SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

School of Computing Science
Professional Master in Big Data
Instructor: **Greg Baker**

Project Report

CMPT 732: Big Data Lab 1, Fall 2019

Group: **Yelp Analysis**
Zhixuan Chi, zca92
Hengzhi Wu, hwa137
Nguyen Cao, nguyenc
Madana Krishnan V.K, mvadakan

Contents

1. Problem

We are living in the social networking world where individuals are encouraged to leave written reviews, star ratings and sharing photos of their experience about businesses they visit. In fact, some businesses have become popular within a very short time because a famous influencer on the Internet gave them a single positive review. Therefore, we are interested in answering the question: **Have popular businesses been reviewed positively by some popular users?**

We believe that answering the above question is valid and useful for business owners as well as end users because it would bring the following business values:

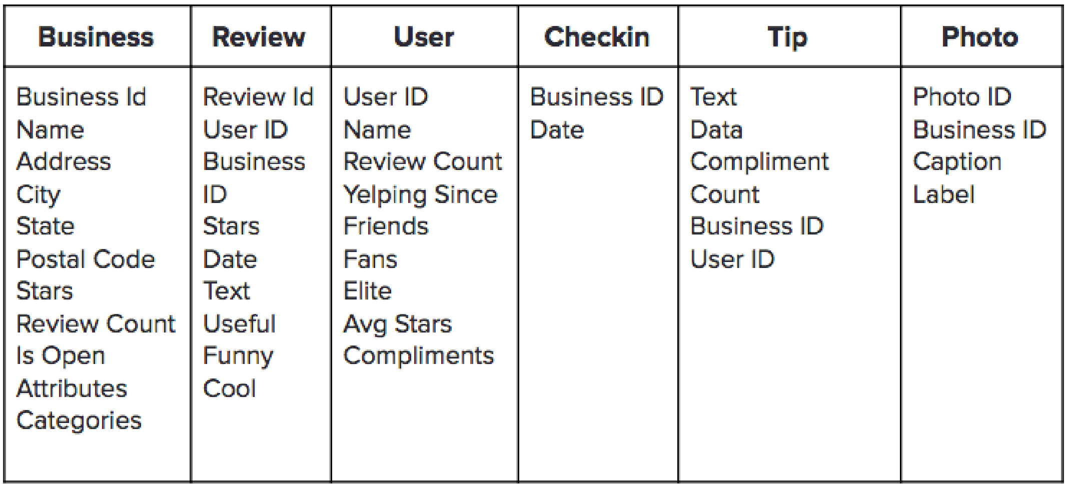
- For business owners, they would be interested in knowning whether reviews of well-known influencers would help making positive impact to their businesses.
- For end users, knowing which businesses have been reviewed positively by famous, trusted people would help them make better choices for their needs.

2. Methodology

Our approach to answer the question of interests is to use publicly available **review datasets**, analyze such data using our computation resources in order to find the **correlation** between *how popular a business is* and *the number of popular users have given positive reviews* about that business.

2.1. Datasets

Yelp Dataset: a public review dataset with about **8GB** in size and distributed as structured JSON files. Two biggest files are about users and reviews which is about 2GB and 5GB corerspondingly. The following table provides an overview of the structure of each file.



Business	Review	User	Checkin	Tip	Photo
Business Id Name Address City State Postal Code Stars Review Count Is Open Attributes Categories	Review Id User ID Business ID Stars Date Text Useful Funny Cool	User ID Name Review Count Yelping Since Friends Fans Elite Avg Stars Compliments	Business ID Date	Text Data Compliment Count Business ID User ID	Photo ID Business ID Caption Label

- In order to answer the question, we will only need **Business**, **User** and **Review** data.
- We also use **Checkin** as a feature when defining how popular a business is.

TripAdvisor & Google Reviews: two additional online services we use to search for the ratings and reviews of about 100 restaurants manually extracted from Yelp data. The collected data is used to learn automatically the weights of business popularity model.

2.2. Data Modeling

Popular Business: a business is considered popular depending on its **stars** ratings, the **number reviews** it got, whether it is still **opening**, whether its reviews are **useful**, **funny**, **cool** and got many **checkins**. These data is available from **Business**, **Review** and **Checkin** files of Yelp dataset. Each business will be associated with a score measuring how popularity it is based on these defining features.

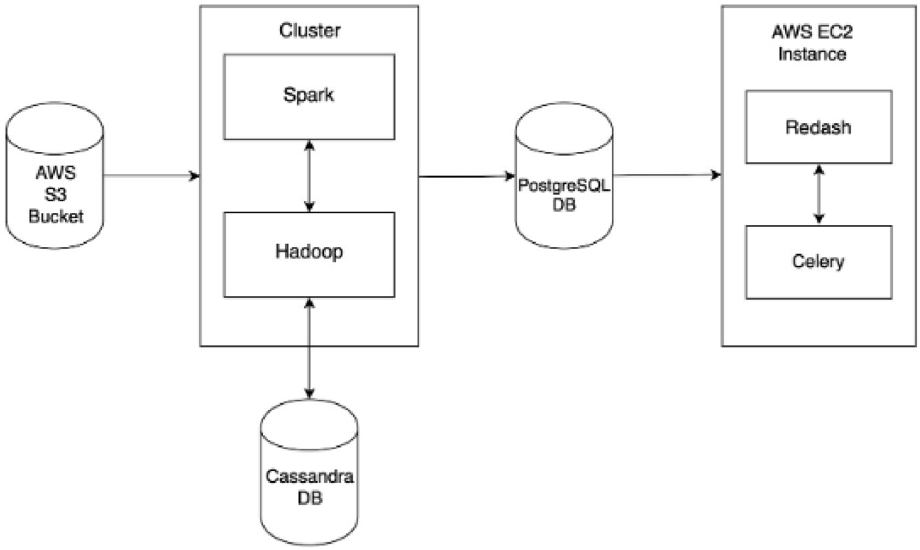
Popular User: a user is considered popular depending on his/her **elite** status, the number of **fans**, **friends**, the number of **reviews** he/she made, and how many **compliment** he/she has received. These data is available from **User** file of Yelp dataset. Each user will be associated with a score measuring how popularity he/she is based on these defining features.

We want to measure how many popular users have given positive reviews to those popular businesses. The question of interest is answer by computing the **correlation** between two ordered lists:

- The ordered list of businesses with values as their **popularity scores**, sorted by their popularity scores.
- The ordered list of businesses with values as their **number of popular users given positive reviews**, sorted by the businesses' popularity scores.

3. Implementation

3.1. System Architecture



- AWS S3 bucket is used to stored raw Yelp data JSON files.
- Spark and Hadoop setup in our lab cluster help in performing large-scale analysis on the data.
- Cassandra is used as a temporary data store for storing intermediate data between analysis steps as well as served as faster data access for Spark jobs compared to sequential data access of HDFS files.
- PostgreSQL DB is setup on an AWS EC2 instance to store results of analysis of Spark jobs. Since the results would not be large, so we do not worry much about the scalability of PostgreSQL DB.

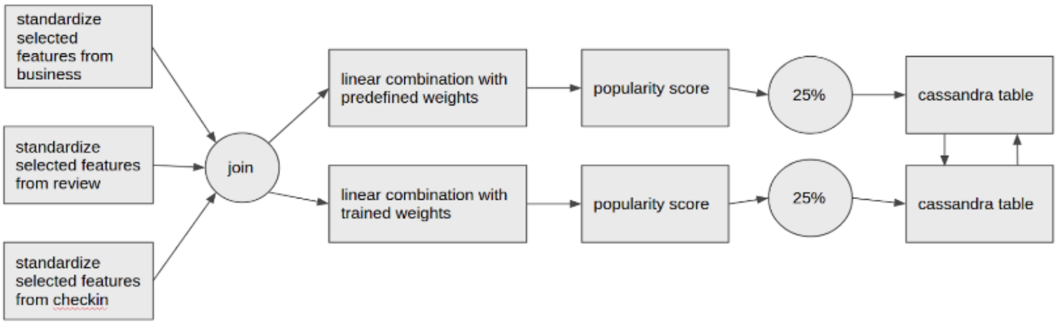
- Redash, which is hosted using another AWS EC2 instance is used for visualizing the analysis in the form of figures, charts, and tables.
- Celery works as a group of task executors in order to perform data query against Redash requests.

3.2. ETL

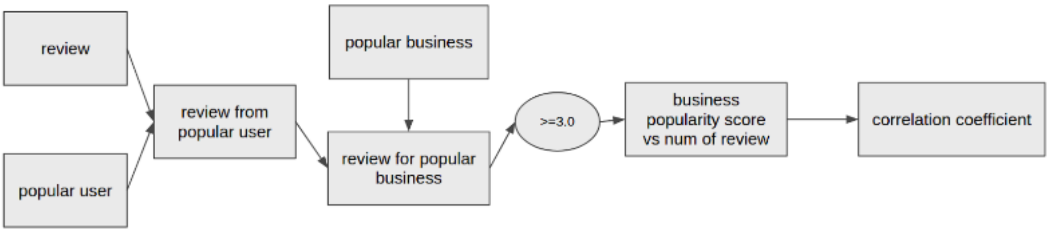
- A HDFS command copies only needed JSON files from AWS S3 bucket and stores into Hadoop HDFS.
- Spark jobs load copied JSON files from Hadoop HDFS and stores into Cassandra tables.
- Spark jobs load data from Cassandra tables to build popular businesses, popular users and store them into corresponding Cassandra tables. The features from businesses and users tables are normalized before performing analysis to ensure that the values among all the features are in a similar case.
- Spark jobs load result data from Cassandra tables and store into Postgres DB for visualization using Redash

3.3. Data Analysis

- Each business/user is associated with a popularity score based on several features in the Data Modeling section. The analysis process is conducted as in the below diagram.

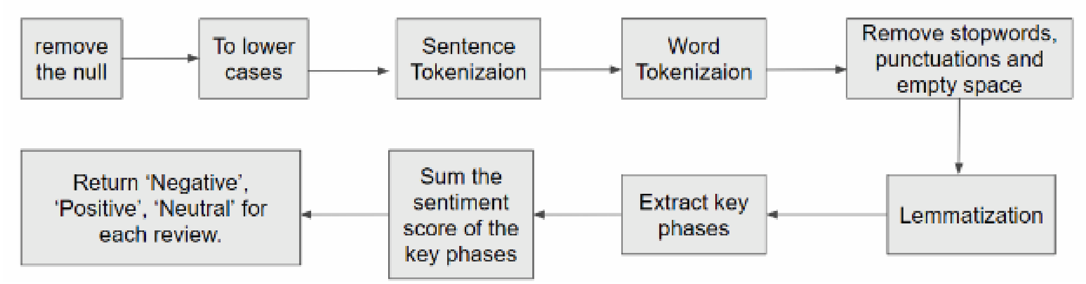

standardize selected features from business
standardize selected features from review
standardize selected features from checkin
join
linear combination with predefined weights
linear combination with trained weights
popularity score
popularity score
25%
25%
cassandra table
cassandra table

- Base on review data, we measure how many popular users have reviewed positively (with rating larger than certain value, for example 3) to a business. Then the correlation is computed based on the business popularity score and its corresponding number of popular users have reviewed it positively



3.4. Advanced Analytics

- **Sentiment Analysis of User Reviews**: We have done more feature engineering to get better popular business model by using NLP on text review data. We used NLTK with SparkML to assign each review into three categories: **Positive**, **Negative** and **Neutral** and use that as additional feature in our business popularity model and other analysis.


remove the null
To lower cases
Sentence Tokenizaion
Word Tokenizaion
Remove stopwords, punctuations and empty space
Lemmatization
Extract key phases
Sum the sentiment score of the key phases
Return 'Negative', 'Positive', 'Neutral' for each review.

- **Regression Model to learn business popularity weights**: We also want to learn the weights of features in our business popularity model automatically instead of manually assigning values to these features. To do that, we use other data about 100 restaurants in Toronto collected from TripAdvisor and Google Reviews. This data is served as label to regression model with input data from our corresponding Yelp data.

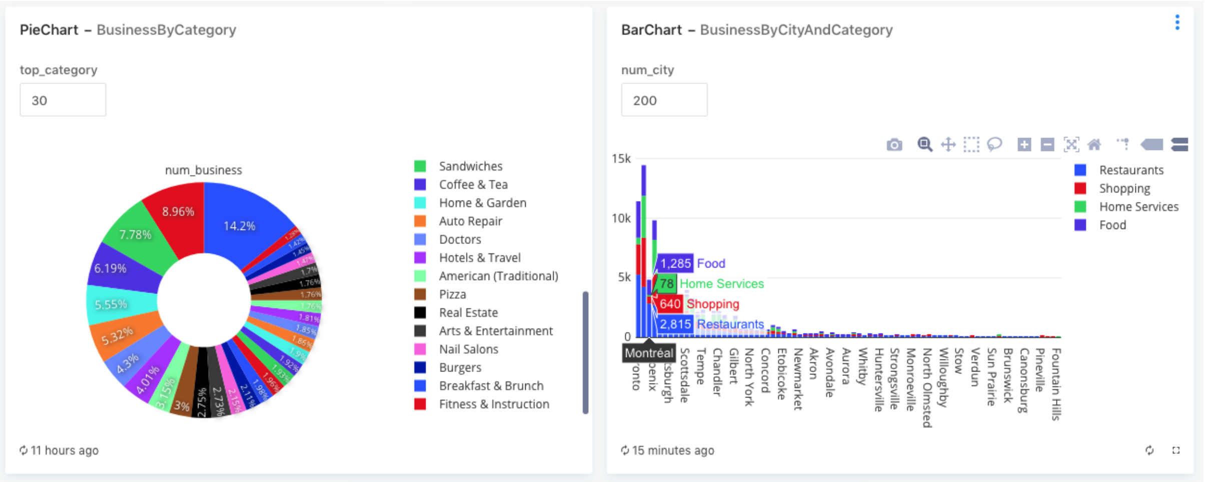
4. Results

4.1. Business Insights

Live Dashboard (click to view online): Yelp Business

We are able to get general insights about businesses from Yelp dataset

- Total number of businesses in which how many of them are actively opening and have been closed.
- Top cities of active businesses.
- Top categories of active businesses.
- Top cities with their corresponding categories of active businesses



PieChart – BusinessByCategory

top_category

30

BarChart – BusinessByCityAndCategory

num_city

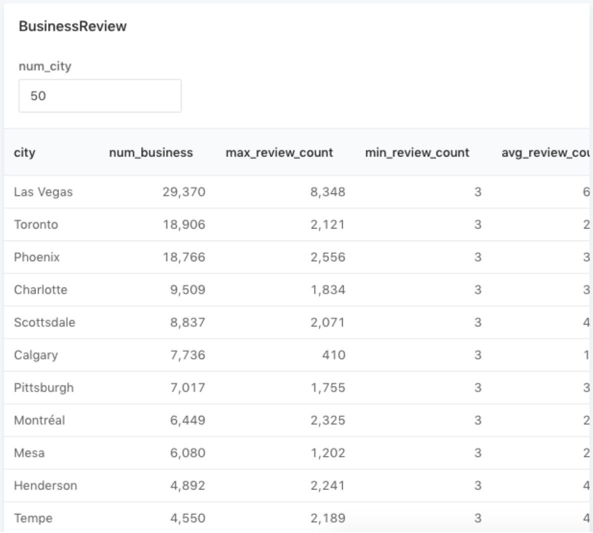
200

4.2. Review Insights

Live Dashboard (click to view online): Yelp Review

We are able to get review insights from Yelp dataset

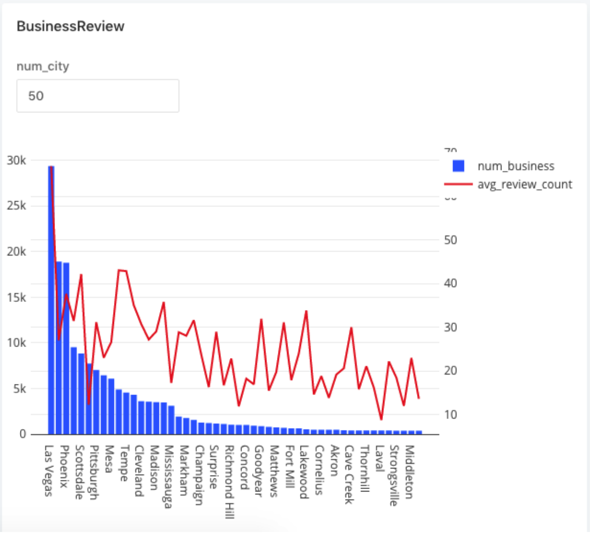
- Top cities based on the averaged number of reviews.
- Top categories with averaged number of postive, negative and neutral reviews.



BusinessReview

num_city: 50

city	num_business	max_review_count	min_review_count	avg_review_cou
Las Vegas	29,370	8,348	3	6
Toronto	18,906	2,121	3	2
Phoenix	18,766	2,556	3	3
Charlotte	9,509	1,834	3	3
Scottsdale	8,837	2,071	3	4
Calgary	7,736	410	3	1
Pittsburgh	7,017	1,755	3	3
Montréal	6,449	2,325	3	2
Mesa	6,080	1,202	3	2
Henderson	4,892	2,241	3	4
Tempe	4,550	2,189	3	4



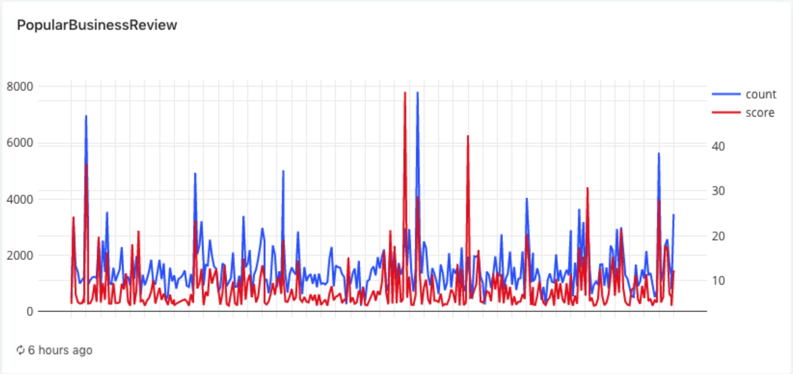
BusinessReview

num_city: 50

4.3. Business Review Insights

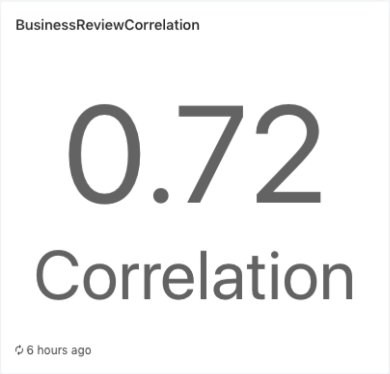
We are able to get business review insights from Yelp dataset

- Highly recommendated businesses are reviewed positively by popular users.
- Top categories show that there is even higher correlation between highly recommendated businesses and their positive reviews by popular users.



PopularBusinessReview

6 hours ago

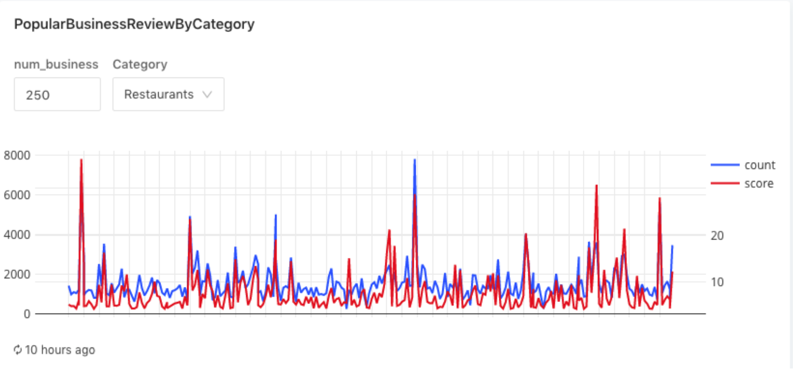


BusinessReviewCorrelation

0.72

Correlation

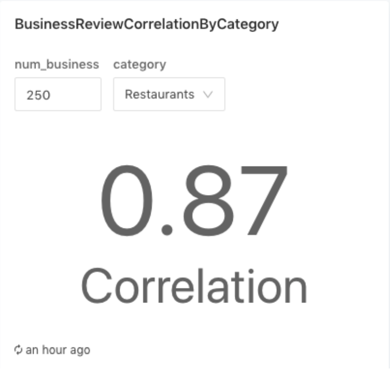
6 hours ago



PopularBusinessReviewByCategory

num_business: 250 Category: Restaurants

10 hours ago



BusinessReviewCorrelationByCategory

num_business: 250 category: Restaurants

0.87

Correlation

an hour ago