CIS 4560 - Yelp Dataset Term Paper

Team 3 - Michael Brumwell, Marco Rodriguez, & Chris Alipio
Department of Information Systems, California State University - LA
Los Angeles, CA

Abstract: This document goes into detail about how we used Yelp's data set from the Yelp Dataset Challenge. We used Hadoop and Pig to import and create the data structures for the data to analyze user sentiment the lowest five rated restaurants.

1. Introduction

We based our project on user sentiment because we wanted to get a better understanding of the data behind the reviews of highly rated and lowly rated restaurants. For restaurants that are looking to improve, they can use the data from their reviews to identify any potential problems or areas of improvement. Through analysis of user sentiment, restaurants can use key points as guides to improve themselves. Sentiment analysis is a great tool because it allows analysis of text. In this case, we analyze text reviews of restaurants on Yelp to find words that can yield inferences on what aspects of a restaurant customers find dissatisfactory.

1.1 Apache Hadoop

For this project, we will be using the Apache Hadoop platform, which is a suite of open-source software developed for handling large amounts of data using a network of computers. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using a thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was developed based on the Google File System (GFS) and Google's Map Reduce algorithm. Hadoop operates on the MapReduce algorithm, which processes data in parallel by splitting it and distributes the workload throughout different nodes. [2]

1.2 Apache Pig

Included in the Hadoop platform is Apache Pig. It is a high level platform that allows users to create programs that run on Hadoop's MapReduce. It is similar to SQL for relational database systems. [3]

1.3 AFINN Dictionary

The AFINN lexicon is a list of English terms manually rated for valence with an integer between -5 (negative) and +5 (positive) by Finn Årup Nielsen between 2009 and 2011.[4] This is crucial for extracting any inferences on different aspects of a restaurant

```
-bash-4.2$ cat AFINN.txt
abandon -2
abandoned       -2
abandons        -2
abducted        -2
abduction       -2
abductions      -2
abhor   -3
abhorred        -3
abhorrent       -3
abhors  -3
abilities       2
ability 2
aboard  1
absentee        -1
absentees       -1
absolve 2
absolved        2
absolves        2
absolving       2
absorbed        1
abuse   -3
abused  -3
abuses  -3
abusive -3
accept  1
accepted        1
accepting       1
accepts 1
accident        -2
accidental      -2
accidentally    -2
accidents       -2
accomplish      2
accomplished    2
accomplishes    2
accusation      -2
accusations     -2
accuse  -2
accused -2
accuses -2
accusing        -2
ache    -2
achievable      1
aching  -2
acquit  2
acquits 2
acquitted       2
acquitting      2
acrimonious     -3
active  1
adequate        1
```

Figure 1. Partial output of the AFINN dictionary

2. Problem Definition

Restaurants that are seeking to improve their service can use the reviews posted by yelp to identify areas of improvement. Yelp reviews provide a useful source of information that businesses can use to improve areas they're weak at or get ideas based on reviews from popular restaurants. Data analysis of user reviews is a strong way to find ways to improve.

3. Proposed Methodology

Our project will follow the following steps:

1.We will download and load into our EMR cluster the dataset provided by Yelp and the AFINN dictionary.

2. A script to extract the 5 lowest rated restaurants will be written.

3. Once those restaurants have been determined, we gather all reviews associated to those restaurants

4. We will compare all reviews to the dictionary and extract all negative words and their associated ratings.

4. Implementation and Results

All calculations were done on an Amazon EC2 instance with an Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz CPU running Amazon's AMI Linux image with control of an Amazon EMR cluster of 3 nodes Running Amazon's version of Hadoop version 2.8.5 with Pig version 0.17.0.

Yelp's dataset comprises of 6 json files listed below. For this project, we will only be using the 'business.json', 'review.json' and the 'user.json'

Figure 2. JSON files contained in Yelp dataset.

To upload them to the EC2 instance, we will use the scp command in the Windows command prompt as such. [5]

Figure 3. Usage of scp to upload a file

Now that all files are uploaded. Create relations using these files in Pig. We will begin by creating a relation using the business.json file and finding the 5 lowest rated restaurants in Toronto.

4.1 Determining the 5 lowest rated restaurants

To begin the analysis outlined in this project, we want to determine the lowest 5 rated restaurants in the city of Toronto. We will begin by loading the creating a relation in Pig out of the business.json file and performing ETL operations on it.

```
business = LOAD 'business.json' using JsonLoader('
        business_id:chararray,
        name:chararray,
        address:chararray,
        city:chararray,
        state:chararray,
        postalcode:chararray,
        latitude:float,
        longitude:float,
        stars:float,
        review_count:int,
        is_open:int,
        attributes:(GoodForKids:chararray),
        categories:chararray,
        hours:(day:chararray, hours:chararray)
');



low_rated_bus = FILTER business by stars <=2.0;
low_rated_city = FILTER low_rated_bus BY city == 'Toronto';
low_rated_categ = FILTER low_rated_city BY (categories matches '.*Restaurant*.');


foreach_business = FOREACH low_rated_categ GENERATE
        business_id,
        name,
        city,
        latitude,
        longitude,
        stars,
        categories;

business_ordered = ORDER foreach_business BY stars ASC;


limit_business = LIMIT business_ordered 5;


DUMP limit_business;

STORE limit_business into 'lowfive' using PigStorage(',');
```

Figure 4. Pig script to determine the 5 lowest rated restaurants in the city of Toronto.

```
(pFiHYTqH5t9cWqYN-NbQ-w,Pacific Wok And Grill,1571 Sandhurst Circle,Toronto,ON,43.80978,-79.26944,1.0,Chinese
, Restaurants)
(CsAj3Rsw7HuExJl4OG-iZw,Market@416,2 Eireann Quay,Toronto,ON,43.631813,-79.39785,1.0,Cafes, Restaurants)
(grmdcTqmUph_mi8rhTSuzw,Pastacceria,101 College Street,Toronto,ON,43.659927,-79.38866,1.0,Fast Food, Restaura
nts)
(NSe_Fsyb939ueLd8WhuLPg,Night Market,4850 Yonge Street,Toronto,ON,43.7626,-79.41151,1.0,Asian Fusion, Barbequ
e, Korean, Restaurants)
(rI_DhGggyQZwr0Rsh9bXTg,Fat Bastard Burrito Co,628 King Street W,Toronto,ON,43.644382,-79.40108,1.0,Fast Food
, Tex-Mex, Mexican, Restaurants)
grunt>
```

Figure 5. Results of the previous script

The results will be saved to a csv file that will be later used to perform the sentiment analysis. This file will be named lowFive.csv

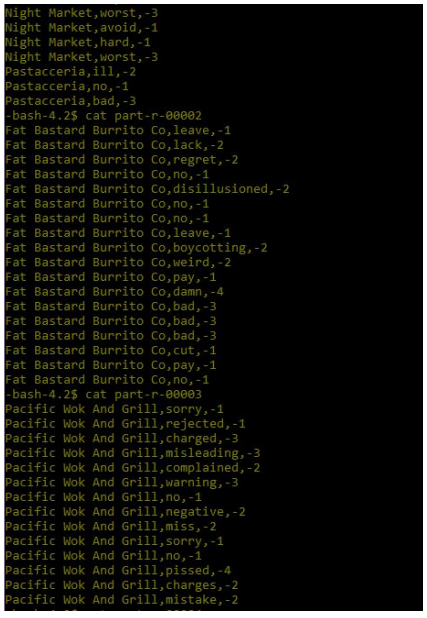
4.2 Performing Sentiment Analysis

Now that we have the 5 lowest rated restaurants, we will Perform ETL operations on the review.json, user.json, and lowfive.csv files. We will find all reviews that were made for these restaurants and analyze the text of the reviews with the AFINN dictionary of words. We will display all negatively rated words that can provide clues as to what a restaurant is doing wrong.

```
--LOADING USER.JSON, REVIEW.JSON, BUSINESS.JSON FILES INTO PIG
users = LOAD 'user.json' using JsonLoader('user_id:chararray,user_name:chararray,review_count:int,yelping_since:chararray,friends:
ars:chararray,compliment_hot:int,compliment_more:int,compliment_profile:int,compliment_cute:int,compliment_list:int,compliment_not
:int');

review = LOAD 'review.json' using JsonLoader('review_id:chararray,user_id:chararray,business_id:chararray,stars:chararray,cool:cha

lowfive = LOAD 'lowfive.csv' using PigStorage(',') AS (business_id:chararray,bus_name:chararray,city:chararray,latitude:float,long

-- LOADING AFINN WORD DICTIONARY FOR
dictionary = load 'AFINN.txt' using PigStorage('\t') AS(word:chararray,rating:int);



--SELECTING ONLY THE FIELDS THAT ARE NEEDED FORM THE LOWFIVE RELATION

foreach_business = FOREACH lowfive GENERATE business_id,bus_name;

--RELATION TO JUST HAVE THE USER ID AND NAME OF A USER FROM THE USERS RELATION

foreach_user = FOREACH users GENERATE user_id, user_name;


--RELATION TO JUST HAVE THE USER ID, BUSINESS ID, TEXT OF REVIEW, DATE OF REVIEW FOR A REVIEW FROM THE REVIEWS RELATION

foreach_review = FOREACH review GENERATE user_id, business_id, text, date;


--COMBINING BOTH ABOVE RELATIONS(USERS/REVIEW) TO HAVE A USER NAME ASSOCIATED TO A REVIEW

user_reviews = JOIN foreach_user BY user_id, foreach_review BY user_id;


--ELIMINATING REDUNDANT INFORMATION FROM PREVIOUS RELATION

fr_user_reviews = FOREACH user_reviews GENERATE foreach_user::user_name,
foreach_review::business_id, foreach_review::text, foreach_review::date;

--FR_USER_REVIEWS AND FOREACH_BUSINESS WILL BE THE RELATIONS THAT ARE JOINED

businesses_reviews = JOIN fr_user_reviews BY business_id, foreach_business BY business_id;


--ELIMINATING UNNECCESARY FIELDS FROM FINAL RELATION

toronto_busrev = FOREACH businesses_reviews GENERATE
    foreach_business::bus_name as bus_name,
    fr_user_reviews::foreach_user::user_name as user_name ,
    fr_user_reviews::foreach_review::text as text,
    FLATTEN(TOKENIZE(fr_user_reviews::foreach_review::text)) AS word,
    fr_user_reviews::foreach_review::date as date;

word_rating = JOIN toronto_busrev BY word left outer, dictionary BY word using 'replicated';

word_rating_filter = FILTER word_rating BY dictionary::rating <0;

low_rated_words = FOREACH word_rating_filter GENERATE toronto_busrev::bus_name, dictionary::word, dictionary::rating;

STORE low_rated_words into 'lowRatedWords' using PigStorage (',');
```

Figure 5. Pig script to perform sentiment analysis.

As with the previous script, the results will be saved in a csv file. The results are shown below.



```
Night Market,worst,-3
Night Market,avoid,-1
Night Market,hard,-1
Night Market,worst,-3
Pastacceria,ill,-2
Pastacceria,no,-1
Pastacceria,bad,-3
-bash-4.2$ cat part-r-00002
Fat Bastard Burrito Co,leave,-1
Fat Bastard Burrito Co,lack,-2
Fat Bastard Burrito Co,regret,-2
Fat Bastard Burrito Co,no,-1
Fat Bastard Burrito Co,disillusioned,-2
Fat Bastard Burrito Co,no,-1
Fat Bastard Burrito Co,no,-1
Fat Bastard Burrito Co,leave,-1
Fat Bastard Burrito Co,boycotting,-2
Fat Bastard Burrito Co,weird,-2
Fat Bastard Burrito Co,pay,-1
Fat Bastard Burrito Co,damn,-4
Fat Bastard Burrito Co,bad,-3
Fat Bastard Burrito Co,bad,-3
Fat Bastard Burrito Co,bad,-3
Fat Bastard Burrito Co,cut,-1
Fat Bastard Burrito Co,pay,-1
Fat Bastard Burrito Co,no,-1
-bash-4.2$ cat part-r-00003
Pacific Wok And Grill,sorry,-1
Pacific Wok And Grill,rejected,-1
Pacific Wok And Grill,charged,-3
Pacific Wok And Grill,misleading,-3
Pacific Wok And Grill,complained,-2
Pacific Wok And Grill,warning,-3
Pacific Wok And Grill,no,-1
Pacific Wok And Grill,negative,-2
Pacific Wok And Grill,miss,-2
Pacific Wok And Grill,sorry,-1
Pacific Wok And Grill,no,-1
Pacific Wok And Grill,pissed,-4
Pacific Wok And Grill,charges,-2
Pacific Wok And Grill,mistake,-2
```

Figure 6. Results of the sentiment analysis.

5. Conclusions

As seen in Figure 6, a list of words found in reviews and their scores have been found. For example, a review for Pacific Wok and Grill find the word “misleading.” This can cause that establishment to investigate any discrepancies between it’s advertised menu and it’s actual service. Big Data analysis creates many possibilities that were not there with traditional data processing methods, such as relational databases.

References

[1] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[2] Apache Hadoop Retrieved from https://hadoop.apache.org

[3] Hadoop: Apache Pig Retrieved from http://pig.apache.org/

[4] AFINN Sentiment Lexicon Retrieved from http://corpustext.com/reference/sentiment_afinn.html

[5] How to Use SCP Command to Securely Transfer Files(2019) Retrieved from https://linuxize.com/post/how-to-use-scp-command-to-securely-transfer-files/

[] Tableau Software. (2019). Changing the way you think about data. Retrieved from https://www.tableau.com/

[] Yelp Inc. (2019). Yelp dataset challenge. Retrieved from https://www.yelp.com/dataset/challenge