CIS4560 Term Project Tutorial

Authors: Daniel Kang, Klayton Trinh, Michael Kintanar, Danny Xie, Ryan Hathcock
Instructor: Jongwook Woo
Date: 12/20/2020

Lab Tutorial

12/20/2020

Reddit Data Analysis using NGrams

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Retrieve data from the repository
- Upload to HDFS
- Create tables correlated to analysis of data
- SQL commands to perform the analysis.
- Visualization

Platform Spec

- # of nodes: 3
- Total Memory Size: 180GB

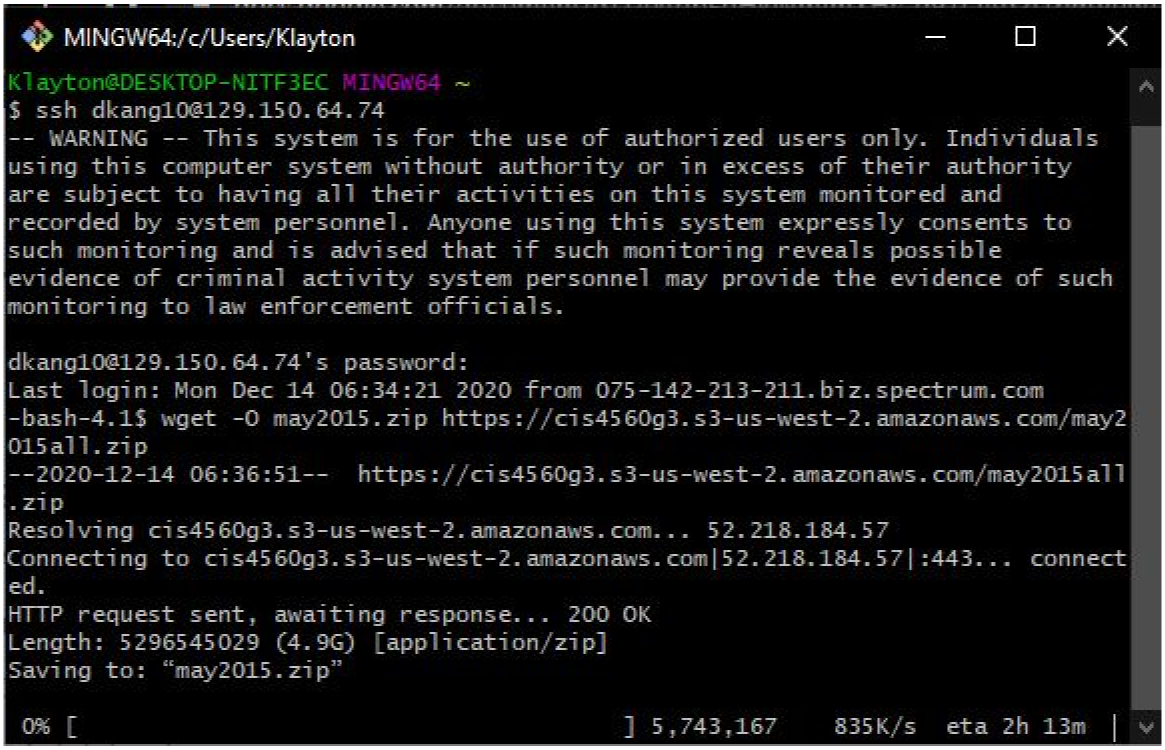
- Cluster Size: 802.6GB

Step 1: Connect to Oracle Cloud and Retrieve Data from Amazon s3

1. Connect to Oracle Cloud. Please change the text in red to the appropriate user.

   $ ssh dkang10@129.150.64.74

2. $ wget -O may2015all.zip https://cis4560g3.s3-us-west-2.amazonaws.com/may2015all.zip



```
Klayton@DESKTOP-NITF3EC MINGW64 ~
$ ssh dkang10@129.150.64.74
-- WARNING -- This system is for the use of authorized users only. Individuals
using this computer system without authority or in excess of their authority
are subject to having all their activities on this system monitored and
recorded by system personnel. Anyone using this system expressly consents to
such monitoring and is advised that if such monitoring reveals possible
evidence of criminal activity system personnel may provide the evidence of such
monitoring to law enforcement officials.

dkang10@129.150.64.74's password:
Last login: Mon Dec 14 06:34:21 2020 from 075-142-213-211.biz.spectrum.com
-bash-4.1$ wget -O may2015.zip https://cis4560g3.s3-us-west-2.amazonaws.com/may2
015all.zip
--2020-12-14 06:36:51--  https://cis4560g3.s3-us-west-2.amazonaws.com/may2015all
.zip
Resolving cis4560g3.s3-us-west-2.amazonaws.com... 52.218.184.57
Connecting to cis4560g3.s3-us-west-2.amazonaws.com|52.218.184.57|:443... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 5296545029 (4.9G) [application/zip]
Saving to: “may2015.zip”

 0% [                                        ] 5,743,167    835K/s   eta 2h 13m
```

3. Once the Data is retrieved we must unzip the file.

   $ unzip may2015.zip

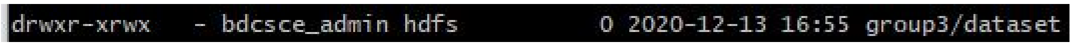
4. After that we can create our HDFS directory.

   $ hdfs dfs -mkdir group3

   $ hdfs dfs -mkdir group3/dataset

5. Confirm the directory was made.

   $ hdfs dfs -ls group3



```
drwxr-xrwx   - bdcsce_admin hdfs          0 2020-12-13 16:55 group3/dataset
```

6. Insert the file into the created directory.

   $ HDFS DFS -put may_2015_trim.csv group 3/dataset

Step 2: Connect to Beeline and Create Tables

7. Run the following HDFS command to make your beeline command works.

   $ hdfs dfs -chmod -R o+w .

8. Open Beeline, then connect to hive servers.

   $ beeline

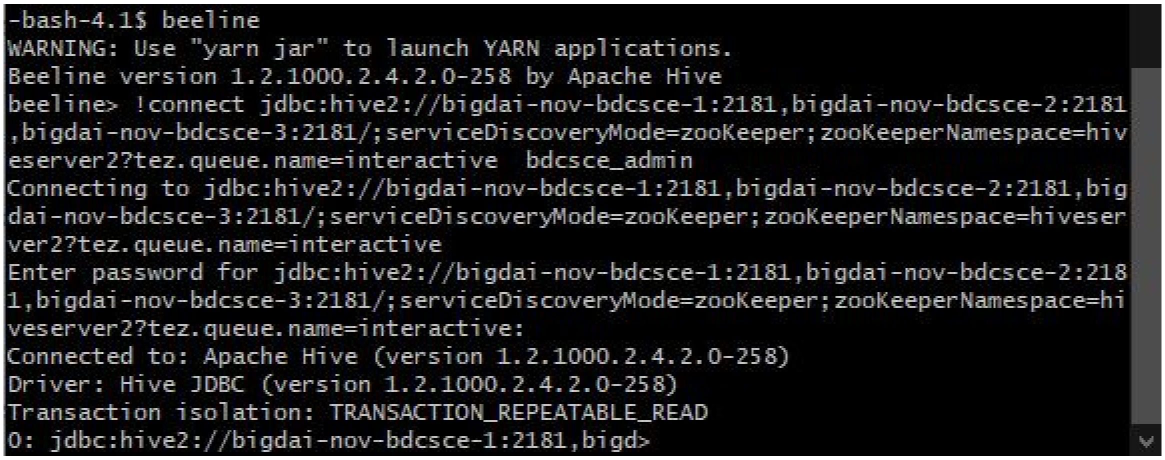
   beeline> !connect

   jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181,bigdai-nov-bdcsce-3:2181/;ser

   viceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive

   bdcsce_admin

   When prompted for a password press the “Enter” key.



```
-bash-4.1$ beeline
WARNING: Use "yarn jar" to launch YARN applications.
Beeline version 1.2.1000.2.4.2.0-258 by Apache Hive
beeline> !connect jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181
,bigdai-nov-bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiv
eserver2?tez.queue.name=interactive  bdcsce_admin
Connecting to jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181,big
dai-nov-bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveser
ver2?tez.queue.name=interactive
Enter password for jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:218
1,bigdai-nov-bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hi
veserver2?tez.queue.name=interactive:
Connected to: Apache Hive (version 1.2.1000.2.4.2.0-258)
Driver: Hive JDBC (version 1.2.1000.2.4.2.0-258)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigd>
```

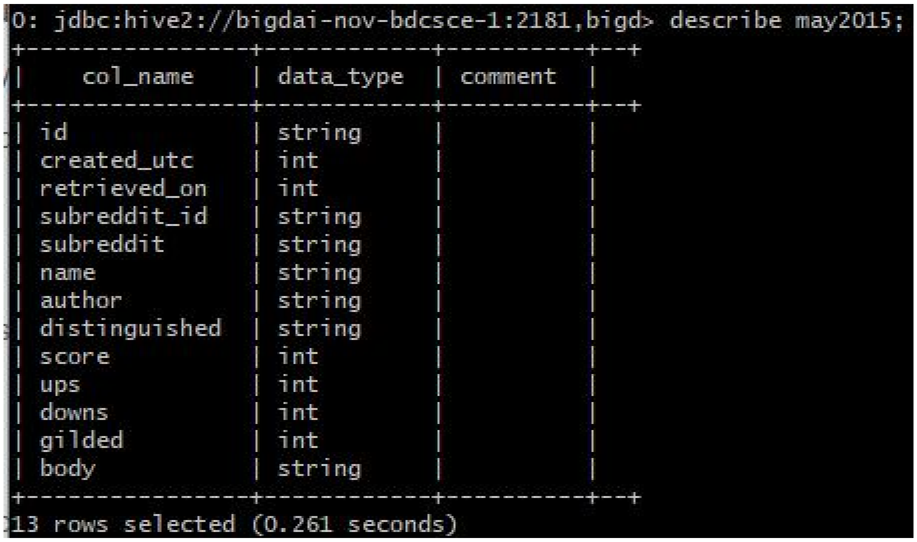
9. Once in beeline let's use our database and create some tables. First table is for general information collecting. Please change the text in red to the appropriate user.

```
create external table may2015 (
        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
)
row format delimited
fields terminated by '`'
lines terminated by '\n'
location '/user/dkang10/group3/dataset'
tblproperties('skip.header.line.count'='1');
```

10. Confirm the table is in the database.

```
show tables;

describe may2015;
```



```
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)
```

```
select * from may2015 limit 10;
```

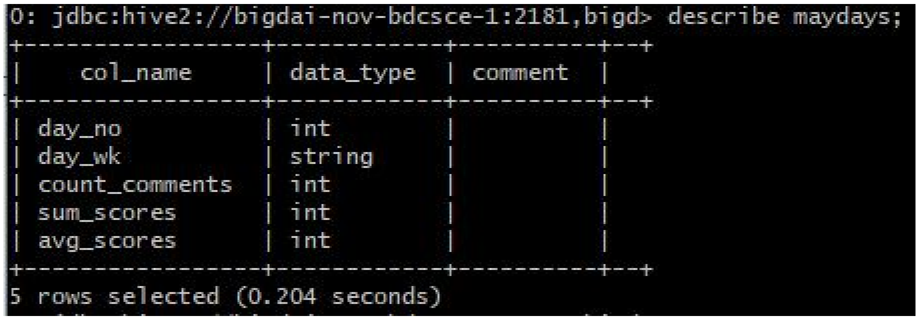
11. Create a second table that deals with temporal data.

```
create table maydays(
        day_no int,
        day_wk string,
        count_comments int,
        sum_scores int,
        avg_scores int
)
row format delimited
fields terminated by ','
        lines terminated by '\n';
```

12. Confirm the table is in the database.

```
show tables;

describe maydays;
```



```
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)
```

13. Populate the table with the days in May

```
insert into table maydays
select '1', 'fri', count(body), sum(score), avg(score) from may2015
where created_utc>1430438400 and created_utc<1430524799;

insert into table maydays
select '2', 'sat', count(body), sum(score), avg(score) from may2015
where created_utc>1430524800 and created_utc<1430611199;

insert into table maydays
select '3', 'sun', count(body), sum(score), avg(score) from may2015
where created_utc>1430611200 and created_utc<1430697599;

insert into table maydays
```

```
select '4', 'mon', count(body), sum(score), avg(score) from may2015
where created_utc>1430697600 and created_utc<1430783999;

insert into table maydays
select '5', 'tue', count(body), sum(score), avg(score) from may2015
where created_utc>1430784000 and created_utc<1430870399;

insert into table maydays
select '6', 'wed', count(body), sum(score), avg(score) from may2015
where created_utc>1430870400 and created_utc<1430956799;

insert into table maydays
select '7', 'thu', count(body), sum(score), avg(score) from may2015
where created_utc>1430956800 and created_utc<1431043199;

insert into table maydays
select '8', 'fri', count(body), sum(score), avg(score) from may2015
where created_utc>1431043200 and created_utc<1431129599;

insert into table maydays
select '9', 'sat', count(body), sum(score), avg(score) from may2015
where created_utc>1431129600 and created_utc<1431215999;

insert into table maydays
select '10', 'sun', count(body), sum(score), avg(score) from may2015
where created_utc>1431216000 and created_utc<1431302399;

insert into table maydays
select '11', 'mon', count(body), sum(score), avg(score) from may2015
where created_utc>1431302400 and created_utc<1431388799;

insert into table maydays
select '12', 'tue', count(body), sum(score), avg(score) from may2015
where created_utc>1431388800 and created_utc<1431475199;

insert into table maydays
select '13', 'wed', count(body), sum(score), avg(score) from may2015
where created_utc>1431475200 and created_utc<1431561599;

insert into table maydays
select '14', 'thu', count(body), sum(score), avg(score) from may2015
where created_utc>1431561600 and created_utc<1431647999;

insert into table maydays
select '15', 'fri', count(body), sum(score), avg(score) from may2015
where created_utc>1431648000 and created_utc<1431734399;

insert into table maydays
select '16', 'sat', count(body), sum(score), avg(score) from may2015
```

```
where created_utc>1431734400 and created_utc<1431820799;

insert into table maydays
select '17', 'sun', count(body), sum(score), avg(score) from may2015
where created_utc>1431820800 and created_utc<1431907199;

insert into table maydays
select '18', 'mon', count(body), sum(score), avg(score) from may2015
where created_utc>1431907200 and created_utc<1431993599;

insert into table maydays
select '19', 'tue', count(body), sum(score), avg(score) from may2015
where created_utc>1431993600 and created_utc<1432079999;

insert into table maydays
select '20', 'wed', count(body), sum(score), avg(score) from may2015
where created_utc>1432080000 and created_utc<1432166399;

insert into table maydays
select '21', 'thu', count(body), sum(score), avg(score) from may2015
where created_utc>1432166400 and created_utc<1432252799;

insert into table maydays
select '22', 'fri', count(body), sum(score), avg(score) from may2015
where created_utc>1432252800 and created_utc<1432339199;

insert into table maydays
select '23', 'sat', count(body), sum(score), avg(score) from may2015
where created_utc>1432339200 and created_utc<1432425599;

insert into table maydays
select '24', 'sun', count(body), sum(score), avg(score) from may2015
where created_utc>1432425600 and created_utc<1432511999;

insert into table maydays
select '25', 'mon', count(body), sum(score), avg(score) from may2015
where created_utc>1432512000 and created_utc<1432598399;

insert into table maydays
select '26', 'tue', count(body), sum(score), avg(score) from may2015
where created_utc>1432598400 and created_utc<1432684799;

insert into table maydays
select '27', 'wed', count(body), sum(score), avg(score) from may2015
where created_utc>1432684800 and created_utc<1432771199;

insert into table maydays
select '28', 'thu', count(body), sum(score), avg(score) from may2015
where created_utc>1432771200 and created_utc<1432857599;
```

```
insert into table maydays
select '29', 'fri', count(body), sum(score), avg(score) from may2015
where created_utc>1432857600 and created_utc<1432943999;

insert into table maydays
select '30', 'sat', count(body), sum(score), avg(score) from may2015
where created_utc>1432944000 and created_utc<1433030399;

insert into table maydays
select '31', 'sun', count(body), sum(score), avg(score) from may2015
where created_utc>1433030400 and created_utc<1433116799;
```

14. Confirm the “maydays” has been properly created and populated.

```
select * from maydays;
```

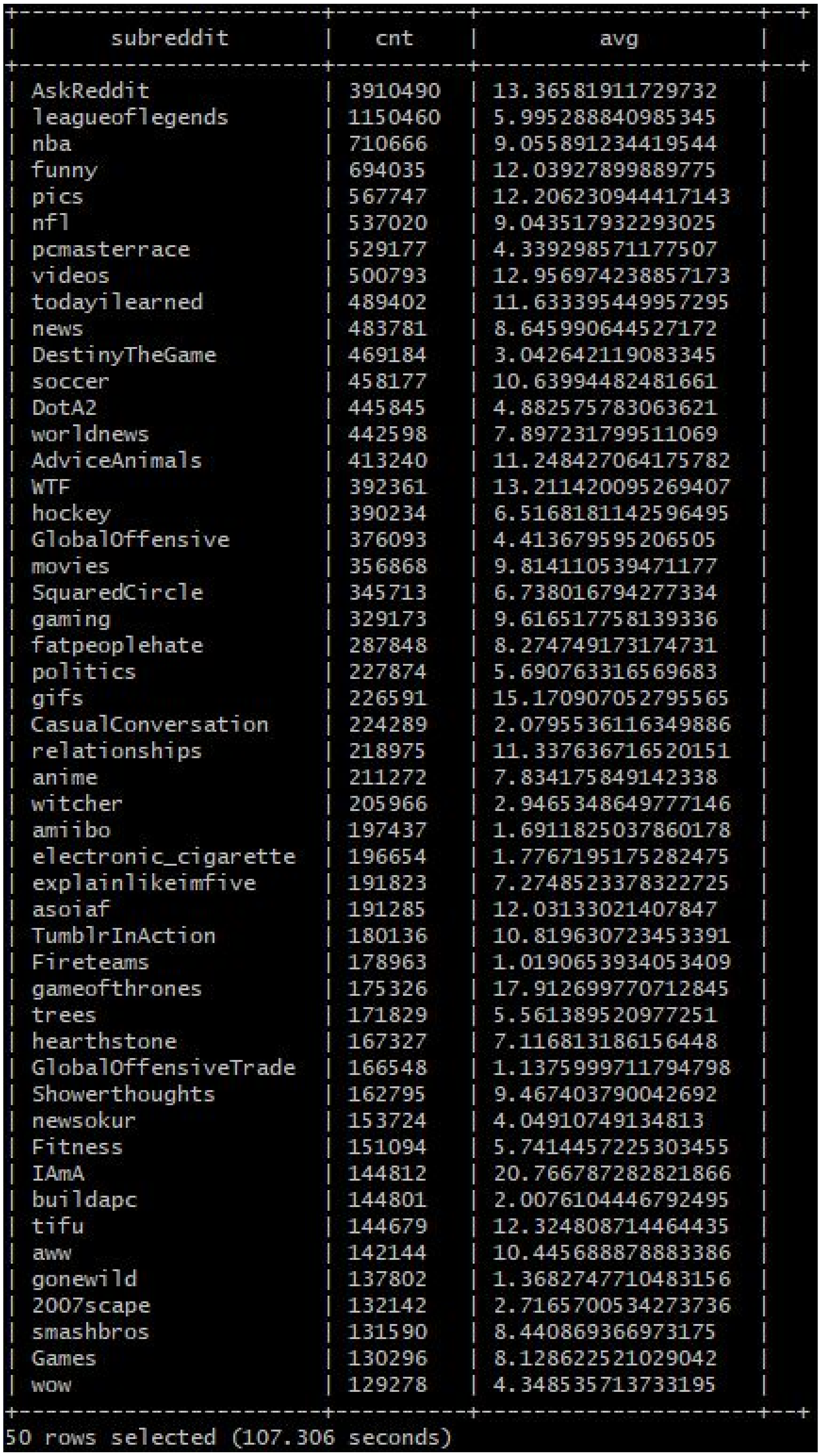
15. Create a new table for us to use for visualization at the end of the tutorial. Please change the text in red to the appropriate user.

```
create table mayday_export(
	day_no int,
	day_wk string,
	count_comments int,
	sum_scores int,
	avg_scores int
)
row format delimited
fields terminated by ','
lines terminated by '\n'
stored as textfile location '/user/dkang10/group3/export';
```

Step 3: Analysis of the Data gathered

16. Display a table that shows top 50 subreddits sorted by number of comments.

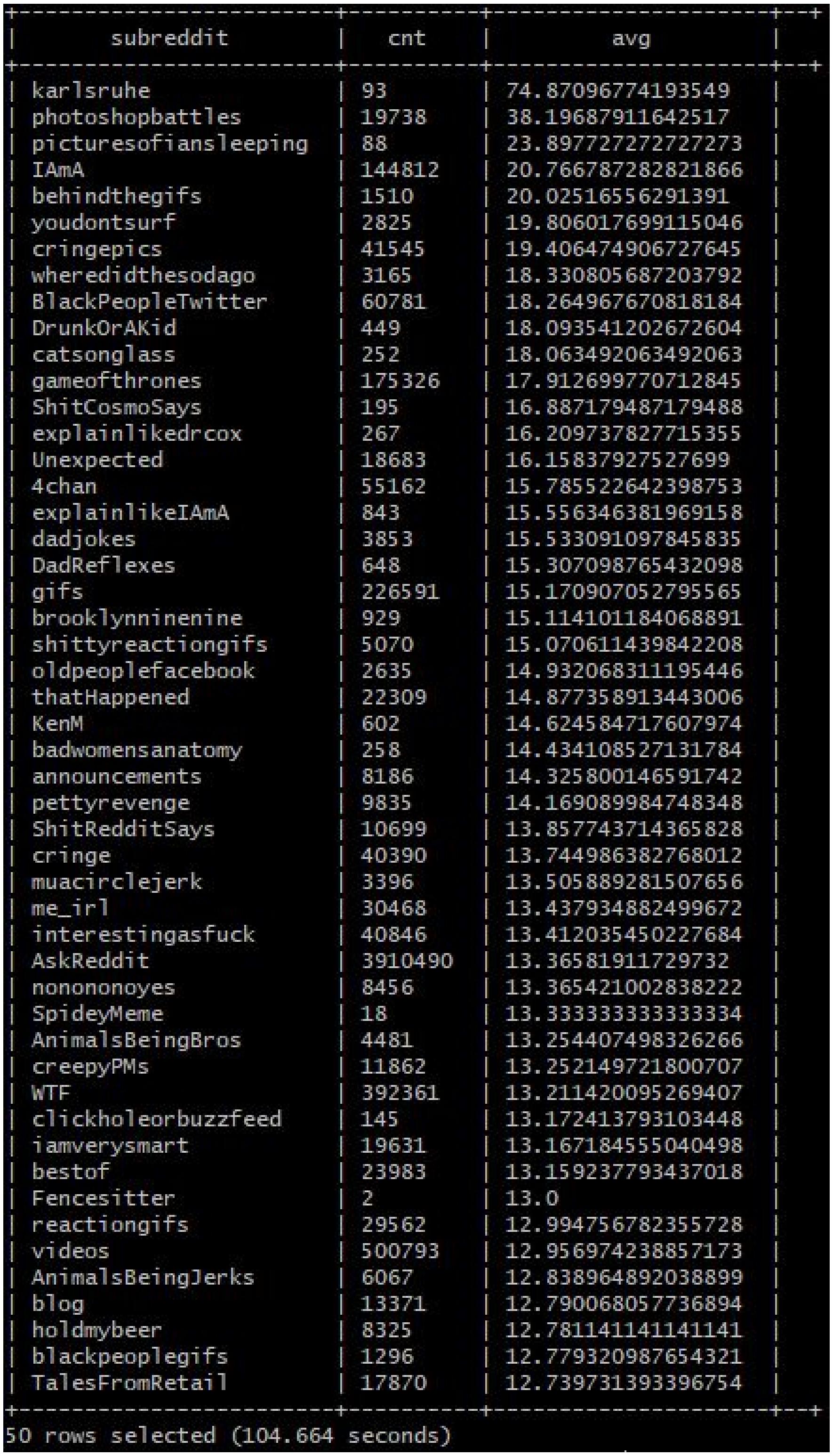
```
select subreddit, count(score) as cnt, avg(score) as avg from may2015
group by subreddit order by cnt desc limit 50;
```



```
+----------------------+----------+----------------------+--+
|       subreddit      |   cnt    |          avg         |
+----------------------+----------+----------------------+--+
| AskReddit            | 3910490  | 13.36581911729732    |
| 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   | 7.2748523378322725   |
| 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   | 5.7414457225303455   |
| IAmA                 | 144812   | 20.766787282821866   |
| buildapc             | 144801   | 2.0076104446792495   |
| tifu                 | 144679   | 12.324808714464435   |
| aww                  | 142144   | 10.445688878883386   |
| gonewild             | 137802   | 1.3682747710483156   |
| 2007scape            | 132142   | 2.7165700534273736   |
| smashbros            | 131590   | 8.440869366973175    |
| Games                | 130296   | 8.128622521029042    |
| wow                  | 129278   | 4.348535713733195    |
+----------------------+----------+----------------------+--+
50 rows selected (107.306 seconds)
```

17. Display a table that shows the top 50 subreddits by number of upvotes per comment.

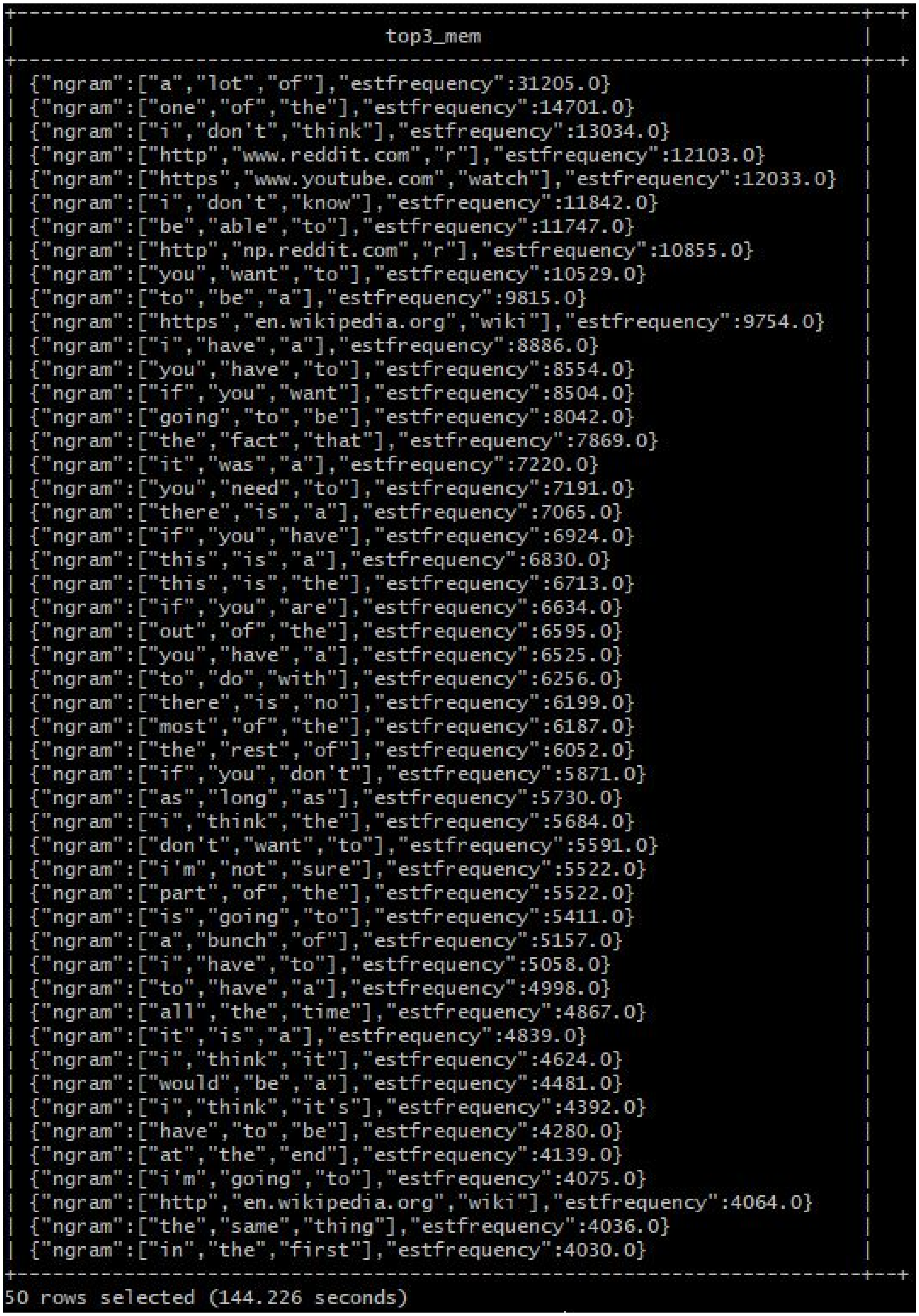
select subreddit, count(score) as cnt, avg(score) as avg from may2015
group by subreddit order by avg desc limit 50;



```
+------------------------+----------+----------------------+--+
|       subreddit        |   cnt    |         avg          |
+------------------------+----------+----------------------+--+
| karlsruhe              | 93       | 74.87096774193549    |
| photoshopbattles       | 19738    | 38.19687911642517    |
| picturesofiansleeping  | 88       | 23.897727272727273   |
| IAmA                   | 144812   | 20.766787282821866   |
| behindthegifs          | 1510     | 20.02516556291391    |
| youdontsurf            | 2825     | 19.806017699115046   |
| cringepics             | 41545    | 19.406474906727645   |
| wheredidthesodago      | 3165     | 18.330805687203792   |
| BlackPeopleTwitter     | 60781    | 18.264967670818184   |
| DrunkOrAKid            | 449      | 18.093541202672604   |
| catsonglass            | 252      | 18.063492063492063   |
| gameofthrones          | 175326   | 17.912699770712845   |
| ShitCosmoSays          | 195      | 16.887179487179488   |
| explainlikedrcox       | 267      | 16.209737827715355   |
| Unexpected             | 18683    | 16.15837927527699    |
| 4chan                  | 55162    | 15.785522642398753   |
| explainlikeIAmA        | 843      | 15.556346381969158   |
| dadjokes               | 3853     | 15.533091097845835   |
| DadReflexes            | 648      | 15.307098765432098   |
| gifs                   | 226591   | 15.170907052795565   |
| brooklynninenine       | 929      | 15.114101184068891   |
| shittyreactiongifs     | 5070     | 15.070611439842208   |
| oldpeoplefacebook      | 2635     | 14.932068311195446   |
| thatHappened           | 22309    | 14.877358913443006   |
| KenM                   | 602      | 14.624584717607974   |
| badwomensanatomy       | 258      | 14.434108527131784   |
| announcements          | 8186     | 14.325800146591742   |
| pettyrevenge           | 9835     | 14.169089984748348   |
| ShitRedditSays         | 10699    | 13.857743714365828   |
| cringe                 | 40390    | 13.744986382768012   |
| muacirclejerk          | 3396     | 13.505889281507656   |
| me_irl                 | 30468    | 13.437934882499672   |
| interestingasfuck      | 40846    | 13.412035450227684   |
| AskReddit              | 3910490  | 13.36581911729732    |
| nonononoyes            | 8456     | 13.365421002838222   |
| SpideyMeme             | 18       | 13.333333333333334   |
| AnimalsBeingBros       | 4481     | 13.254407498326266   |
| creepyPMs              | 11862    | 13.252149721800707   |
| WTF                    | 392361   | 13.211420095269407   |
| clickholeorbuzzfeed    | 145      | 13.172413793103448   |
| iamverysmart           | 19631    | 13.167184555040498   |
| bestof                 | 23983    | 13.159237793437018   |
| Fencesitter            | 2        | 13.0                 |
| reactiongifs           | 29562    | 12.994756782355728   |
| videos                 | 500793   | 12.956974238857173   |
| AnimalsBeingJerks      | 6067     | 12.838964892038899   |
| blog                   | 13371    | 12.790068057736894   |
| holdmybeer             | 8325     | 12.781141141141141   |
| blackpeoplegifs        | 1296     | 12.779320987654321   |
| TalesFromRetail        | 17870    | 12.739731393396754   |
+------------------------+----------+----------------------+--+
50 rows selected (104.664 seconds)
```

18. Display the top 3gram on Memorial Day, Monday 25th.

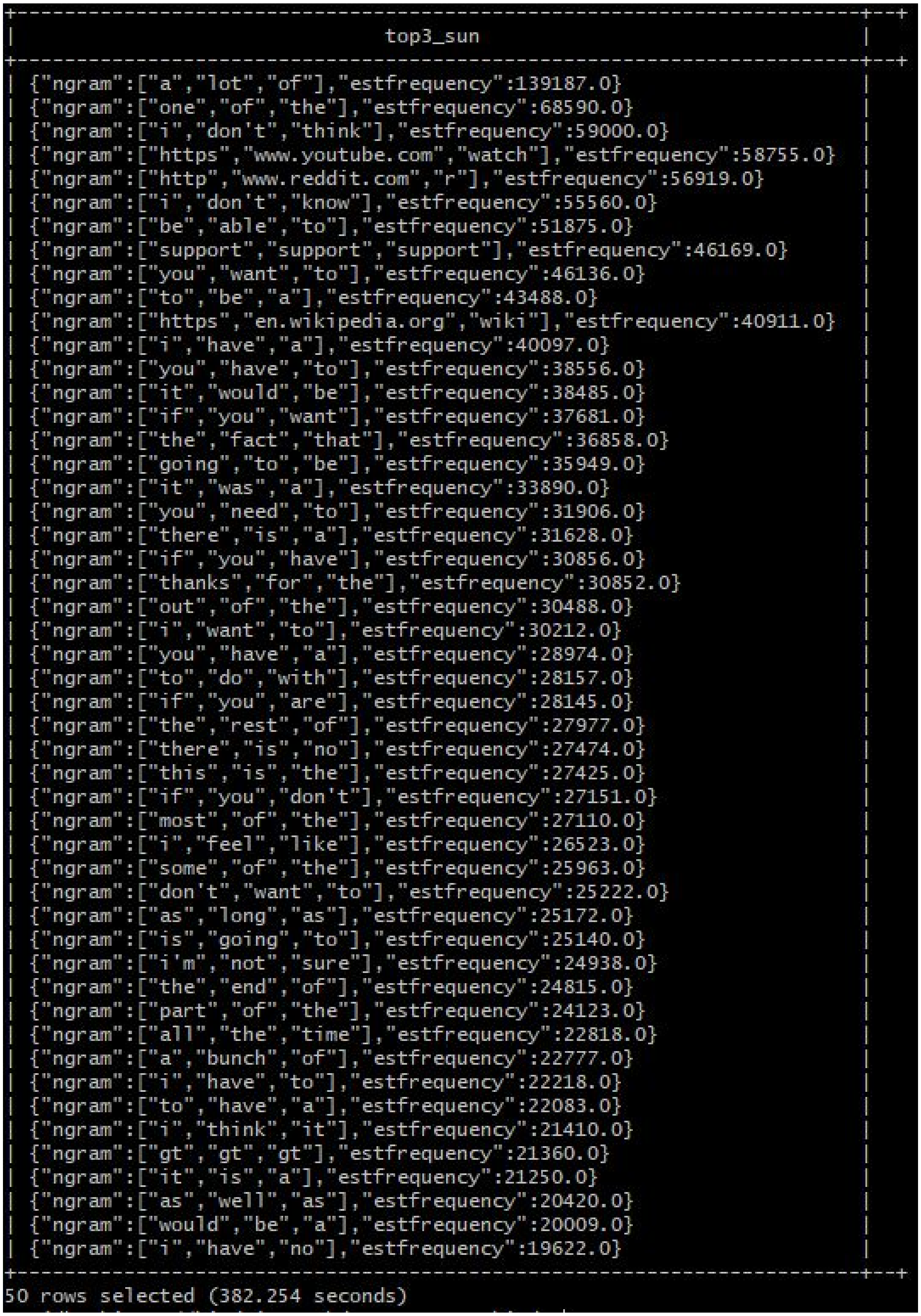
select explode(ngrams(sentences(lower(body)), 3, 50)) as top3_mem from may2015
where created_utc>1432512000 and created_utc<1432598399;



```
+-----------------------------------------------------------------------------+--+
|                                  top3_mem                                   |
+-----------------------------------------------------------------------------+--+
| {"ngram":["a","lot","of"],"estfrequency":31205.0}                            |
| {"ngram":["one","of","the"],"estfrequency":14701.0}                          |
| {"ngram":["i","don't","think"],"estfrequency":13034.0}                       |
| {"ngram":["http","www.reddit.com","r"],"estfrequency":12103.0}               |
| {"ngram":["https","www.youtube.com","watch"],"estfrequency":12033.0}         |
| {"ngram":["i","don't","know"],"estfrequency":11842.0}                        |
| {"ngram":["be","able","to"],"estfrequency":11747.0}                          |
| {"ngram":["http","np.reddit.com","r"],"estfrequency":10855.0}                |
| {"ngram":["you","want","to"],"estfrequency":10529.0}                         |
| {"ngram":["to","be","a"],"estfrequency":9815.0}                              |
| {"ngram":["https","en.wikipedia.org","wiki"],"estfrequency":9754.0}          |
| {"ngram":["i","have","a"],"estfrequency":8886.0}                             |
| {"ngram":["you","have","to"],"estfrequency":8554.0}                          |
| {"ngram":["if","you","want"],"estfrequency":8504.0}                          |
| {"ngram":["going","to","be"],"estfrequency":8042.0}                          |
| {"ngram":["the","fact","that"],"estfrequency":7869.0}                        |
| {"ngram":["it","was","a"],"estfrequency":7220.0}                             |
| {"ngram":["you","need","to"],"estfrequency":7191.0}                          |
| {"ngram":["there","is","a"],"estfrequency":7065.0}                           |
| {"ngram":["if","you","have"],"estfrequency":6924.0}                          |
| {"ngram":["this","is","a"],"estfrequency":6830.0}                            |
| {"ngram":["this","is","the"],"estfrequency":6713.0}                          |
| {"ngram":["if","you","are"],"estfrequency":6634.0}                           |
| {"ngram":["out","of","the"],"estfrequency":6595.0}                           |
| {"ngram":["you","have","a"],"estfrequency":6525.0}                           |
| {"ngram":["to","do","with"],"estfrequency":6256.0}                           |
| {"ngram":["there","is","no"],"estfrequency":6199.0}                          |
| {"ngram":["most","of","the"],"estfrequency":6187.0}                          |
| {"ngram":["the","rest","of"],"estfrequency":6052.0}                          |
| {"ngram":["if","you","don't"],"estfrequency":5871.0}                         |
| {"ngram":["as","long","as"],"estfrequency":5730.0}                           |
| {"ngram":["i","think","the"],"estfrequency":5684.0}                          |
| {"ngram":["don't","want","to"],"estfrequency":5591.0}                        |
| {"ngram":["i'm","not","sure"],"estfrequency":5522.0}                         |
| {"ngram":["part","of","the"],"estfrequency":5522.0}                          |
| {"ngram":["is","going","to"],"estfrequency":5411.0}                          |
| {"ngram":["a","bunch","of"],"estfrequency":5157.0}                           |
| {"ngram":["i","have","to"],"estfrequency":5058.0}                            |
| {"ngram":["to","have","a"],"estfrequency":4998.0}                            |
| {"ngram":["all","the","time"],"estfrequency":4867.0}                         |
| {"ngram":["it","is","a"],"estfrequency":4839.0}                              |
| {"ngram":["i","think","it"],"estfrequency":4624.0}                           |
| {"ngram":["would","be","a"],"estfrequency":4481.0}                           |
| {"ngram":["i","think","it's"],"estfrequency":4392.0}                         |
| {"ngram":["have","to","be"],"estfrequency":4280.0}                           |
| {"ngram":["at","the","end"],"estfrequency":4139.0}                           |
| {"ngram":["i'm","going","to"],"estfrequency":4075.0}                         |
| {"ngram":["http","en.wikipedia.org","wiki"],"estfrequency":4064.0}           |
| {"ngram":["the","same","thing"],"estfrequency":4036.0}                       |
| {"ngram":["in","the","first"],"estfrequency":4030.0}                         |
+-----------------------------------------------------------------------------+--+
50 rows selected (144.226 seconds)
```

19. Display the top 3gram on Sunday.

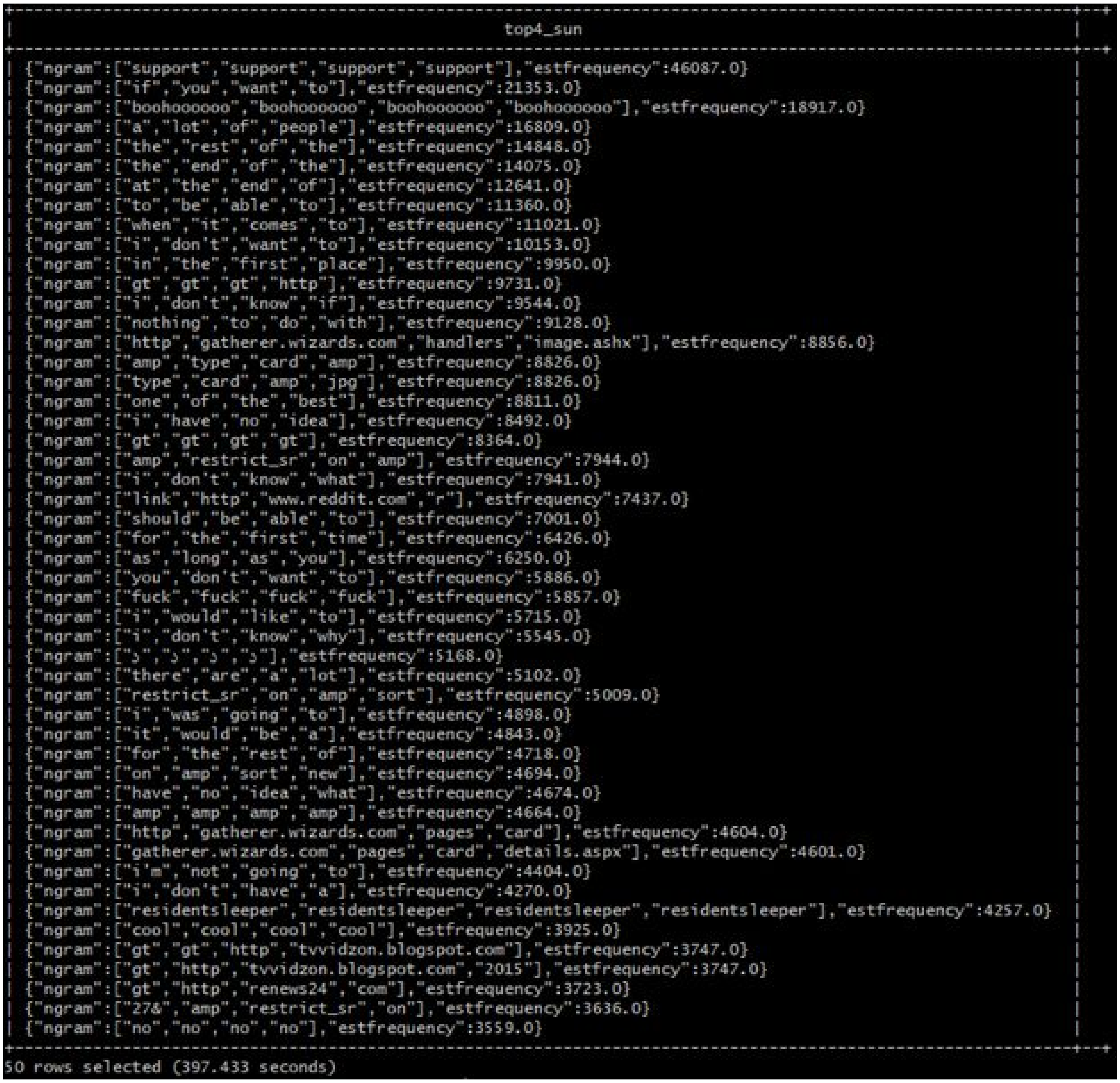
```
select explode(ngrams(sentences(lower(body)), 3, 50)) as top3_sun from may2015
where (created_utc>1430611200 and created_utc<1430697599)
or (created_utc>1431216000 and created_utc<1431302399)
or (created_utc>1431820800 and created_utc<1431907199)
or (created_utc>1432425600 and created_utc<1432511999)
or (created_utc>1433030400 and created_utc<1433116799);
```



```
+--------------------------------------------------------------------------+--+
|                                 top3_sun                                 |
+--------------------------------------------------------------------------+--+
| {"ngram":["a","lot","of"],"estfrequency":139187.0}                       |
| {"ngram":["one","of","the"],"estfrequency":68590.0}                      |
| {"ngram":["i","don't","think"],"estfrequency":59000.0}                   |
| {"ngram":["https","www.youtube.com","watch"],"estfrequency":58755.0}     |
| {"ngram":["http","www.reddit.com","r"],"estfrequency":56919.0}           |
| {"ngram":["i","don't","know"],"estfrequency":55560.0}                    |
| {"ngram":["be","able","to"],"estfrequency":51875.0}                      |
| {"ngram":["support","support","support"],"estfrequency":46169.0}         |
| {"ngram":["you","want","to"],"estfrequency":46136.0}                     |
| {"ngram":["to","be","a"],"estfrequency":43488.0}                         |
| {"ngram":["https","en.wikipedia.org","wiki"],"estfrequency":40911.0}     |
| {"ngram":["i","have","a"],"estfrequency":40097.0}                        |
| {"ngram":["you","have","to"],"estfrequency":38556.0}                     |
| {"ngram":["it","would","be"],"estfrequency":38485.0}                     |
| {"ngram":["if","you","want"],"estfrequency":37681.0}                     |
| {"ngram":["the","fact","that"],"estfrequency":36858.0}                   |
| {"ngram":["going","to","be"],"estfrequency":35949.0}                     |
| {"ngram":["it","was","a"],"estfrequency":33890.0}                        |
| {"ngram":["you","need","to"],"estfrequency":31906.0}                     |
| {"ngram":["there","is","a"],"estfrequency":31628.0}                      |
| {"ngram":["if","you","have"],"estfrequency":30856.0}                     |
| {"ngram":["thanks","for","the"],"estfrequency":30852.0}                  |
| {"ngram":["out","of","the"],"estfrequency":30488.0}                      |
| {"ngram":["i","want","to"],"estfrequency":30212.0}                       |
| {"ngram":["you","have","a"],"estfrequency":28974.0}                      |
| {"ngram":["to","do","with"],"estfrequency":28157.0}                      |
| {"ngram":["if","you","are"],"estfrequency":28145.0}                      |
| {"ngram":["the","rest","of"],"estfrequency":27977.0}                     |
| {"ngram":["there","is","no"],"estfrequency":27474.0}                     |
| {"ngram":["this","is","the"],"estfrequency":27425.0}                     |
| {"ngram":["if","you","don't"],"estfrequency":27151.0}                    |
| {"ngram":["most","of","the"],"estfrequency":27110.0}                     |
| {"ngram":["i","feel","like"],"estfrequency":26523.0}                     |
| {"ngram":["some","of","the"],"estfrequency":25963.0}                     |
| {"ngram":["don't","want","to"],"estfrequency":25222.0}                   |
| {"ngram":["as","long","as"],"estfrequency":25172.0}                      |
| {"ngram":["is","going","to"],"estfrequency":25140.0}                     |
| {"ngram":["i'm","not","sure"],"estfrequency":24938.0}                    |
| {"ngram":["the","end","of"],"estfrequency":24815.0}                      |
| {"ngram":["part","of","the"],"estfrequency":24123.0}                     |
| {"ngram":["all","the","time"],"estfrequency":22818.0}                    |
| {"ngram":["a","bunch","of"],"estfrequency":22777.0}                      |
| {"ngram":["i","have","to"],"estfrequency":22218.0}                       |
| {"ngram":["to","have","a"],"estfrequency":22083.0}                       |
| {"ngram":["i","think","it"],"estfrequency":21410.0}                      |
| {"ngram":["gt","gt","gt"],"estfrequency":21360.0}                        |
| {"ngram":["it","is","a"],"estfrequency":21250.0}                         |
| {"ngram":["as","well","as"],"estfrequency":20420.0}                      |
| {"ngram":["would","be","a"],"estfrequency":20009.0}                      |
| {"ngram":["i","have","no"],"estfrequency":19622.0}                       |
+--------------------------------------------------------------------------+--+
50 rows selected (382.254 seconds)
```

20. Display the top 4gram on Sunday.

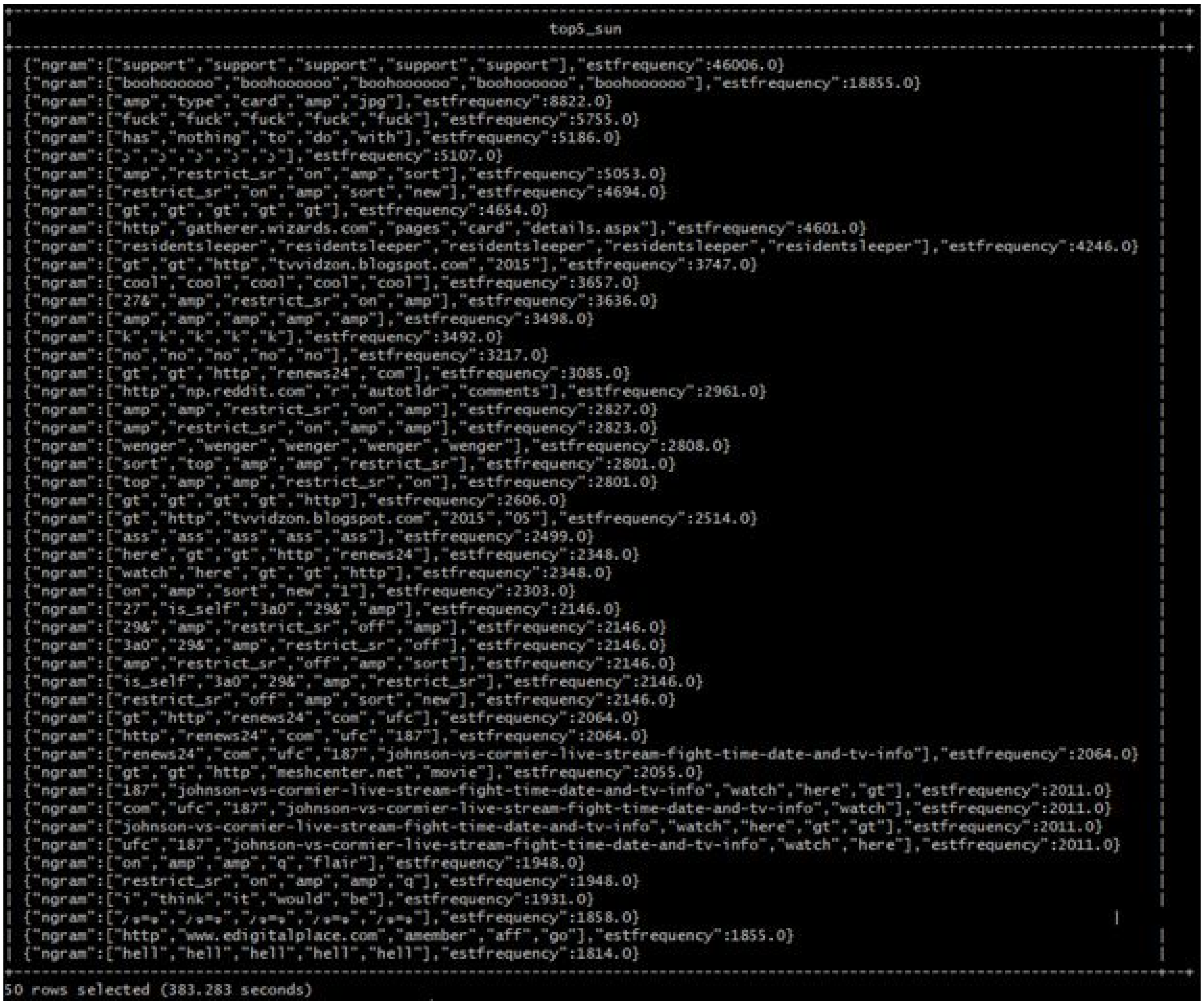
```
select explode(ngrams(sentences(lower(body)), 4, 50)) as top4_sun from may2015
where (created_utc>1430611200 and created_utc<1430697599)
or (created_utc>1431216000 and created_utc<1431302399)
or (created_utc>1431820800 and created_utc<1431907199)
or (created_utc>1432425600 and created_utc<1432511999)
or (created_utc>1433030400 and created_utc<1433116799);
```



```
+-----------------------------------------------------------------------------------------------------------+--+
|                                                 top4_sun                                                  |
+-----------------------------------------------------------------------------------------------------------+--+
| {"ngram":["support","support","support","support"],"estfrequency":46087.0}
| {"ngram":["if","you","want","to"],"estfrequency":21353.0}
| {"ngram":["boohoooooo","boohoooooo","boohoooooo","boohoooooo"],"estfrequency":18917.0}
| {"ngram":["a","lot","of","people"],"estfrequency":16809.0}
| {"ngram":["the","rest","of","the"],"estfrequency":14848.0}
| {"ngram":["the","end","of","the"],"estfrequency":14075.0}
| {"ngram":["at","the","end","of"],"estfrequency":12641.0}
| {"ngram":["to","be","able","to"],"estfrequency":11360.0}
| {"ngram":["when","it","comes","to"],"estfrequency":11021.0}
| {"ngram":["i","don't","want","to"],"estfrequency":10153.0}
| {"ngram":["in","the","first","place"],"estfrequency":9950.0}
| {"ngram":["gt","gt","gt","http"],"estfrequency":9731.0}
| {"ngram":["i","don't","know","if"],"estfrequency":9544.0}
| {"ngram":["nothing","to","do","with"],"estfrequency":9128.0}
| {"ngram":["http","gatherer.wizards.com","handlers","image.ashx"],"estfrequency":8856.0}
| {"ngram":["amp","type","card","amp"],"estfrequency":8826.0}
| {"ngram":["type","card","amp","jpg"],"estfrequency":8826.0}
| {"ngram":["one","of","the","best"],"estfrequency":8811.0}
| {"ngram":["i","have","no","idea"],"estfrequency":8492.0}
| {"ngram":["gt","gt","gt","gt"],"estfrequency":8364.0}
| {"ngram":["amp","restrict_sr","on","amp"],"estfrequency":7944.0}
| {"ngram":["i","don't","know","what"],"estfrequency":7941.0}
| {"ngram":["link","http","www.reddit.com","r"],"estfrequency":7437.0}
| {"ngram":["should","be","able","to"],"estfrequency":7001.0}
| {"ngram":["for","the","first","time"],"estfrequency":6426.0}
| {"ngram":["as","long","as","you"],"estfrequency":6250.0}
| {"ngram":["you","don't","want","to"],"estfrequency":5886.0}
| {"ngram":["fuck","fuck","fuck","fuck"],"estfrequency":5857.0}
| {"ngram":["i","would","like","to"],"estfrequency":5715.0}
| {"ngram":["i","don't","know","why"],"estfrequency":5545.0}
| {"ngram":["ɔ","ɔ","ɔ","ɔ"],"estfrequency":5168.0}
| {"ngram":["there","are","a","lot"],"estfrequency":5102.0}
| {"ngram":["restrict_sr","on","amp","sort"],"estfrequency":5009.0}
| {"ngram":["i","was","going","to"],"estfrequency":4898.0}
| {"ngram":["it","would","be","a"],"estfrequency":4843.0}
| {"ngram":["for","the","rest","of"],"estfrequency":4718.0}
| {"ngram":["on","amp","sort","new"],"estfrequency":4694.0}
| {"ngram":["have","no","idea","what"],"estfrequency":4674.0}
| {"ngram":["amp","amp","amp","amp"],"estfrequency":4664.0}
| {"ngram":["http","gatherer.wizards.com","pages","card"],"estfrequency":4604.0}
| {"ngram":["gatherer.wizards.com","pages","card","details.aspx"],"estfrequency":4601.0}
| {"ngram":["i'm","not","going","to"],"estfrequency":4404.0}
| {"ngram":["i","don't","have","a"],"estfrequency":4270.0}
| {"ngram":["residentsleeper","residentsleeper","residentsleeper","residentsleeper"],"estfrequency":4257.0} |
| {"ngram":["cool","cool","cool","cool"],"estfrequency":3925.0}
| {"ngram":["gt","gt","http","tvvidzon.blogspot.com"],"estfrequency":3747.0}
| {"ngram":["gt","http","tvvidzon.blogspot.com","2015"],"estfrequency":3747.0}
| {"ngram":["gt","http","renews24","com"],"estfrequency":3723.0}
| {"ngram":["27&","amp","restrict_sr","on"],"estfrequency":3636.0}
| {"ngram":["no","no","no","no"],"estfrequency":3559.0}
+-----------------------------------------------------------------------------------------------------------+--+
50 rows selected (397.433 seconds)
```

21. Display the top 5gram on Sunday.

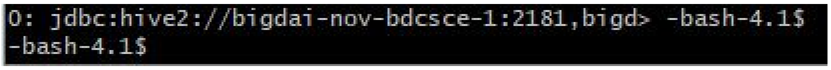
```
select explode(ngrams(sentences(lower(body)), 5, 50)) as top5_sun from may2015
where (created_utc>1430611200 and created_utc<1430697599)
or (created_utc>1431216000 and created_utc<1431302399)
or (created_utc>1431820800 and created_utc<1431907199)
or (created_utc>1432425600 and created_utc<1432511999)
or (created_utc>1433030400 and created_utc<1433116799);
```



```
+----------------------------------------------------------------------------------------------------------------------+--+
|                                                      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":["ɔ","ɔ","ɔ","ɔ","ɔ"],"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":["/ɐ=ɐ","/ɐ=ɐ","/ɐ=ɐ","/ɐ=ɐ","/ɐ=ɐ"],"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)
```

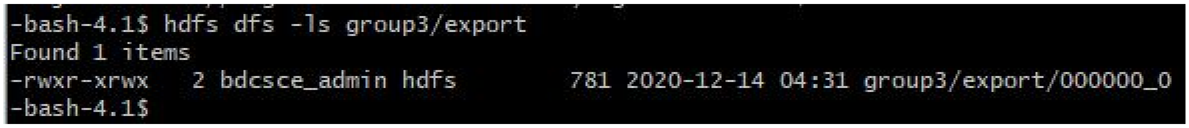
Step 4: Downloading data into your PC

22. Press “Ctrl” + “C” to exit beeline.



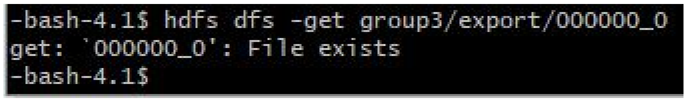
```
0: jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigd> -bash-4.1$
-bash-4.1$
```

23. Confirm file exported from hive table. Refer to step 15.



```
-bash-4.1$ hdfs dfs -ls group3/export
Found 1 items
-rwxr-xrwx   2 bdcsce_admin hdfs        781 2020-12-14 04:31 group3/export/000000_0
-bash-4.1$
```

24. Retrieve the file from hdfs to local.



```
-bash-4.1$ hdfs dfs -get group3/export/000000_0
get: `000000_0': File exists
-bash-4.1$
```

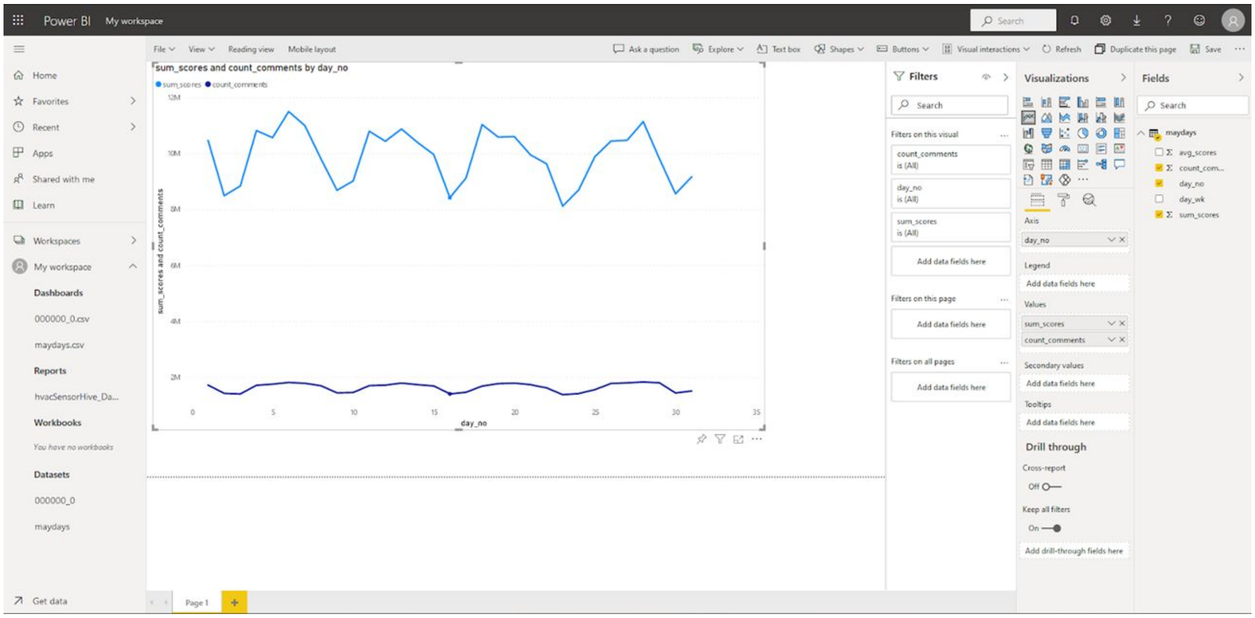
25. Open a new terminal to execute scp command to download the file to your local PC. Please change the text in red to the appropriate user.

    scp dkang10@129.150.64.74:/home/dkang10/000000_0 maydays.csv

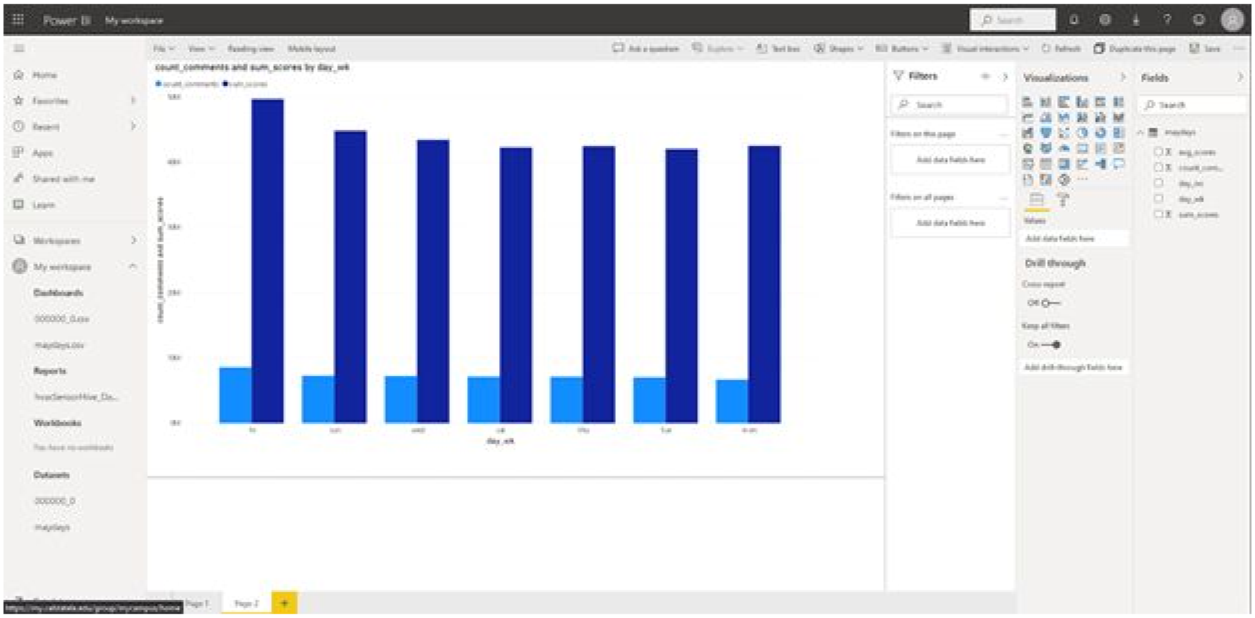
Step 5: Visualization

1. Once the file is on your local PC open it on Microsoft PowerBI. The app can be accessed at app.powerbi.com if you do not have the application on your workstation.

2. Once loaded to PowerBI lets create a graph for the “Total activity of a linear time scale from May 1 to May 31”



3. Lets click add a new page and create a new visual that shows “Activity for each day of the week sorted by comment count”



THIS IS THE END OF THE LAB.

References

1. URL of Data Source: https://www.kaggle.com/reddit/reddit-comments-may-2015

2. URL of your Github:

   https://github.com/r-hathcock/CIS4560-01_Grp03_TermProject

3. “Twitter Sentiment Analysis and n-gram with Hadoop and Hive SQL”,

   https://gist.github.com/umbertogriffo/a512baaf63ce0797e175