CS179G Parts 2 + 3 Report

Vinayak Gajjewar Rayyaan Mustafa Kevin Nguyen Anthony Gao

Requirements

This project aims to run a comprehensive data analysis on the Yelp review and business datasets and display this analysis in a visually appealing interface. The data provided by Yelp contains two datasets, a business dataset and a review dataset. The data analysis portion includes data cleaning and computing several statistics about the data using Apache Spark. To complete this project, we need to install MongoDB onto the EC2 instance to store our dataset. Likewise, we need to install Apache Spark onto our system.

Design

For our database, we decided to use MongoDB. We chose to use a non-relational database because it fit our data format better. Relational databases don't support arrays, so without any preprocessing of the data, MySQL would not support the "categories" attribute that contains an array of strings in the business dataset. The two datasets we use have a great variety of data since not every entry has the same attributes. NoSQL databases such as MongoDB are better suited to handle this complexity.

We wanted our data to be able to be filtered by average rating by state, average rating by cuisine, average review sentiment by state, and average review sentiment by cuisine. Additionally, we wanted to collect various statistics about the data such as the minimum or maximum number of stars given to any particular restaurant category. We gather this information using spark and store it in a file. Then we port this information over to our front end. Our front-end is a 2d map interface in which users can pan, zoom, and click on select locations to find more information about it. Additionally, users will be able to filter the points on the map using textbox inputs.

Implementation

We used the open-source mapping library OpenLayers in our front end because it allows us to easily map points and color-code them. We wanted to show the data in a mapping format to show the data in a spatial format. We use raw HTML, CSS, JS as well as node.js to build out the website.

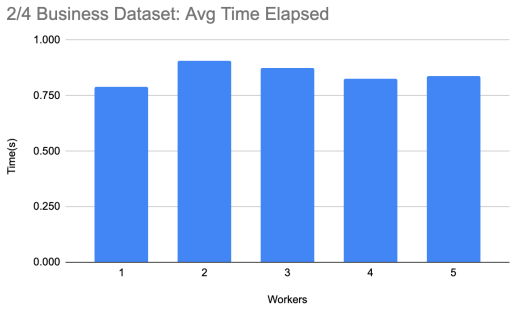
We use python and PySpark to process the data from the json file and store it as a collection in the MongoDB database. We processed each filter by aggregating the rating or review sentiment

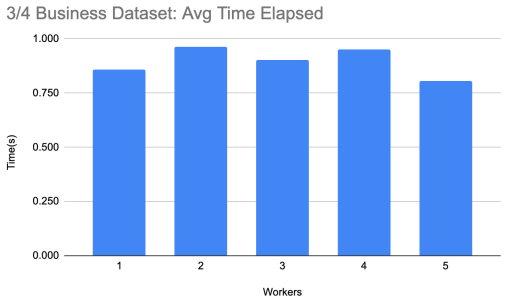
by column. For the category column (for the cuisine), we had to write an extra user function that picked the first cuisine that was listed for the aggregate to be grouped by.

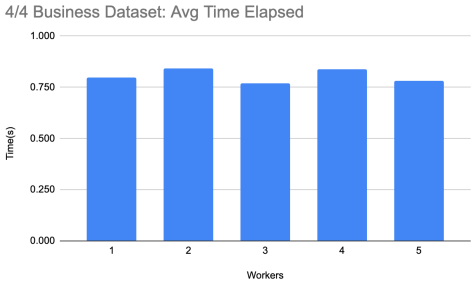
We use TF-IDF for the sentiment analysis to convert the textual data into numerical features we can train our model on. First, we use regular expressions to remove punctuation. Then apply Spark Stopwords remover and Tokenizer the review text. We train a linear Linear Support Vector Classification (SVC) to perform the classification.

Evaluation

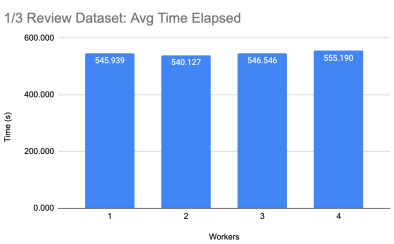
We can see below the difference in time when calculating the average rating by state with datasets of 4 different sizes utilizing a different number of workers each time.

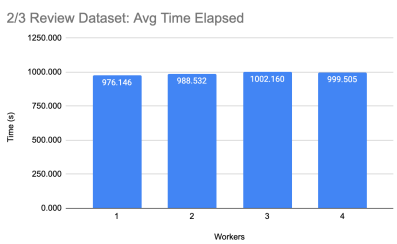


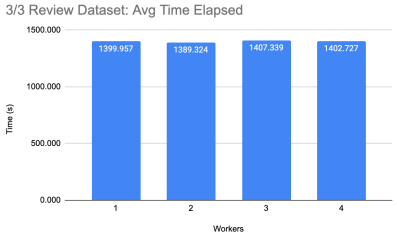




Likewise, here are the graphs for running the sentiment model on the review dataset.



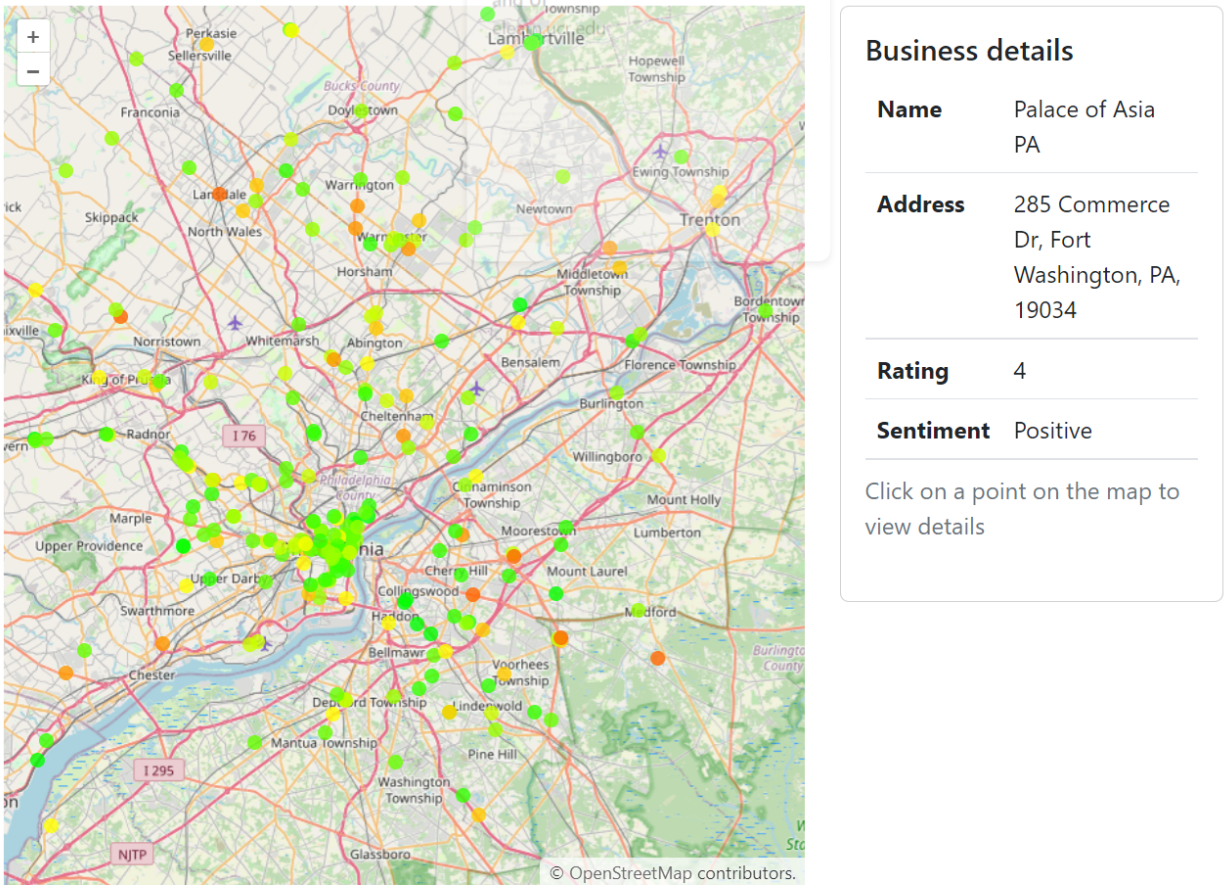




We believe the negligible difference in time between using a different number of workers and different dataset sizes is a result of the small dataset size. The small dataset size doesn't give spark enough time to showcase its ability to speed up processes.

With the larger dataset, we still do not see much difference between increasing the number of workers. Perhaps this is due to the all the wide dependency required when shuffling the partition to perform calculations for loading and store to mongoDB. However, we do see the effects of size of the dataset which increases elapsed time to run the script proportional to the dataset size.

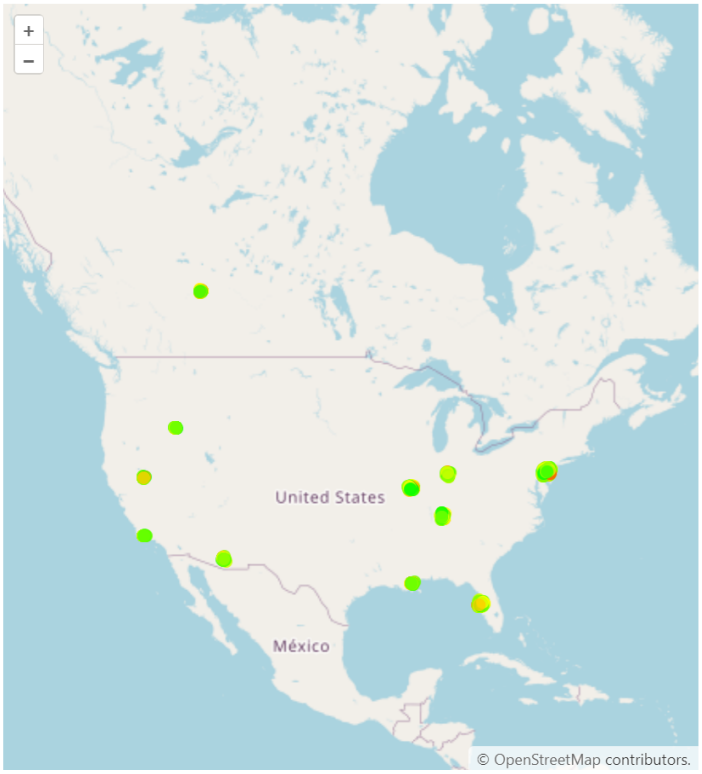
Screenshots



Business details

Name	Palace of Asia PA
Address	285 Commerce Dr, Fort Washington, PA, 19034
Rating	4
Sentiment	Positive

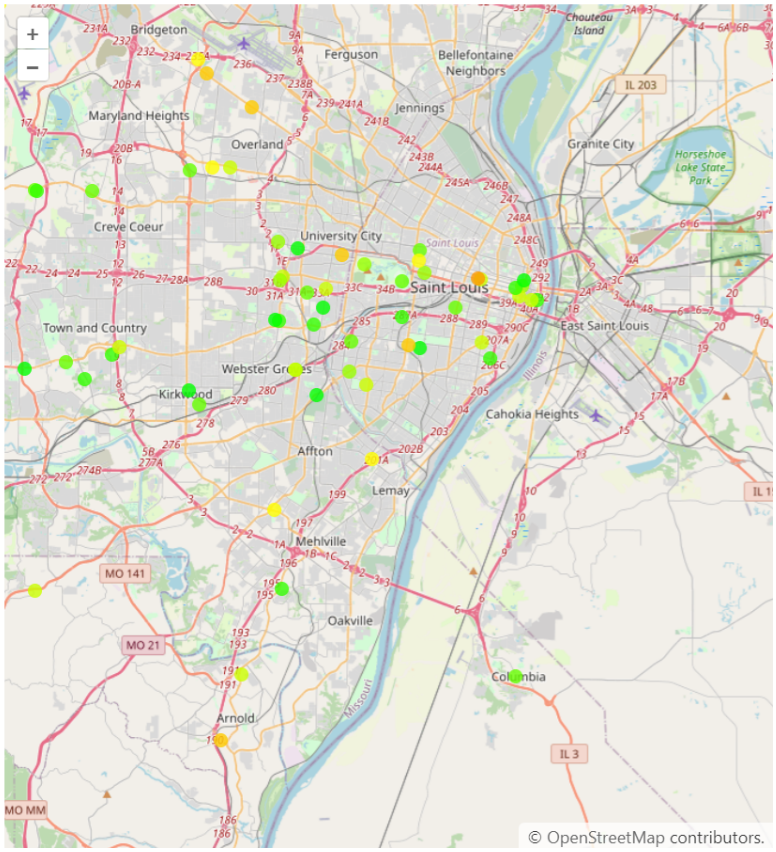
Click on a point on the map to view details



Business details

Name	
Address	
Rating	
Sentiment	

Click on a point on the map to view details



Business details

Name	Subway
Address	1 Brookings Dr, Saint Louis, MO, 63105
Rating	2
Sentiment	Negative

Click on a point on the map to view details

Contributions

Vinayak	Worked on front-end interface, sentimental analysis model, presentation
Rayyaan	Part 2 data processing contributions, presentation, write-up, setting up ec2 with required packages
Kevin	Part 2 contributions, presentation, sentiment analysis model, elapsed time data collection
Anthony	Part 2 contributions, Part 3 contributions, presentation, report