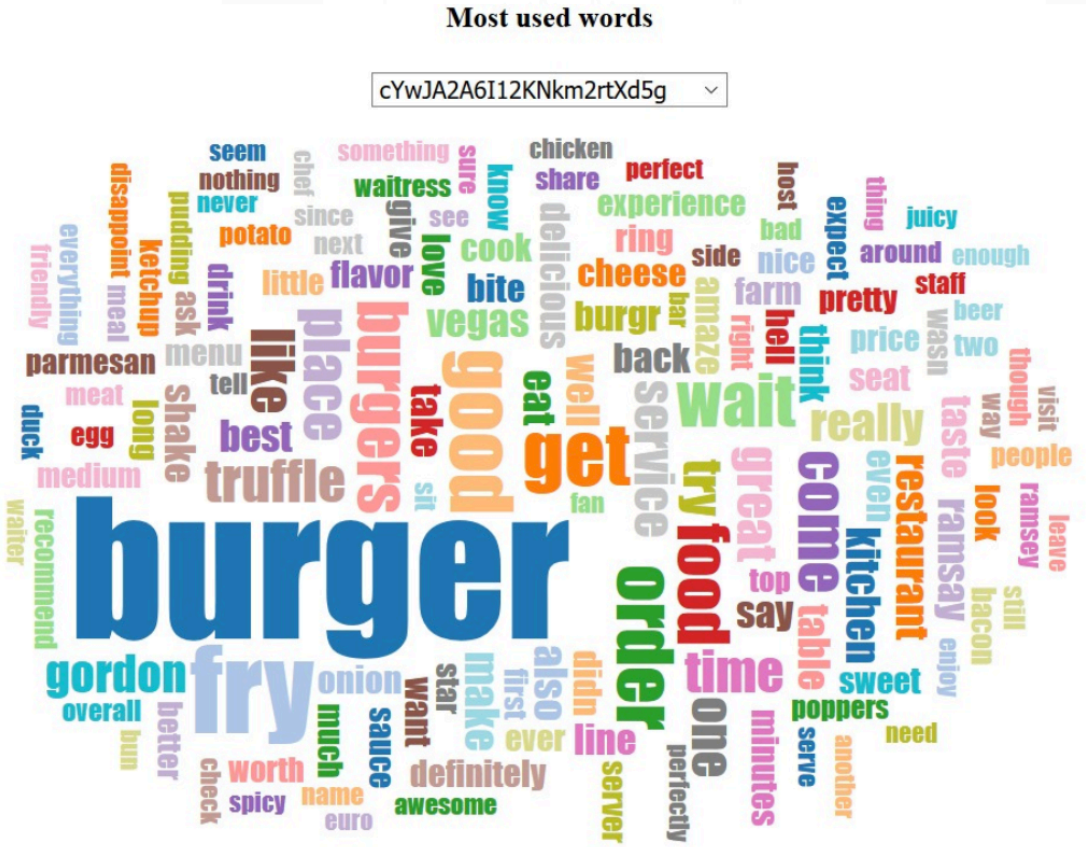


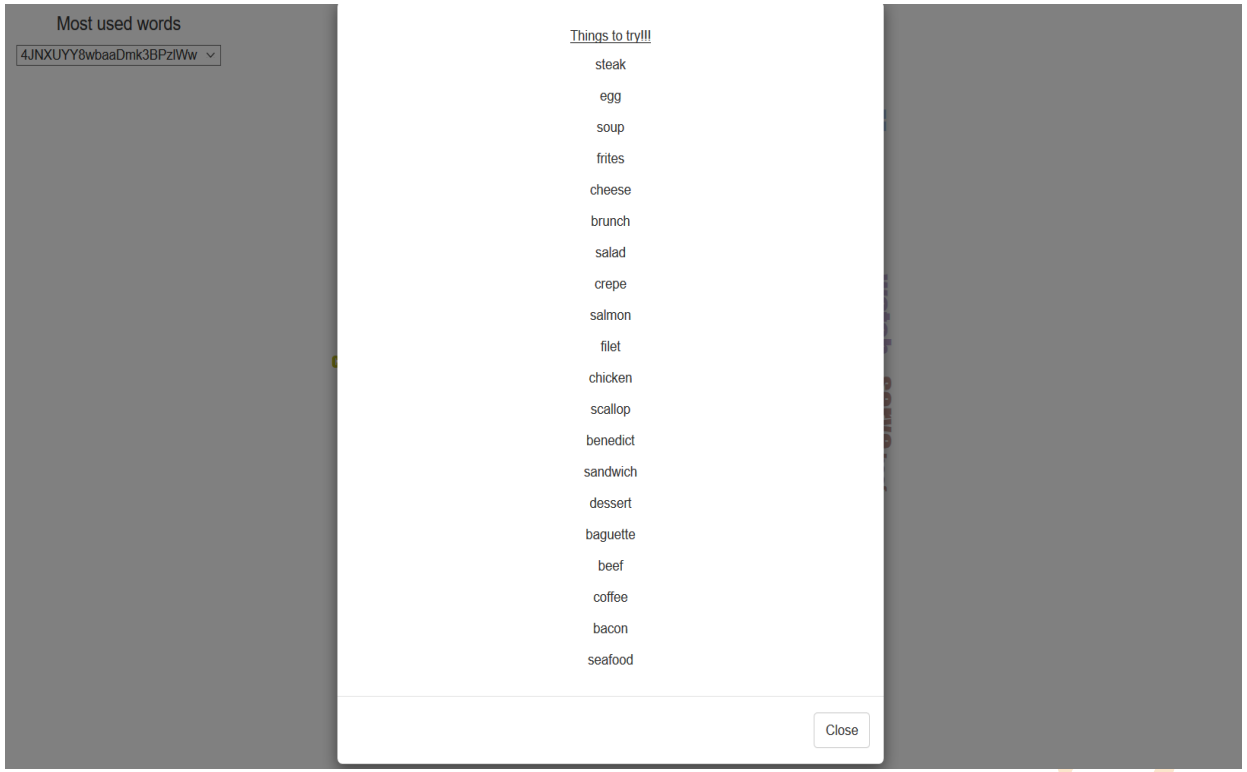


3. WordClouds: positive and negative factors



4. Food Recommendations





Obstacles

1 | **Size of the data and speed of processing**

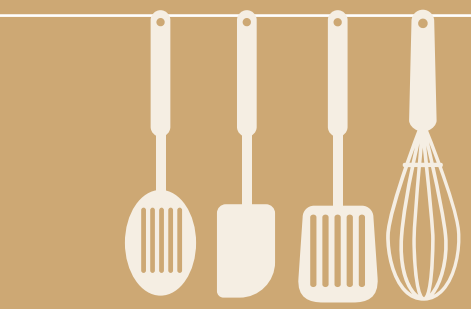
The data was too big to be handled outside the Hadoop cluster. It took about 3-4 hours to run a single analytic on individual business-ids.

2 | **Discrepancies in the data**

Some of the addresses did not match with the zip-codes provided.

3 | **Irregular languages and formats in reviews**

The reviews had comments in different languages and included emoticons which made them harder to decipher.



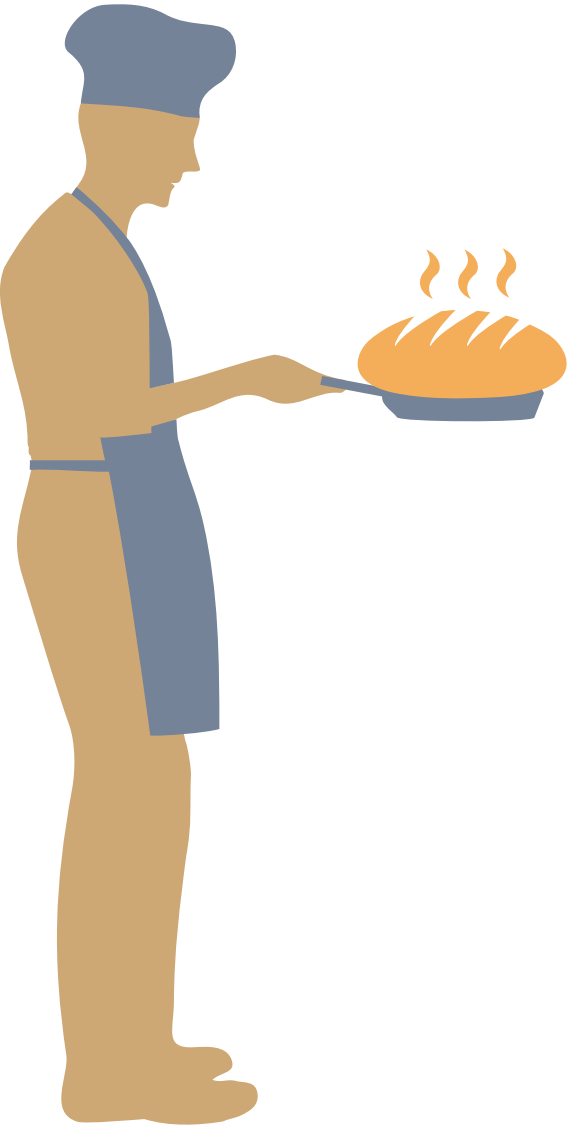
Summary

The application does an in-depth analysis of the Yelp dataset and generates several actionable insights using PySpark, NTLK and other data analytic tools.

It provides the user with food recommendations, best time to visit a place, the average ratings of a business over a period as well as the general consensus among the public regarding the business by highlight the pros and cons of the place.

It also provides visualization of the results using JQuery and d3.

Acknowledgements



We would like to thank Professor Suzanne McIntosh for her constant support and guidance.

We are also thankful to NYU HPC for providing access to Dumbo and other big data tools and for helping us throughout the project.

Lastly, we would like to express our gratitude to the graders and fellow classmates for resolving our queries promptly.

References

1. https://www.yelp.com/dataset/challenge
2. Anandhan, L. Shuib, M. A. Ismail, and G. Mujtaba, “Social Media Recommender Systems: Review and Open Research Issues,” IEEE Access, vol. 6, pp. 15608–15628, 2018.
3. Nayebzadeh, M., Moazzam, A., Saba, A., Abdolrahimpour, H., & Shahab, E. (2017). An investigation on social network recommender systems and collaborative filtering techniques. arXiv:1708.00417
4. He, Jianming, and Wesley W. Chu. "A social network- based recommender system (SNRS)." In Data Mining for Social Network Data, pp. 47-74. Springer US, 2010.
5. M.Pazzani, D.Billsus,Brusilovsky P, Kobsa A., Nejdl W. (Eds.), The Adaptive Web: Methods and Strategies of Web Personalization, Springer Berlin Heidelberg, Berlin, Heidelberg (2007), pp. 325-341
6. Imane El Alaoui, Youssef Gahi, Rochdi Messoussi, Youness Chaabi, Alexis Todoskoff, Abdessamad Kobi. “A novel adaptable approach for sentiment analysis on big social data”link.springer.com/article/10.1186/s40537-018- 0120-0
7. Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai. “Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges”
8. https://spark.apache.org/docs/latest/mllib-feature- extraction.html
9. https://web.stanford.edu/class/cs124/lec/sentiment.pdf

Thank you!