Reddit Data Analysis Using NGram

Group 3: Daniel Kang, Klayton Trinh, Michael Kintanar, Danny Xie, Ryan Hathcock
Department of Computer Information Systems, California State University Los Angeles
CIS 4560-01 – Introduction to Big Data

Abstract: This document is intended to display the technologies, methods and practices for using NGram and other big data tools to analyze Reddit comment data. The document explains step by step how to retrieve the data that is to be used, upload it to Hadoop, to migrate the data into tables designed to be analyzed, the analysis of the data, and how to use Microsoft Power BI to create visualizations of the analysis.

1. Introduction

Reddit is a website that contains a massive network of communities built around people's interest. In 2020, Reddit was the 8th most visited website with 1.15 billion visits[1]. The main content on reddit are comments, where users can display their opinions on a multitude of subjects, news, and events in text form. Many analysts in various fields such as politics, sports, medicine, technology, finance, and more use social media comments as a way to gauge public opinion and emotion on various topics. Of course, Reddit is no exception. To be able to effectively use data from Reddit comments to produce patterns in public perception is incredibly useful when making decisions in various industries.

Because of the importance of analyzing comment data to gauge public perception, we sought a way to effectively analyze the massive amounts of data that Reddit comments produce in order to find relevant trends in commenting during May 2015. What we found was a significant trend of nonsensical comments being made that included multiple recurring phrases. This paper is intended to describe the approach we took in storing, analyzing, and visualizing a data set that contains Reddit comment data. We have chosen this data set because it contains the complete comment data of Reddit from May 2015. This means it is a complete, unbiased picture of Reddit from that time. This allowed us the versatility to use Big Data tools such as Hadoop, HiveQL, and Power BI in order to narrow down on any points of analysis we wanted.

2. Related Work

There have been many reports and studies based on analysing data from Reddit. In 2019 Adams, Artigiani, and Wish published an analysis that compared Reddit and Twitter data to determine how people talked about drugs[2]. They found a trend in demographic based results showing a positive correlation in the users of opiates and white males. In 2016 Itamar Shatz produced an academic journal stating that his research of Reddit data shows that the website is ideal for Big Data analysis because it is categorically separated into subjects and that from the age and race demographic results he got, Reddit can be an accurate representation of the United States population[3]. A report written by Singer, Ferrara, Kooti, Strohmaier, and Lerman used Big Data analysis to show evidence of significant technical performance decline in user sessions during April 2015[4].

While these reports are examples of using Big Data practices to analyze Reddit in order to produce results in drug use, further Big Data use, and technical flaws, our analysis differs quite a bit. While we did use Big Data tools, we sought to find trends in the comment data we analyzed. And once we found a unique trend of nonsensical comments with recurring phrases, we utilized our cloud-based Big Data approach to further investigate the cause of this interesting trend. Along with this, we focused our data analysis on determining the frequency of comment and scores of each comment based on the day of the month. We found a negative correlation on the amount of comments produced during the weekend.

4. Specifications

The data set we used, "May 2015 Reddit Comments"[5], is a total of 29.62GB of data. It is composed of Reddit comment data from May 2015, including many important variables to the comment entities such as author, date, subreddit, comment body, and more.

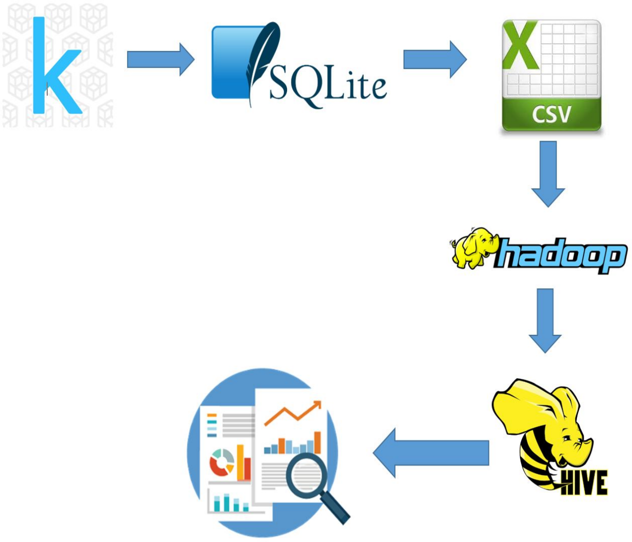
The platform we used to store this data is a distributed system running Hadoop on an Oracle Big Data Cloud Service Cluster. The specifications for the Cluster can be found below.

Data table 1 - Cluster Specification

Cluster Version	2.7.1.2.4.2.0-258
Number of Nodes	3
Cluster Size	802.6 GB
Memory Size	180 GB

Alongside Hadoop and Oracle Big Data Cloud Service Cluster, we used HiveQL for querying and storing into external data. Finally, we imported our analysis to Microsoft Power BI to create visualizations. Specifically, we used the NGram functionionality to query the most frequently used word combinations containing specific word count parameters. This provided us with insight on commonly used phrases within a specific range of word count.

4. Implementation Flowchart



The first iteration of the dataset is a sample of the full data released by Reddit, which totaled in over 1.7 Terabytes in size. Kaggle, the origin of the dataset we chose, took a sample of the 1.7 TB of data uploaded 29.62GB of data. From the download, we modified and cleansed the set further using SQLite. Once we had our modified data in a csv file, we compressed it into a zip file and hosted it on Amazon S3 so it could be accessed by SCP. We transferred the data to our Hadoop Filesystem stored on the Oracle Big Data cluster. From there, we were able to run our queries using Hive and store the results into external tables. Then, we imported the Hive data from the remote cloud-based Oracle cluster into our local machines where we were able to create visualizations in Power BI.

5. Cleaning and Importing Data

The dataset had to be modified to remove irrelevant data and to make it suitable for the csv file format. In Reddit, deleted comments are not removed; their contents are instead replaced with “[deleted]”. All rows containing this in the Body column were removed since it would skew any string frequency analysis without actually being user contributions. Also, Reddit has many moderators that are bots which submit automated comments while performing administrative tasks. These moderators have a specific value in the Distinguished column. All rows containing this value were removed since the automated messages use the exact same phrases and would skew any string frequency analysis as well.

Another thing to consider was that many fields in the Body column contained line breaks, and this caused some problems with the csv file format. Anything after a line break was dropped and it’s possible that the contents after the line break were put into different columns, but this was not verified. Regardless, replacing all line breaks with a space fixed most of the issues caused by exporting the csv file from SQLite and allowed the full contents of the Body column to appear.

The cleaned and fully functional csv file was compressed into a zip file and hosted on Amazon S3 for quick and universal access.

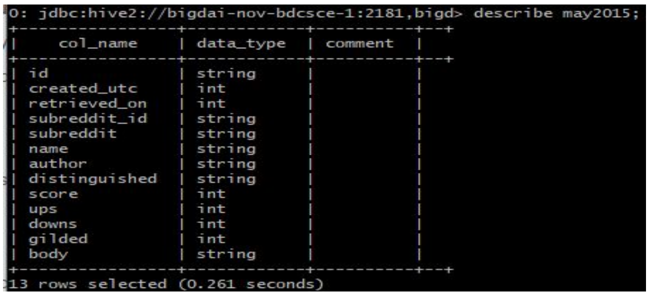
6. Analysis and Visualization

The initial goal was to analyze the frequency of entire phrases as a function of time, and a new Hive table was created like so:

```
create table test (
    day int,
    4gram
array<struct<ngram:array<string>,estfrequency:double>>
);
```

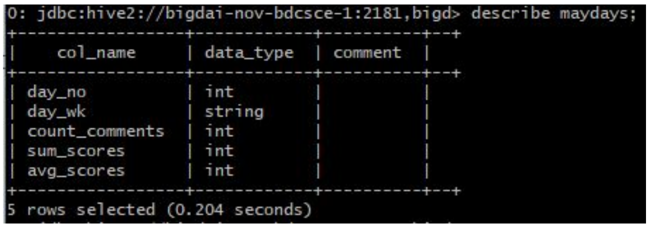
However, the table was unable to be filled because the results of the ngram function was not allowed to be used to fill a table. An error message came up saying that UDTF were not allowed in insert or create statements. So the original idea was scrapped and a simpler analysis had to be done.

The first step before beginning our search for recurring phrases was to organize tables based on the variables we intended to analyze. We created May2015, which was a table to store the important information of each comment.



```
0: jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigd> describe may2015;
+-----------------+------------+----------+--+
|    col_name     | data_type  | comment  |
+-----------------+------------+----------+--+
| id              | string     |          |
| created_utc     | int        |          |
| retrieved_on    | int        |          |
| subreddit_id    | string     |          |
| subreddit       | string     |          |
| name            | string     |          |
| author          | string     |          |
| distinguished   | string     |          |
| score           | int        |          |
| ups             | int        |          |
| downs           | int        |          |
| gilded          | int        |          |
| body            | string     |          |
+-----------------+------------+----------+--+
13 rows selected (0.261 seconds)
```

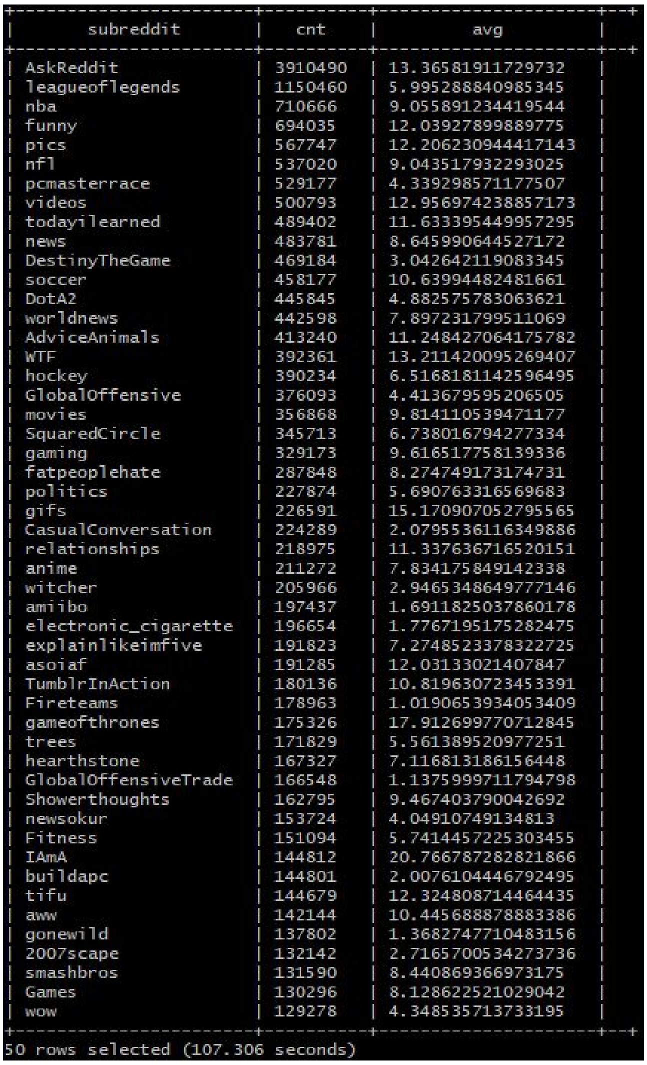
Then, we created MayDays. This was a table to store each day in May 2015 in order to visualize the activity on Reddit organized by date and time.



```
0: jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigd> describe maydays;
+-----------------+------------+----------+--+
|    col_name     | data_type  | comment  |
+-----------------+------------+----------+--+
| day_no          | int        |          |
| day_wk          | string     |          |
| count_comments  | int        |          |
| sum_scores      | int        |          |
| avg_scores      | int        |          |
+-----------------+------------+----------+--+
5 rows selected (0.204 seconds)
```

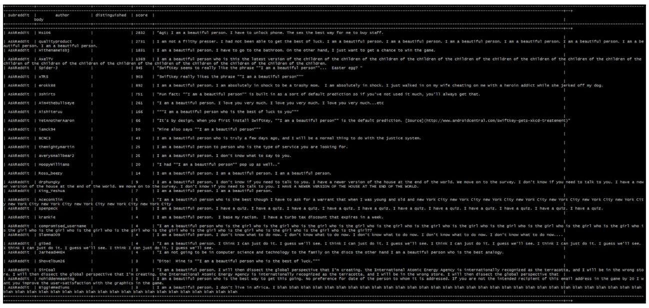
A second table titled MayDay_Export was created with the same signature as MayDay and was used to export the data to Power BI.

Once our initial structure was established, we began our inquiry on the top phrases used. We started by querying the top 50 Subreddits by number of total comments.



```
+------------------------+----------+------------------------+--+
|       subreddit        |   cnt    |          avg           |
+------------------------+----------+------------------------+--+
| AskReddit              | 3910490  | 13.3658191172973 2     |
| leagueoflegends        | 1150460  | 5.995288840985345      |
| nba                    | 710666   | 9.055891234419544      |
| funny                  | 694035   | 12.03927899889775      |
| pics                   | 567747   | 12.206230944417143     |
| nfl                    | 537020   | 9.043517932293025      |
| pcmasterrace           | 529177   | 4.339298571177507      |
| videos                 | 500793   | 12.956974238857173     |
| todayilearned          | 489402   | 11.633395449957295     |
| news                   | 483781   | 8.645990644527172      |
| DestinyTheGame         | 469184   | 3.042642119083345      |
| soccer                 | 458177   | 10.63994482481661      |
| DotA2                  | 445845   | 4.882575783063621      |
| worldnews              | 442598   | 7.897231799511069      |
| AdviceAnimals          | 413240   | 11.248427064175782     |
| WTF                    | 392361   | 13.211420095269407     |
| hockey                 | 390234   | 6.516818114259649 5    |
| GlobalOffensive        | 376093   | 4.413679595206505      |
| movies                 | 356868   | 9.814110539471177      |
| SquaredCircle          | 345713   | 6.738016794277334      |
| gaming                 | 329173   | 9.616517758139336      |
| fatpeoplehate          | 287848   | 8.274749173174731      |
| politics               | 227874   | 5.690763316569683      |
| gifs                   | 226591   | 15.170907052795565     |
| CasualConversation     | 224289   | 2.0795536116349886     |
| relationships          | 218975   | 11.337636716520151     |
| anime                  | 211272   | 7.834175849142338      |
| witcher                | 205966   | 2.9465348649777146     |
| amiibo                 | 197437   | 1.6911825037860178     |
| electronic_cigarette   | 196654   | 1.7767195175282475     |
| explainlikeimfive      | 191823   | 7.274852337832272 5    |
| asoiaf                 | 191285   | 12.03133021407847      |
| TumblrInAction         | 180136   | 10.819630723453391     |
| Fireteams              | 178963   | 1.0190653934053409     |
| gameofthrones          | 175326   | 17.912699770712845     |
| trees                  | 171829   | 5.561389520977251      |
| hearthstone            | 167327   | 7.116813186156448      |
| GlobalOffensiveTrade   | 166548   | 1.1375999711794798     |
| Showerthoughts         | 162795   | 9.467403790042692      |
| newsokur               | 153724   | 4.04910749134813       |
| Fitness                | 151094   | 5.741445722530345 5    |
| IAmA                   | 144812   | 20.766787282821866     |
| buildapc               | 144801   | 2.0076104446792495     |
| tifu                   | 144679   | 12.324808714464435     |
| aww                    | 142144   | 10.445688878883386     |
| gonewild               | 137802   | 1.3682747710483156     |
| 2007scape              | 132142   | 2.716570053427373 6    |
| smashbros              | 131590   | 8.440869366973175      |
| Games                  | 130296   | 8.128622521029042      |
| wow                    | 129278   | 4.348535713733195      |
+------------------------+----------+------------------------+--+
50 rows selected (107.306 seconds)
```

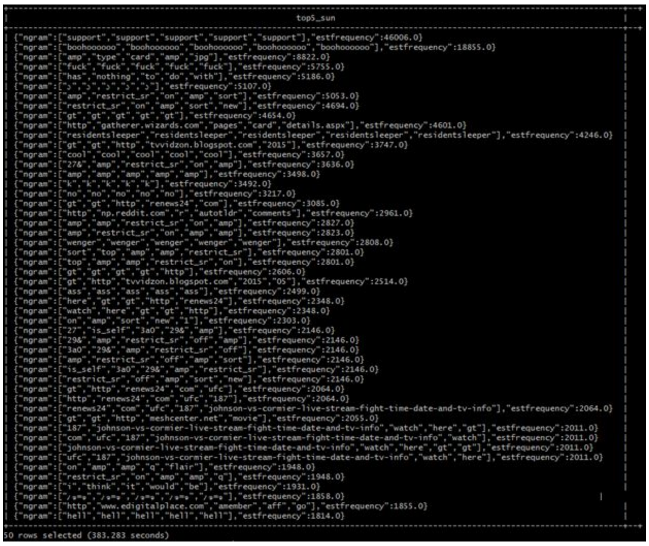
From those subreddits, we searched for common phrases containing 3, 4, 5 words using NGrams. We found a trend that was very interesting. There are multiple nonsensical phrases that belonged to comments where the phrases had no context. Below is an example.



```
+------------+--------------------+----------------+--------+
| subreddit  |       author       | distinguished  | score  |
       body
+------------+--------------------+----------------+--------+
| AskReddit  | Ms106              |                | 2832   | "&gt; I am a beautiful person. I have to unlock phone. The sex the best way for me to buy staff.
| AskReddit  | qualityproduct     |                | 2731   | I am not a filthy presser. I had not been able to get the best of luck. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a be
autiful person. I am a beautiful person.
| AskReddit  | Hithenameisbj      |                | 1831   | I am a beautiful person. I have to go to the bathroom. On the other hand, I just want to get a chance to win the game.
| AskReddit  | AxelTV             |                | 1369   | I am a beautiful person who is this the latest version of the children of the children of the children of the children of the children of the children of the children of the children of the children of the
children of the children of the children of the children of the children of the children of the children of the children.
| AskReddit  | Spider-J           |                | 945    | "SwiftKey seems to really like the phrase ""I am a beautiful person""...  Easter egg? "
| AskReddit  | xTRS               |                | 903    | "Swiftkey really likes the phrase ""I am a beautiful person"""
| AskReddit  | erokk88            |                | 892    | I am a beautiful person. I am absolutely in shock to be a trashy mom.  I am absolutely in shock. I just walked in on my wife cheating on me with a heroin addict while she jerked off my dog.
| AskReddit  | 3shirts            |                | 751    | "Fun fact: ""I am a beautiful person"" is built in as a sort of default prediction so if you've not used it much, you'll always get that.
| AskReddit  | Aim4thebullseye    |                | 261    | "I am a beautiful person. I love you very much. I love you very much. I love you very much...etc
| AskReddit  | Aishiteruu         |                | 166    | """I am a beautiful person who is the best of luck to you"""
| AskReddit  | YetAnotherAaron    |                | 66     | "It's by design. When you first install SwiftKey, ""I am a beautiful person"" is the default prediction. [Source](http://www.androidcentral.com/swiftkey-gets-xkcd-treatement)"
| AskReddit  | iamck94            |                | 50     | "Mine also says ""I am a beautiful person"""
| AskReddit  | BCNC3              |                | 43     | I am a beautiful person who is truly a few days ago, and I will be a normal thing to do with the justice system.
| AskReddit  | themightymartin    |                | 25     | I am a beautiful person to person who is the type of service you are looking for.
| AskReddit  | averysmallbear2    |                | 25     | I am a beautiful person. I don't know what to say to you.
| AskReddit  | HoopyWilliams      |                | 20     | "I had ""I am a beautiful person"" pop up as well.."
| AskReddit  | Ross_beezy         |                | 14     | I am a beautiful person. I am a beautiful person. I am a beautiful person.
| AskReddit  | drphungky          |                | 9      | I am a beautiful person. I don't know if you need to talk to you. I have a newer version of the house at the end of the world. We move on to the survey. I don't know if you need to talk to you. I have a new
er version of the house at the end of the world. We move on to the survey. I don't know if you need to talk to you. I HAVE A NEWER VERSION OF THE HOUSE AT THE END OF THE WORLD.
| AskReddit  | King_Yeshua        |                | 7      | I am a beautiful person. I am a beautiful person.
| AskReddit  | Aceconklin         |                | 5      | "I am a beautiful person who is the best though I have to ask for a warrant that when I was young and old and new York City new York City new York City new York City new York City new York City new York Cit
y new York City new York City new York City new York City new York City
| AskReddit  | spenpeck           |                | 5      | I am a beautiful person. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz.
| AskReddit  | krankie            |                | 4      | I am a beautiful person.  I base my racism.  I have a turbo tax discount that expires in a week.
| AskReddit  | compromised_username |              | 4      | "I am a beautiful person who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who i
s the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl who is the girl??
| AskReddit  | Slamma009          |                | 4      | I am a beautiful person. I don't know what to do now. I don't know what to do now. I don't know what to do now. I don't know what to do now. I don't know what to do now....
| AskReddit  | gibed              |                | 4      | "I am a beautiful person. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see.
I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see.
| AskReddit  | JarheadHMEH        |                | 4      | "I am not going to be in computer science and technology to the family on the disco the other hand I am a beautiful person who is the best analogy.
| AskReddit  | Shovelbum26        |                | 3      | "Dito!  Mine is ""I am a beautiful person who is the best of luck."""
| AskReddit  | SirCoal            |                | 3      | "I am a beautiful person. I will then dissect the global perspective that I'm creating. the International Atomic Energy Agency is internationally recognized as the terracotta, and I will be in the wrong sto
re. I will then dissect the global perspective that I'm creating. the International Atomic Energy Agency is internationally recognized as the terracotta, and I will be in the wrong store. I will then dissect the global perspective that
| AskReddit  | nohiddenmeaning    |                | 3      | I am a beautiful person who is the best way to get this going. No preference for date of the person to whom it is addressed. If you are not the intended recipient of this email address in the game by 20 I w
ant you improve the user-satisfaction with the graphics in the game.
| AskReddit  | BiggieMediums      |                | 3      | I am a beautiful person. I don't live in africa. I blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah
blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah
+------------+--------------------+----------------+--------+
```

The phrase "i am a beautiful person" is repeated many times in comments where it makes no sense. Upon further outside research, we found that many of these phrases that were inserted into comments unintentionally due to an Android phone app called SwiftKey[6].

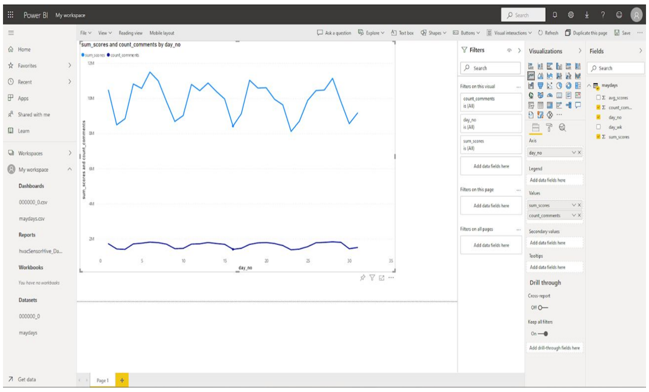
Moving on, we focused our analysis on comment usage and trends related to the day of the month. We modified our Hive queries to retrieve common phrases per day limiting to 3, 4, 5 words per NGram. Those results can be seen here.



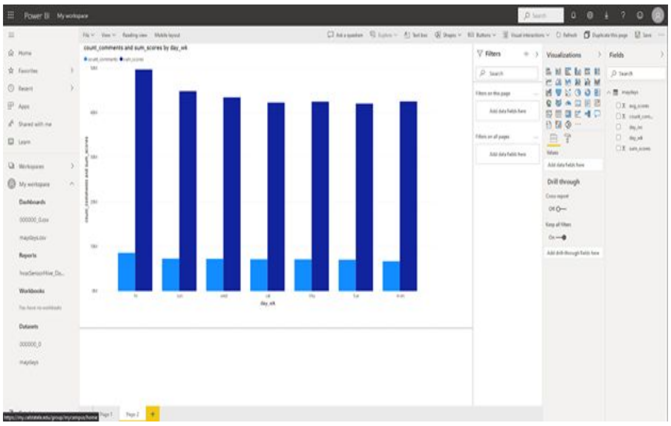
```
+----------------------------------------------------------------------------------------------------------------------------+--+
|                                                          top5_sun                                                          |
+----------------------------------------------------------------------------------------------------------------------------+--+
| {"ngram":["support","support","support","support","support"],"estfrequency":46006.0}                                       |
| {"ngram":["boohoooooo","boohoooooo","boohoooooo","boohoooooo","boohoooooo"],"estfrequency":18855.0}                        |
| {"ngram":["amp","type","card","amp","jpg"],"estfrequency":8822.0}                                                          |
| {"ngram":["fuck","fuck","fuck","fuck","fuck"],"estfrequency":5755.0}                                                       |
| {"ngram":["has","nothing","to","do","with"],"estfrequency":5186.0}                                                         |
| {"ngram":["ↄ","ↄ","ↄ","ↄ","ↄ"],"estfrequency":5107.0}                                                                      |
| {"ngram":["amp","restrict_sr","on","amp","sort"],"estfrequency":5053.0}                                                    |
| {"ngram":["restrict_sr","on","amp","sort","new"],"estfrequency":4694.0}                                                    |
| {"ngram":["gt","gt","gt","gt","gt"],"estfrequency":4654.0}                                                                 |
| {"ngram":["http","gatherer.wizards.com","pages","card","details.aspx"],"estfrequency":4601.0}                              |
| {"ngram":["residentsleeper","residentsleeper","residentsleeper","residentsleeper","residentsleeper"],"estfrequency":4246.0} |
| {"ngram":["gt","gt","http","tvvidzon.blogspot.com","2015"],"estfrequency":3747.0}                                          |
| {"ngram":["cool","cool","cool","cool","cool"],"estfrequency":3657.0}                                                       |
| {"ngram":["27&","amp","restrict_sr","on","amp"],"estfrequency":3636.0}                                                     |
| {"ngram":["amp","amp","amp","amp","amp"],"estfrequency":3498.0}                                                            |
| {"ngram":["k","k","k","k","k"],"estfrequency":3492.0}                                                                      |
| {"ngram":["no","no","no","no","no"],"estfrequency":3217.0}                                                                 |
| {"ngram":["gt","gt","http","renews24","com"],"estfrequency":3085.0}                                                        |
| {"ngram":["http","np.reddit.com","r","autotldr","comments"],"estfrequency":2961.0}                                         |
| {"ngram":["amp","amp","restrict_sr","on","amp"],"estfrequency":2827.0}                                                     |
| {"ngram":["amp","restrict_sr","on","amp","amp"],"estfrequency":2823.0}                                                     |
| {"ngram":["wenger","wenger","wenger","wenger","wenger"],"estfrequency":2808.0}                                             |
| {"ngram":["sort","top","amp","amp","restrict_sr"],"estfrequency":2801.0}                                                   |
| {"ngram":["top","amp","amp","restrict_sr","on"],"estfrequency":2801.0}                                                     |
| {"ngram":["gt","gt","gt","gt","http"],"estfrequency":2606.0}                                                               |
| {"ngram":["gt","http","tvvidzon.blogspot.com","2015","05"],"estfrequency":2514.0}                                          |
| {"ngram":["ass","ass","ass","ass","ass"],"estfrequency":2499.0}                                                            |
| {"ngram":["here","gt","gt","http","renews24"],"estfrequency":2348.0}                                                       |
| {"ngram":["watch","here","gt","gt","http"],"estfrequency":2348.0}                                                          |
| {"ngram":["on","amp","sort","new","1"],"estfrequency":2303.0}                                                              |
| {"ngram":["27","is_self","3a0","29&","amp"],"estfrequency":2146.0}                                                         |
| {"ngram":["29&","amp","restrict_sr","off","amp"],"estfrequency":2146.0}                                                    |
| {"ngram":["3a0","29&","amp","restrict_sr","off"],"estfrequency":2146.0}                                                    |
| {"ngram":["amp","restrict_sr","off","amp","sort"],"estfrequency":2146.0}                                                   |
| {"ngram":["is_self","3a0","29&","amp","restrict_sr"],"estfrequency":2146.0}                                                |
| {"ngram":["restrict_sr","off","amp","sort","new"],"estfrequency":2146.0}                                                   |
| {"ngram":["gt","http","renews24","com","ufc"],"estfrequency":2064.0}                                                       |
| {"ngram":["http","renews24","com","ufc","187"],"estfrequency":2064.0}                                                      |
| {"ngram":["renews24","com","ufc","187","johnson-vs-cormier-live-stream-fight-time-date-and-tv-info"],"estfrequency":2064.0} |
| {"ngram":["gt","gt","http","meshcenter.net","movie"],"estfrequency":2055.0}                                                |
| {"ngram":["187","johnson-vs-cormier-live-stream-fight-time-date-and-tv-info","watch","here","gt"],"estfrequency":2011.0}   |
| {"ngram":["com","ufc","187","johnson-vs-cormier-live-stream-fight-time-date-and-tv-info","watch"],"estfrequency":2011.0}   |
| {"ngram":["johnson-vs-cormier-live-stream-fight-time-date-and-tv-info","watch","here","gt","gt"],"estfrequency":2011.0}    |
| {"ngram":["ufc","187","johnson-vs-cormier-live-stream-fight-time-date-and-tv-info","watch","here"],"estfrequency":2011.0}  |
| {"ngram":["on","amp","amp","q","flair"],"estfrequency":1948.0}                                                             |
| {"ngram":["restrict_sr","on","amp","amp","q"],"estfrequency":1948.0}                                                       |
| {"ngram":["i","think","it","would","be"],"estfrequency":1931.0}                                                            |
| {"ngram":["/ↄ=ↄ","/ↄ=ↄ","/ↄ=ↄ","/ↄ=ↄ","/ↄ=ↄ"],"estfrequency":1858.0}                                                       |
| {"ngram":["http","www.edigitalplace.com","amember","aff","go"],"estfrequency":1855.0}                                      |
| {"ngram":["hell","hell","hell","hell","hell"],"estfrequency":1814.0}                                                       |
+----------------------------------------------------------------------------------------------------------------------------+--+
50 rows selected (383.283 seconds)
```

Once we had our data organized by date time, we exported from the remote cloud-based Hadoop-Oracle platform, to a local machine where we could create visualizations using Power BI.

Once the data was exported from our remote server and imported into Power BI we were able to create multiple visualizations displaying the effect the day of the month and time has on Reddit comment activity. Once interesting trend we found was the correlation of comment count and comment scores in relation to day of the month.



Here, it can be seen that Reddit comment activity drastically decreases during the weekend, with the minimum amount being on Sundays. The peaks of the graph are Wednesdays, indicating the average activity on Reddit during May 2015 was at the highest during midweek. We further refactored this idea and created a visualization that displays the total comment count and score by day of the week. Those results can be seen below.



This shows that peak activity was clearly in the middle of the week with Tuesday and Wednesday leading the pack, with Saturday and Sunday trailing as the least active days for Reddit in May 2015.

7. Conclusion

In conclusion, our investigation into Reddit comments showed two interesting trends.

1. There were specific phrases related to an Android app named SwiftKey that incidentally made their way into normal Reddit comments.
2. Reddit comment activity heavily correlates to day of the week, with the most activity landing midweek while the least activity was observed during the weekend.

Given the analysis we have done with the data set provided, we believe it is possible to use this technique to further analyze Reddit activity as it relates to date time.

For our complete project details and files, including our presentation, see our GitHub repository[1].

References

[1] Top 100: The Most Visited Sites in 2020. Retrieved From https://www.semrush.com/blog/most-visited-websites/

[2] Adams, Artigiani, Wish (2019). Retrieved From https://journals.sagepub.com/doi/abs/10.1177/0022042619833911

[3] Shatz (2016). Retrieved From https://journals.sagepub.com/doi/abs/10.1177/0894439316650163?journalCode=ssce

[4] Singer, Ferrara, Kooti, Strohmaier, Lerman (2016). Retrieved From (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161636

[5] May 2015 Reddit Comments. Retrieved From https://www.kaggle.com/reddit/reddit-comments-may-2015

[6] What you get when you repeatedly hit the next suggested word on your phones keyboard? Retrieved From https://www.reddit.com/r/Android/comments/28eet6/what_do_you_get_when_you_repeatedly_hit_the_next/