Team 8

Final Project Report

Yelp Review Analysis in Spark

Connor Jones, Nagavineela Mellachervu, Nilabh Sinha,
Young Jae Woo, Zhen Wu
May 1st, 2018

Contents

Abstract

Consumer review websites, such as Yelp, have become popular and trustworthy enough that consumers rely on them to make purchase decisions. By taking publicly available data provided by Yelp. Inc on Kaggle, we applied scalable data analysis techniques to get insight on the word frequency distribution of Yelp reviews. The main tool we used is Spark. Our study based on restaurant reviews and focused on review length, top-10 words and sentiments We studied the overall review analysis, review analysis by rating, by business type and then by user interaction. Our result shows that: 1. The ratio of positive words to negative words decreases significantly as the star rating increase. 2. “Food” and “service” are two main dimensions that customers judge restaurants. 3. People tend to express positive sentiments in their reviews. 4. As star rating increases, the average review length decreases.

Problem Statement

Understanding customer preferences is critical when it comes to restaurant operations. When a business has the capital to expand or renovate, they have the opportunity to address customer complaints, expand on strengths, and improve the customer experience overall. Without customer data, a restaurant business is spending is not making the best use of its limited capital. Analyzing customer data is necessary for restaurant businesses to act strategically.

Traditionally, customer data was gathered by word-of-mouth. Although this approach seems genuine, there are shortcomings. A restauranteur may only get honest feedback from the establishment's most devoted patrons. The customers who have bad experiences may choose avoid confrontation and leave their feedback unvoiced.

But in the age of social media, customers can leave reviews online without social pressure or fear of negative response. Almost every person in the United States has access to smartphones, so there are few barriers to leaving a review online.

However, with the sheer volume of reviews being submitted every day, it can be difficult to impossible to read all of them. Although sites like Yelp offer simple aggregated star ratings, they do not show the full picture of what a restaurant's strengths and weaknesses are.

We seek to use Spark to analyze the text of Yelp restaurant reviews to get a global view of business strengths and weaknesses. We will first do this by counting words by review rating and restaurant cuisine type. We will then add up positive and negative words for each group to get an idea of the sentiment being expressed for each grouping.

Dataset Overview

The dataset from Kaggle is a subset of Yelp's businesses, reviews, and user data.

It contains 7 CSV files. We used 2 of them for our project.

Yelp_review.csv

This file contains around 5,200,000 user reviews from yelp.

yelp_review.csv (3.53GB)								
Review_id	user_id	business_id	stars	date	text	useful	funny	cool
vkVSCC7xljjrAI4UGfnKEQ	bv2nCi5Qv5vroFiqKGopiw	AEx2SYEUJmTxVVB18LICwA	5	5/28/2016	Super simple place but amazing nonetheless. It's been around since the 30's and they still serve the same thing they started with: a bologna and salami sandwich with mustard. Staff was very helpful and friendly.	0	0	0

Yelp_business.csv

This file contains information on 174,000 businesses.

yelp_business.csv(30.2MB)											
business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories
Gu-xs3NIQTj3Mj2xYoN2aw	"Maxim Bakery & Restaurant"	"9665 Bayview Avenue, Unit 1-4"	Richmond Hill	ON	L4C 9V4	43.8675648	-79.4126618	3.5	34	1	French;Food;Bakeries;Restaurants

Summary of data fields:

Business Data:

- business_id—a unique identifier for business locations listings on Yelp.
- name—the colloquial name of the business.
- address—the street address of the business.
- stars—refers only to the overall star rating for the business.
- review_count—the number of people that have reviewed this business.
- categories—a semicolon separated string that contains a class description of the business.

Review Data:

- review_id—a unique identifier for every review on Yelp.
- user_id—a unique identifier for every user on Yelp.
- business_id— the unique identifier for the business that the review is associated with.
- stars—this stars field refers to the star rating of the review itself.
- text—a string of the review text itself.
- useful/funny/cool— an interaction field between users.

Preliminary Work

The data is available on Kaggle in a comma separated value format. If a user attempts to load this file into Spark without any preprocessing, Spark will interpret each '\n' character as a line break and each comma as a new split. With this parsing, any analysis is impossible because each "row" will have a different length.

There is a fix to this problem available in Spark 2.0. However, we wanted to utilize the Analytics Research Cluster in order to run Spark in cluster mode. Therefore, we went about some minor pre-processing using Python.

We loaded the file into Python using pandas.read_csv, which is able to parse the difference between '\n' in text fields and true line breaks. It can also parse the distinction between commas for grammatical purposes in text fields and commas for delimiting values in csv files.

While the data was loaded in Python, we decided to perform the initial join operation in Python. We used a simple "merge" function on the field "business_id" as an inner join. We only kept the following columns:'business_id','name','stars_business','categories','review_id','stars_review','text','useful','funny','cool'.

Once we obtained the merged, cleaned dataset, we wrote it to a csv file using the delimiter "^", which is a better delimiter for this dataset. We, then, uploaded this data to HDFS on the Analytics Research Cluster to do the Spark analysis.

In the following chapter, we discussed our Spark analysis approach in detail.

Spark Analysis

First, we uploaded the combined review/business csv file into Spark as an RDD and used the split function in order to separate each line into a list of 10 distinct elements, which are more easily analyzed.

As another preliminary step, we set out to remove punctuation and stopwords from the text field of the dataset. Punctuation would interfere with the word frequency count analysis because a string such as "food," would be counted separately from the string "food." During this step we also changed each character to its corresponding lowercase character. This step was performed in order to ensure strings such as "Food" were counted along with strings such as "food."

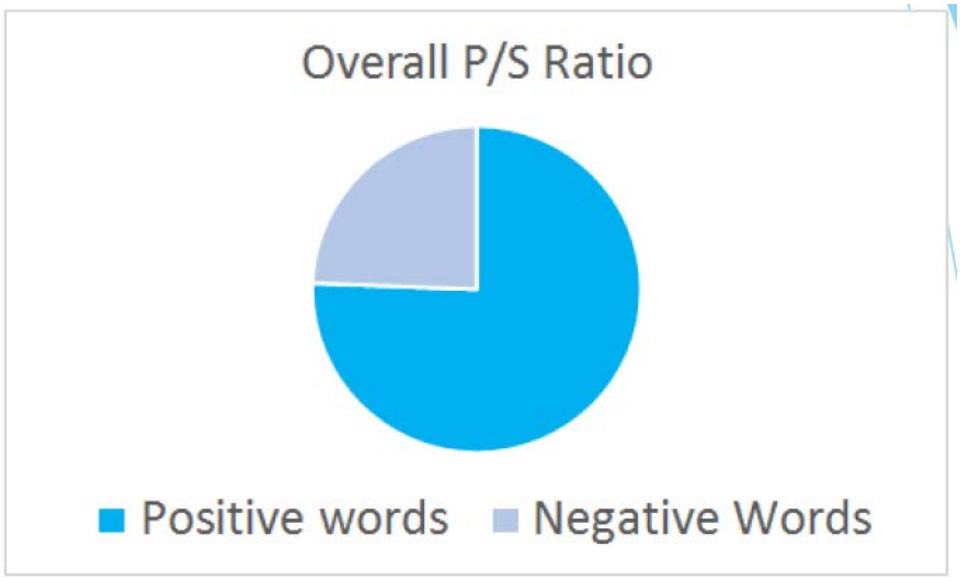
Similarly, we removed meaningless filler words such as "as," "I," and "you." These words are called "stopwords." We used a fairly standard set of stopwords from Python's Natural Language Toolkit (NLTK) package.Each of these operations (altering to lowercase, removing punctuation, and deleting stopwords) was performed with a simple Spark map function. In order to perform analysis by segment, we created several subset RDDs using Spark's filter functions. We then performed the analysis on each RDD in order to get our results.

We used the joined table of yelp_business and yelp_review to generate insights from the customer rating and comments. After reading the data into RDD, we created a (key, value) pair with key as star rating and value as words of each comments. Now, count the frequency of each (key, value) pair and store it in form ((key, value), frequency). Reformulate the key, value pair as (rating, (frequency, word)) so that to group the data based on star rating. In last, sort the values based on frequency and take the top 10 results for each star rating and print the result.

In our dataset, reviews are divided into 3 categories of Cool, Useful and Funny. We have found out the average word count and also the top 10 words for each of these review type. For this analysis we have used the Review csv file which underwent a lot of preprocessing steps like removal of punctuations , stopwords etc . Now, created an RDD containing (Key , value) pair with key as review type and word and value as the number of its occurrences in the review . After taking the data as (Key, Value) pair we have we got the exact frequency for each for different review types, then sorted the words based on their number of occurrences and found out the most repeating words for each of Cool, useful and funny reviews and print the results.

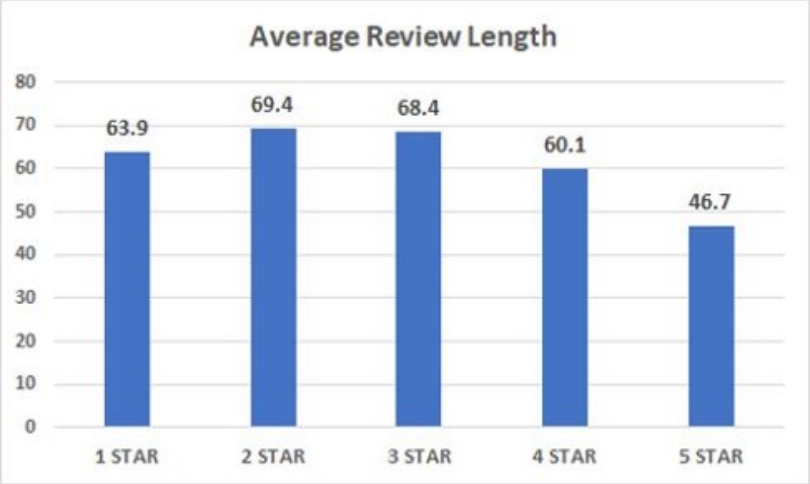
Results

1. Overall
    a. Average Word Count: **57.39**
    b. Top Words: food, good, place, great, service, like, one, time, get, back
    c. Ratio of Positive Words to Negative Words: approximately **3:1**

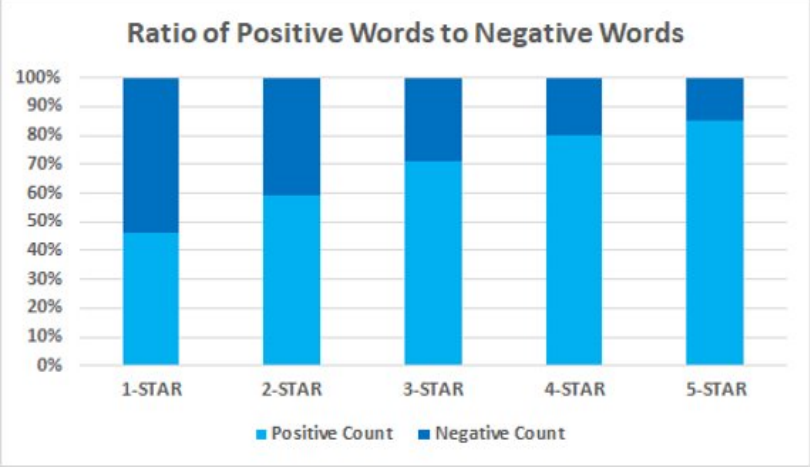


2. By Rating
    a. Average Review Length & Top 10 Words

Category	Average Review Length	Top Words
1-Star	63.9	food, place, service, like, one, us, back, order, get, time
2-Star	69.4	food, good, place, like, service, one, would, time, get, ordered
3-Star	68.4	good, food, place, like, service, would, really, one, get, great
4-Star	60.1	**good**, food, place, **great**, like, service, really, one, get, time
5-Star	46.7	food, **great**, place, **good**, service, **best**, like, one, time, back



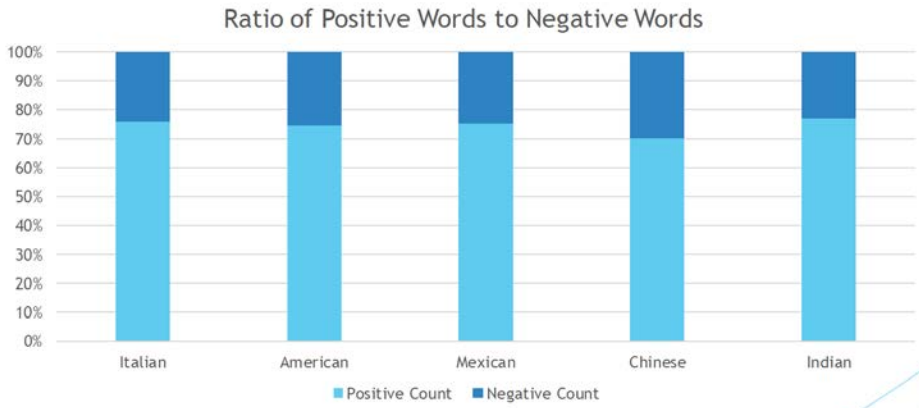
    b. Ratio of Positive Words to Negative Words



3. By Restaurant Type
    a. Average Review Length & Top 10 Words

Rating	Average Review Length	Top Words
Italian	61.0	pizza, food, good, place, great, service, like, one, time, back
American	59.4	food, good, place, great, service, like, one, time, back, get
Mexican	51.3	food, good, place, great, tacos, service, like, mexican, one, back
Chinese	57.2	food, good, place, chicken, chinese, like, service, great, rice, one
Indian	56.2	food, indian, good, chicken, place, service, great, like, restaurant, buffet

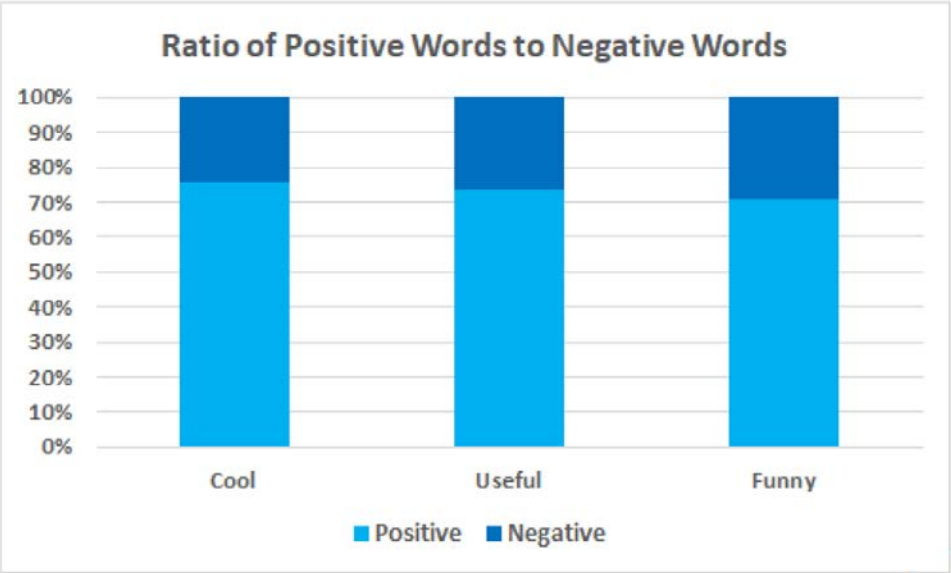
    b. Ratio of Positive Words to Negative Words



4. By cool, useful, funny
    a. Average Review Length & Top 10 Words

Rating	Average Review Length	Top Words
Cool	77.9	Tortillas , made, spicy, cons, cabinets, milestones , right, lived, slice, fee
Useful	73.6	television, pastas, ironically , adorned, ceiling , begin, notice , shut, picadillo ,warmed
Funny	82.4	response, meal, complain ,oily, options, thinks, side, arm, switching, find

b. Ratio of Positive Words to Negative Words



Conclusion

In our study, we based on restaurants reviews and focus on three aspects: the review length, the top-10-word count, and the sentiment analysis. Before doing the analysis, we have some assumptions. Our study verifies some of them, at the same time give us some surprise.

1. One of our assumption is that as star rating increases, the ratio of positive words to negative words should decrease. This suppose is verified by our analysis. The ratio of positive words to negative words decreases significantly as the star rating increase.
2. The other assumption is that when customers go to a restaurant, what he or she cares most should be delicious food and good service. This suppose is also verified perfectly by our analysis. Actually, the words “food” and “service” are consistently among the top words for each group, indicating that those are the main dimensions that customers judge restaurants.
3. The third assumption is the positive and negative words should appear relatively even. But from our results we can see that positive words are approximately 3 times than the negative words. It means that people tend to express positive sentiments in their reviews.
4. At last, we assumed that 1-star and 5-star may have the longest review because when a customer is very satisfied or disappointed with a restaurant, he/she may write a relatively long review to express the strong emotion. But in contrast to our suppose, the 5-star reviews have the shortest average length while the 2-star and 3-star reviews have relatively long average length. Overall, as star rating increases, the average review length decreases.

References

1 Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com, HBS. Boston, MA: Harvard Business School Publishing.

2 Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 04*. doi:10.1145/1014052.1014073

3 Kaggle Yelp Dataset: https://www.kaggle.com/yelp-dataset/yelp-dataset

Appendix

Opinion Lexicon is included in the file as ‘negative-words.txt’ and ‘positive-words.txt’