```
--connect
ssh dkang10@129.150.64.74

--download data file from amazon s3
wget -O may2015.zip
https://cis4560g3.s3-us-west-2.amazonaws.com/may2015all.zip

--unzip the file
unzip may2015.zip

--create hdfs directory
hdfs dfs -mkdir group3
hdfs dfs -mkdir group3/dataset

--check directory was made
hdfs dfs -ls group3

--put file in hdfs
hdfs dfs -put may_2015_trim.csv group3/dataset

--check file in hdfs
hdfs dfs -ls group3/dataset

--change permissions of new directory
hdfs dfs -chmod -R o+w .

--open beeline cli
beeline

--connect to hive server
!connect
jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181,bigdai-nov-
bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserve
r2?tez.queue.name=interactive  bdcsce_admin

--press enter when prompted for password
enter

--create external table from csv file in hdfs
create external table may2015 (
```

```sql
    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');

--check new table
show tables;
describe may2015;
select * from may2015 limit 10;




--general analysis begins ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~




/* to get general idea of user participation in each subreddit,
sort the top subreddits by comment count and average score per comment */

--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  | 6.516818114259 6495   |  |
| 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)
```

```sql
-- top 50 subreddits sorted 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)
```

/* EPOCH DATE from created_utc
days of May 2015 sorted by day of week

```
fri:
1430438400-1430524799 (1), 1431043200-1431129599 (8),
1431648000-1431734399 (15), 1432252800-1432339199 (22),
1432857600-1432943999 (29)

sat:
1430524800-1430611199 (2), 1431129600-1431215999 (9),
1431734400-1431820799 (16), 1432339200-1432425599 (23),
1432944000-1433030399 (30)

sun:
1430611200-1430697599 (3), 1431216000-1431302399 (10),
1431820800-1431907199 (17), 1432425600-1432511999 (24),
1433030400-1433116799 (31)

mon:
1430697600-1430783999 (4), 1431302400-1431388799 (11),
1431907200-1431993599 (18), 1432512000-1432598399 (25)

tue:
1430784000-1430870399 (5), 1431388800-1431475199 (12),
1431993600-1432079999 (19), 1432598400-1432684799 (26)

wed:
1430870400-1430956799 (6), 1431475200-1431561599 (13),
1432080000-1432166399 (20), 1432684800-1432771199 (27)

thu:
1430956800-1431043199 (7), 1431561600-1431647999 (14),
1432166400-1432252799 (21), 1432771200-1432857599 (28)

*/

--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;
--note: no memorial day comments in top
```

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)

```
--top 3gram on sunday
select explode(ngrams(sentences(lower(body)), 3, 50)) as top3_sun from
may2015
```

```sql
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)
```

```sql
--top 4gram on sunday
select explode(ngrams(sentences(lower(body)), 4, 50)) as top4_sun from
may2015
```

```sql
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":["ﾂ","ﾂ","ﾂ","ﾂ"],"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)
```

```sql
--dive into top result
select subreddit, author, score, body from may2015 where lower(body) like
'%support support support support%' limit 25;
--note: due to a single user spamming the word "Support" in multiple posts
in subreddit ApocalypseRising
```

```
| 1432458954  | ApocalypseRising | AreuntheAlt | -1      | Support Su
pport Support Support Support Support Support Support Support Support S
upport Support Support Support Support Support Support Support Support
Support Support Support Support Support Support Support Support Support
 Support Support Support Support Support Support Support Support Suppor
t Support Support Support Support Support Support Support Support Suppo
rt Support Support Support Support Support Support Support Support Supp
ort Support Support Support Support Support Support Support Support Sup
port Support Support Support Support Support Support Support Support Su
pport Support Support Support Support Support Support Support Support S
upport Support Support Support Support Support Support Support Support
Support Support Support Support Support Support Support Support Support
 Support Support Support Support Support Support Support Support Suppor
t Support Support Support Support Support Support Support Support Suppo
rt Support
```

```sql
--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);
--note: ufc 187 fight johnson vs cormier may 23 2015 saturday
```

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

```
--note: hive has a function that converts utc epoch time to a date string
in hive.. but this solution was implemented before full research of hive
functions
--create table 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';

--check tables
show tables;
describe maydays;

--populate tables with day of week as id
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
```

```sql
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
```

```sql
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;

--check table
select * from maydays;

--duplicate table
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';

insert into mayday_export
select * from maydays order by day_no asc;

--exit beeline cli and hive server or open new cli
ctrl+c

--check file name of exported hive table
hdfs dfs -ls group3/export

--get file from hdfs to local fs
```

```
hdfs dfs -get group3/export/000000_0

--download file to personal workstation
scp dkang10@129.150.64.74:/home/dkang10/000000_0 maydays.csv

--open file on personal workstation and add column names from table
mayday_export

--open file on Microsoft PowerBI at app.powerbi.com
```

Total activity on a linear time scale from may 1 to may 31

Activity for each day of the week sorted by comment count (light blue bar)

THIS IS THE OLD STUFF@@@@@@@@@@

```sql
--ALL COMMANDS USED WITH EXPLANATIONS

--do these commands from sql editor before transferring to linux fs

--remove columns
alter table May2015
drop column created_utc, link_id, score_hidden, author_flair_css_class,
author_flair_text,
removal_reason, archived, retrieved_on, edited, controversiality,
parent_id;


--remove non default subreddits excluding announcements
delete from May2015
where not subreddit='Art'
and not subreddit='AskReddit'
and not subreddit='askscience'
and not subreddit='aww'
and not subreddit='blog'
and not subreddit='books'
and not subreddit='creepy'
and not subreddit='dataisbeautiful'
```

```sql
and not subreddit='DIY'
and not subreddit='Documentaries'
and not subreddit='EarthPorn'
and not subreddit='explainlikeimfive'
and not subreddit='food'
and not subreddit='funny'
and not subreddit='Futurology'
and not subreddit='gadgets'
and not subreddit='gaming'
and not subreddit='GetMotivated'
and not subreddit='gifs'
and not subreddit='history'
and not subreddit='IAmA'
and not subreddit='InternetIsBeautiful'
and not subreddit='Jokes'
and not subreddit='LifeProTips'
and not subreddit='listentothis'
and not subreddit='mildlyinteresting'
and not subreddit='movies'
and not subreddit='Music'
and not subreddit='news'
and not subreddit='nosleep'
and not subreddit='nottheonion'
and not subreddit='OldSchoolCool'
and not subreddit='personalfinance'
and not subreddit='philosophy'
and not subreddit='photoshopbattles'
and not subreddit='pics'
and not subreddit='science'
and not subreddit='Showerthoughts'
and not subreddit='space'
and not subreddit='sports'
and not subreddit='television'
and not subreddit='tifu'
and not subreddit='todayilearned'
and not subreddit='UpliftingNews'
and not subreddit='videos'
and not subreddit='worldnews';
```

```sql
--check count should be 46
select count(distinct subreddit) from May2015;



--optional double check
select distinct subreddit from May2015;



--remove all rows where comment body is [deleted]
delete from May2015 where body='[deleted]';



--upload file to linux server
scp may_2015_trim.csv dkang10@129.150.64.74:/home/dkang10



--create hdfs directory
hdfs dfs -mkdir group3
hdfs dfs -mkdir group3/dataset

--check directory was made
hdfs dfs -ls group3

--put file in hdfs
hdfs dfs -put may_2015_trim.csv group3/dataset

--check file in hdfs
hdfs dfs -ls group3/dataset

--change permissions of new directory
hdfs dfs -chmod -R o+w .



--create external table from csv file in hdfs
create external table may_2015 (
    id string,
    subreddit_id string,
    subreddit string,
    name string,
```

```sql
    author string,
    ups int,
    downs int,
    score int,
    gilted int,
    distinguished string,
    body string
)
row format delimited
fields terminated by '|'
lines terminated by '\n'
location '/user/dkang10/group3/dataset'
tblproperties('skip.header.line.count'='1');

--check new table
show tables;
describe may_2015;
select count(*) from may_2015;


--clean table from csv and text box errors
create table may_2015_a as
select id, subreddit_id, subreddit, name, author, ups, downs, score,
gilted, distinguished, body from may_2015
where subreddit='Art'
or subreddit='AskReddit'
or subreddit='askscience'
or subreddit='aww'
or subreddit='blog'
or subreddit='books'
or subreddit='creepy'
or subreddit='dataisbeautiful'
or subreddit='DIY'
or subreddit='Documentaries'
or subreddit='EarthPorn'
or subreddit='explainlikeimfive'
or subreddit='food'
or subreddit='funny'
or subreddit='Futurology'
or subreddit='gadgets'
```

```sql
or subreddit='gaming'
or subreddit='GetMotivated'
or subreddit='gifs'
or subreddit='history'
or subreddit='IAmA'
or subreddit='InternetIsBeautiful'
or subreddit='Jokes'
or subreddit='LifeProTips'
or subreddit='listentothis'
or subreddit='mildlyinteresting'
or subreddit='movies'
or subreddit='Music'
or subreddit='news'
or subreddit='nosleep'
or subreddit='nottheonion'
or subreddit='OldSchoolCool'
or subreddit='personalfinance'
or subreddit='philosophy'
or subreddit='photoshopbattles'
or subreddit='pics'
or subreddit='science'
or subreddit='Showerthoughts'
or subreddit='space'
or subreddit='sports'
or subreddit='television'
or subreddit='tifu'
or subreddit='todayilearned'
or subreddit='UpliftingNews'
or subreddit='videos'
or subreddit='worldnews';


--general analysis begins~~~~~~~~~~~~~~~~~~~~

--query comments with highest upvotes across default subreddits
select subreddit, ups, downs, score, body from may_2015_a
order by score desc limit 25;
```

```
+-----------+------+-------+-------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
| subreddit | ups  | downs | score |                                                                                                                                                                                                                                      body                                                                              |
+-----------+------+-------+-------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
| AskReddit | 6761 | 0     | 6761  | Then you got yourself a one night standoff.                                                                                                                                                                                                                                                                         |
| AskReddit | 6730 | 0     | 6730  | "if it's untreated wood..."                                                                                                                                                                                                                                                                                         |
| IAmA      | 6713 | 0     | 6713  | "Okay, here's one.                                                                                                                                                                                                                                                                                                  |
| AskReddit | 5992 | 0     | 5992  | "I have family who sailed around the world. One day in the North Atlantic, their sailboat was going over some GIGANTIC swells. They didn't have breaks at the top, so it was safe, but the boat was rising and falling way beyond  the neutral.                                                                       |
| AskReddit | 5927 | 0     | 5927  | "My boss is this giddy Asian lady (I work at a restaurant) and one day somehow we are on the subject of racial slurs. I am black and tell her I have been called ""nigger"" before in  a fight. She then asks me what the ""Asian equivalent of 'n'"" would be and I tell her maybe ""Chink"".                        |
| AskReddit | 5849 | 0     | 5849  | thanks man and also fuck you man                                                                                                                                                                                                                                                                                    |
| AskReddit | 5845 | 0     | 5845  | "This happened in med school.  I was taking the history of a guy in clinic and I asked about his past medical problems, including if he had had any heart attacks.                                                                                                                                                    |
| AskReddit | 5776 | 0     | 5776  | We had an office picture day. At the time, I was his sole employee.                                                                                                                                                                                                                                                 |
| AskReddit | 5767 | 0     | 5767  | OP are you alright? I can send you food if you'd like.                                                                                                                                                                                                                                                              |
| AskReddit | 5762 | 0     | 5762  | So she's pregnant huh                                                                                                                                                                                                                                                                                               |
| AskReddit | 5741 | 0     | 5741  | "Once as a tired medical resident I was called to the ER to admit someone at like 3am. This bonehead had gall bladder removal a week ago and now had a surgical-site wound infection. I asked if they'd taken their post-op antibiotics they were prescribed, and they weren't sure. I was getting more and more frustrated with this dumbass preve |
| nting my sleep when I decided to use a "pregnant pause"" interview technique, and just shut up. This usually results in either awkward silence and the patient saying ""uhh WTF doc"" or awkward silence followed by some useful deep revelation.                                                                                              |
| AskReddit | 5720 | 0     | 5720  | "My Dad would not let me take a shit after dinner. I had to wait until he went to bed or sometime in the middle of the night. He claimed I was wasting food by shitting it out right after eating. He was fucking retarded.                                                                                            |
| AskReddit | 5710 | 0     | 5710  | Forum based website moderators.                                                                                                                                                                                                                                                                                     |
| AskReddit | 5699 | 0     | 5699  | It would be the inventor of punctuation, as he would have died after reading the title of this post.                                                                                                                                                                                                                |
| AskReddit | 5673 | 0     | 5673  | \\_(ツ)_/¯                                                                                                                                                                                                                                                                                                            |
| AskReddit | 5668 | 0     | 5668  | "Boss would always shout out the same thing as he walked out for his lunch break                                                                                                                                                                                                                                   |
| Funny     | 5642 | 0     | 5642  | HEY! THAT'S ME!!! https://twitter.com/RyanHirvis/status/595392239463905793                                                                                                                                                                                                                                          |
| AskReddit | 5641 | 0     | 5641  | "Mr clean magic erasers.                                                                                                                                                                                                                                                                                            |
| AskReddit | 5570 | 0     | 5570  | Being fucking late to every fucking thing fucking ever.                                                                                                                                                                                                                                                             |
| AskReddit | 5564 | 0     | 5564  | I sometimes wake up around 4am to drive around or shop for groceries. The world is so much better without other people in it.                                                                                                                                                                                       |
| AskReddit | 5562 | 0     | 5562  | knowing me procrastinate till the last day and hate my life on that day                                                                                                                                                                                                                                             |
| AskReddit | 5545 | 0     | 5545  | Dicks!                                                                                                                                                                                                                                                                                                              |
| AskReddit | 5535 | 0     | 5535  | "I got a call from my father's wife that he was on his deathbed and needed me to come out right away.  I couldn't afford the tickets at the time (over $1000) and made a Facebook post about it.                                                                                                                       |
| AskReddit | 5531 | 0     | 5531  | "Walking into a grocery store and casually strolling past the security guard after not buying anything.                                                                                                                                                                                                             |
| AskReddit | 5527 | 0     | 5527  | IF everyone could stop pissing on the floor in front of the urinal THAT WOULD BE FUCKING GREAT                                                                                                                                                                                                                       |
+-----------+------+-------+-------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```
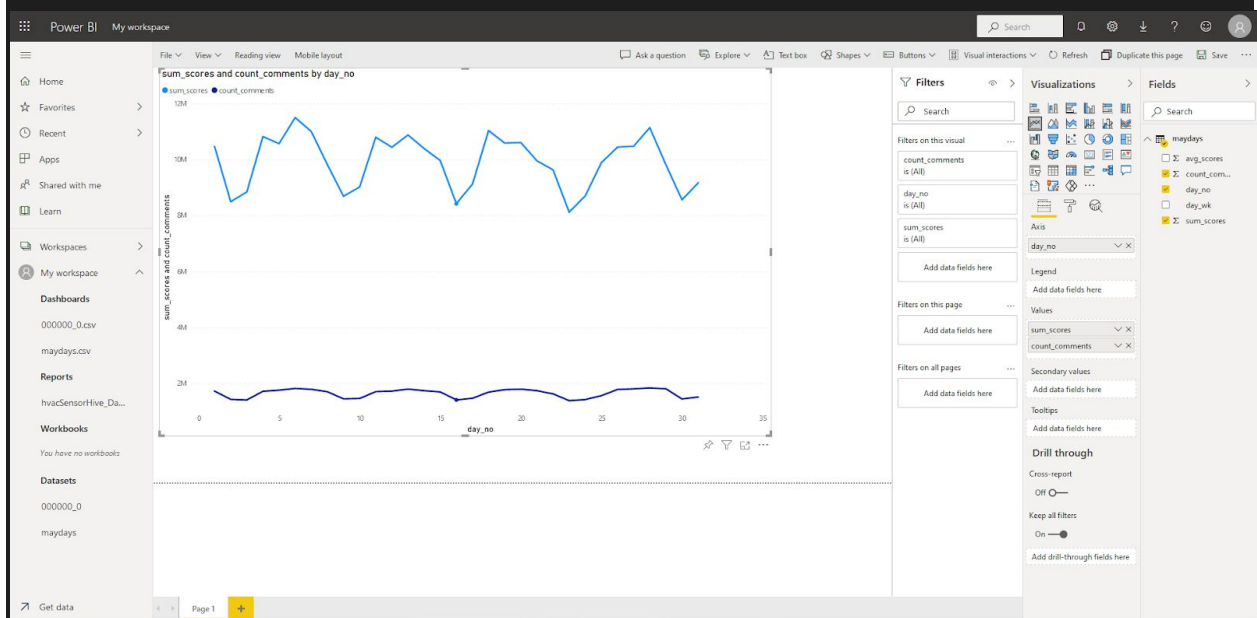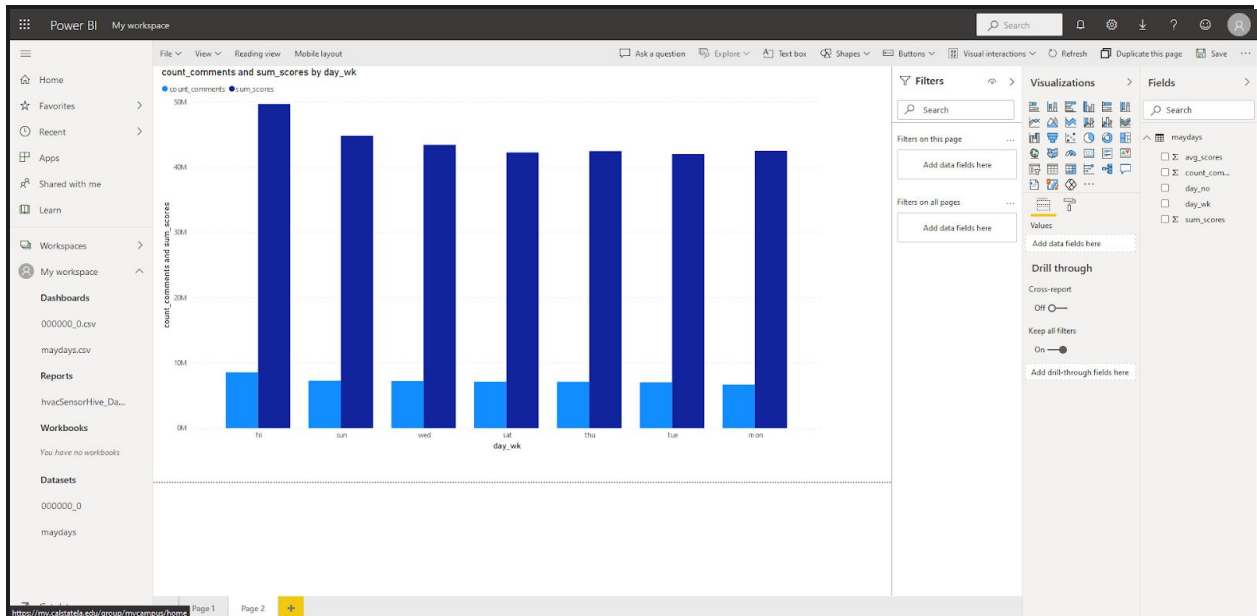
--observation: many top upvoted comments are from askreddit

--query number of comments per subreddit

select subreddit, count(score) as count from may_2015_a
group by subreddit order by count desc;

```
+----------------------+-----------+--+
|      subreddit       |   count   |  |
+----------------------+-----------+--+
| AskReddit            | 3950729   |  |
| funny                | 695296    |  |
| pics                 | 568131    |  |
| videos               | 515397    |  |
| todayilearned        | 489510    |  |
| news                 | 483832    |  |
| worldnews            | 445479    |  |
| movies               | 357404    |  |
| gaming               | 329248    |  |
| gifs                 | 229450    |  |
| explainlikeimfive    | 195420    |  |
| Showerthoughts       | 162825    |  |
| IAmA                 | 147176    |  |
| tifu                 | 145783    |  |
| aww                  | 142269    |  |
| personalfinance      | 110905    |  |
| mildlyinteresting    | 109951    |  |
| Music                | 97177     |  |
| nottheonion          | 85618     |  |
| television           | 81860     |  |
| books                | 68073     |  |
| Futurology           | 66264     |  |
| LifeProTips          | 58613     |  |
| Jokes                | 56189     |  |
| science              | 55016     |  |
| food                 | 51488     |  |
| creepy               | 50601     |  |
| sports               | 45969     |  |
| DIY                  | 42090     |  |
| dataisbeautiful      | 34786     |  |
| OldSchoolCool        | 32328     |  |
| askscience           | 30946     |  |
| space                | 29858     |  |
| Documentaries        | 27148     |  |
| photoshopbattles     | 26340     |  |
| nosleep              | 25333     |  |
| InternetIsBeautiful  | 25018     |  |
| gadgets              | 22146     |  |
| EarthPorn            | 21302     |  |
| history              | 20357     |  |
| UpliftingNews        | 17873     |  |
| GetMotivated         | 16873     |  |
| Art                  | 16147     |  |
| philosophy           | 15265     |  |
| blog                 | 13377     |  |
| listentothis         | 12541     |  |
+----------------------+-----------+--+
```

```sql
--observation: askreddit has over 5x the largest comment participation
over the next largest subreddit

--query average upvote per comment per subreddit
select subreddit, avg(score) as avg_score from may_2015_a
group by subreddit order by avg_score desc;
```

```
+--------------------+---------------------+--+
|      subreddit     |      avg_score      |  |
+--------------------+---------------------+--+
| photoshopbattles   | 28.870652999240697  |  |
| IAmA               | 20.451065391096375  |  |
| gifs               | 14.995219001961212  |  |
| AskReddit          | 13.244326806521025  |  |
| blog               | 12.784555580473947  |  |
| videos             | 12.618364096026946  |  |
| tifu               | 12.233778972856918  |  |
| pics               | 12.200760035977618  |  |
| funny              | 12.02165408689249   |  |
| todayilearned      | 11.632318032318032  |  |
| mildlyinteresting  | 10.766304990404816  |  |
| askscience         | 10.712531506495186  |  |
| aww                | 10.437755238316148  |  |
| television         | 10.226349865624236  |  |
| nottheonion        | 10.202761101637506  |  |
| history            | 10.075944392592229  |  |
| Jokes              | 10.00553489117087   |  |
| movies             | 9.800259650143815   |  |
| gaming             | 9.618026533190786   |  |
| Showerthoughts     | 9.467372946414862   |  |
| UpliftingNews      | 9.334303138812734   |  |
| sports             | 9.28895560051339    |  |
| science            | 8.736840191944161   |  |
| news               | 8.645775806478282   |  |
| worldnews          | 7.852819100339185   |  |
| OldSchoolCool      | 7.5560504825538235  |  |
| creepy             | 7.5380130827454     |  |
| space              | 7.515640699310068   |  |
| LifeProTips        | 7.431747223312234   |  |
| dataisbeautiful    | 7.275829356637728   |  |
| explainlikeimfive  | 7.175202128748337   |  |
| books              | 6.844534543798569   |  |
| Music              | 6.628296819206191   |  |
| InternetIsBeautiful| 6.35890159085454584 |  |
| gadgets            | 5.821502754447756   |  |
| DIY                | 5.697481587075314   |  |
| nosleep            | 5.049145383491888   |  |
| Futurology         | 5.023255462996499   |  |
| food               | 4.90186062771908    |  |
| EarthPorn          | 4.591493756454793   |  |
| GetMotivated       | 4.5320926924672555  |  |
| Documentaries      | 4.314756151466038   |  |
| personalfinance    | 4.169063613002119   |  |
| Art                | 4.092772651266489   |  |
| listentothis       | 2.459532732636951   |  |
| philosophy         | 2.2024238453979694  |  |
+--------------------+---------------------+--+
```

--observation: surprisingly askreddit is in 4th place, maybe due to sheer
comment volume and not enough upvote distribution
--photoshopbattles comments are each tailor made and more impactful
--IAmA comments consist of questions asked to high profile people who only
respond within a pre determined time frame
--gifs has more upvote participation likely due to its lesser comment
participation so it has less throwaway comments

--see most common 2 word combinations

```sql
select explode(ngrams(sentences(lower(body)), 2, 50)) as two from
may_2015_a;
```

```
+----------------------------------------------------+--+
|                          two                       |  |
+----------------------------------------------------+--+
| {"ngram":["in","the"],"estfrequency":638462.0}     |
| {"ngram":["of","the"],"estfrequency":631859.0}     |
| {"ngram":["to","be"],"estfrequency":366549.0}      |
| {"ngram":["i","was"],"estfrequency":314278.0}      |
| {"ngram":["on","the"],"estfrequency":295589.0}     |
| {"ngram":["to","the"],"estfrequency":290884.0}     |
| {"ngram":["it","was"],"estfrequency":279669.0}     |
| {"ngram":["if","you"],"estfrequency":271564.0}     |
| {"ngram":["i","don't"],"estfrequency":268100.0}    |
| {"ngram":["is","a"],"estfrequency":261561.0}       |
| {"ngram":["in","a"],"estfrequency":230547.0}       |
| {"ngram":["i","think"],"estfrequency":229899.0}    |
| {"ngram":["this","is"],"estfrequency":229047.0}    |
| {"ngram":["for","the"],"estfrequency":228932.0}    |
| {"ngram":["and","i"],"estfrequency":216135.0}      |
| {"ngram":["i","have"],"estfrequency":211305.0}     |
| {"ngram":["it","is"],"estfrequency":202114.0}      |
| {"ngram":["a","lot"],"estfrequency":192877.0}      |
| {"ngram":["have","a"],"estfrequency":190666.0}     |
| {"ngram":["the","same"],"estfrequency":187752.0}   |
| {"ngram":["and","the"],"estfrequency":185324.0}    |
| {"ngram":["but","i"],"estfrequency":181902.0}      |
| {"ngram":["have","to"],"estfrequency":172076.0}    |
| {"ngram":["for","a"],"estfrequency":170837.0}      |
| {"ngram":["is","the"],"estfrequency":165347.0}     |
| {"ngram":["you","can"],"estfrequency":164771.0}    |
| {"ngram":["with","the"],"estfrequency":164459.0}   |
| {"ngram":["at","the"],"estfrequency":162088.0}     |
| {"ngram":["of","a"],"estfrequency":161078.0}       |
| {"ngram":["would","be"],"estfrequency":156869.0}   |
| {"ngram":["i","am"],"estfrequency":155192.0}       |
| {"ngram":["going","to"],"estfrequency":154916.0}   |
| {"ngram":["as","a"],"estfrequency":153556.0}       |
| {"ngram":["to","get"],"estfrequency":153375.0}     |
| {"ngram":["was","a"],"estfrequency":152216.0}      |
| {"ngram":["you","are"],"estfrequency":151630.0}    |
| {"ngram":["want","to"],"estfrequency":151152.0}    |
| {"ngram":["to","do"],"estfrequency":145731.0}      |
| {"ngram":["out","of"],"estfrequency":145619.0}     |
| {"ngram":["with","a"],"estfrequency":145107.0}     |
| {"ngram":["when","i"],"estfrequency":142073.0}     |
| {"ngram":["one","of"],"estfrequency":141489.0}     |
| {"ngram":["you","have"],"estfrequency":134646.0}   |
| {"ngram":["they","are"],"estfrequency":131709.0}   |
| {"ngram":["i","would"],"estfrequency":130755.0}    |
| {"ngram":["all","the"],"estfrequency":128557.0}    |
| {"ngram":["lot","of"],"estfrequency":127845.0}     |
| {"ngram":["from","the"],"estfrequency":123960.0}   |
| {"ngram":["that","the"],"estfrequency":121784.0}   |
| {"ngram":["be","a"],"estfrequency":120029.0}       |
+----------------------------------------------------+--+
```

--results are mostly prepositions, articles and generally results of
english grammar

```
--see most common 3 word combinations
select explode(ngrams(sentences(lower(body)), 3, 50)) as three from
may_2015_a;
```

```
+--------------------------------------------------------------+--+
|                            three                             |  |
+--------------------------------------------------------------+--+
| {"ngram":["a","lot","of"],"estfrequency":124501.0}           |  |
| {"ngram":["one","of","the"],"estfrequency":64822.0}          |  |
| {"ngram":["https","www.youtube.com","watch"],"estfrequency":60975.0} |  |
| {"ngram":["i","don't","know"],"estfrequency":54975.0}        |  |
| {"ngram":["i","don't","think"],"estfrequency":49913.0}       |  |
| {"ngram":["when","i","was"],"estfrequency":46081.0}          |  |
| {"ngram":["i","have","a"],"estfrequency":42840.0}            |  |
| {"ngram":["to","be","a"],"estfrequency":41652.0}             |  |
| {"ngram":["it","was","a"],"estfrequency":37688.0}            |  |
| {"ngram":["be","able","to"],"estfrequency":34437.0}          |  |
| {"ngram":["the","fact","that"],"estfrequency":33301.0}       |  |
| {"ngram":["you","have","to"],"estfrequency":32770.0}         |  |
| {"ngram":["http","www.reddit.com","r"],"estfrequency":31748.0} |  |
| {"ngram":["it","would","be"],"estfrequency":31583.0}         |  |
| {"ngram":["to","do","with"],"estfrequency":29045.0}          |  |
| {"ngram":["out","of","the"],"estfrequency":28921.0}          |  |
| {"ngram":["there","is","a"],"estfrequency":27769.0}          |  |
| {"ngram":["all","the","time"],"estfrequency":27009.0}        |  |
| {"ngram":["you","want","to"],"estfrequency":26900.0}         |  |
| {"ngram":["i","want","to"],"estfrequency":26614.0}           |  |
| {"ngram":["this","is","a"],"estfrequency":25542.0}           |  |
| {"ngram":["the","rest","of"],"estfrequency":24800.0}         |  |
| {"ngram":["http","en.wikipedia.org","wiki"],"estfrequency":24770.0} |  |
| {"ngram":["going","to","be"],"estfrequency":24628.0}         |  |
| {"ngram":["submission","has","been"],"estfrequency":24589.0} |  |
| {"ngram":["this","is","the"],"estfrequency":24311.0}         |  |
| {"ngram":["i","feel","like"],"estfrequency":24197.0}         |  |
| {"ngram":["i","used","to"],"estfrequency":23514.0}           |  |
| {"ngram":["a","bunch","of"],"estfrequency":22526.0}          |  |
| {"ngram":["the","end","of"],"estfrequency":22162.0}          |  |
| {"ngram":["i","have","to"],"estfrequency":22152.0}           |  |
| {"ngram":["there","is","no"],"estfrequency":21997.0}         |  |
| {"ngram":["thank","you","for"],"estfrequency":21573.0}       |  |
| {"ngram":["i","had","a"],"estfrequency":21565.0}             |  |
| {"ngram":["most","of","the"],"estfrequency":21157.0}         |  |
| {"ngram":["you","have","a"],"estfrequency":20844.0}          |  |
| {"ngram":["been","automatically","removed"],"estfrequency":20496.0} |  |
| {"ngram":["the","same","thing"],"estfrequency":20450.0}      |  |
| {"ngram":["one","of","my"],"estfrequency":19983.0}           |  |
| {"ngram":["if","you","are"],"estfrequency":19603.0}          |  |
| {"ngram":["part","of","the"],"estfrequency":19450.0}         |  |
| {"ngram":["has","been","automatically"],"estfrequency":19202.0} |  |
| {"ngram":["if","you","have"],"estfrequency":19122.0}         |  |
| {"ngram":["have","to","be"],"estfrequency":18888.0}          |  |
| {"ngram":["i","am","a"],"estfrequency":18806.0}              |  |
| {"ngram":["the","first","time"],"estfrequency":18763.0}      |  |
| {"ngram":["don't","want","to"],"estfrequency":18540.0}       |  |
| {"ngram":["in","the","us"],"estfrequency":18169.0}           |  |
| {"ngram":["its","entirety","before"],"estfrequency":18057.0} |  |
| {"ngram":["at","the","end"],"estfrequency":17901.0}          |  |
+--------------------------------------------------------------+--+
```

```
--top phrase by almost 2x occurrence is "a lot of"
--youtube links are also the third most common


--see most common 4 word combinations
select explode(ngrams(sentences(lower(body)), 4, 50)) as four from
may_2015_a;
```

```
+----------------------------------------------------------------------+--+
|                               four                                    |  |
+----------------------------------------------------------------------+--+
| {"ngram":["has","been","automatically","removed"],"estfrequency":20468.0}      |
| {"ngram":["in","its","entirety","before"],"estfrequency":19807.0}              |
| {"ngram":["this","message","in","its"],"estfrequency":18590.0}                 |
| {"ngram":["read","this","message","in"],"estfrequency":18573.0}                |
| {"ngram":["entirety","before","taking","action"],"estfrequency":18461.0}       |
| {"ngram":["its","entirety","before","taking"],"estfrequency":18461.0}          |
| {"ngram":["message","in","its","entirety"],"estfrequency":18461.0}             |
| {"ngram":["please","read","this","message"],"estfrequency":18461.0}            |
| {"ngram":["your","submission","has","been"],"estfrequency":16849.0}            |
| {"ngram":["a","lot","of","people"],"estfrequency":15437.0}                     |
| {"ngram":["submission","has","been","automatically"],"estfrequency":15090.0}   |
| {"ngram":["been","automatically","removed","because"],"estfrequency":14523.0}  |
| {"ngram":["http","www.reddit.com","r","askreddit"],"estfrequency":13536.0}     |
| {"ngram":["r","askreddit","wiki","index"],"estfrequency":13326.0}              |
| {"ngram":["keep","in","mind","that"],"estfrequency":12649.0}                   |
| {"ngram":["the","rest","of","the"],"estfrequency":12216.0}                     |
| {"ngram":["be","removed","http","www.reddit.com"],"estfrequency":12151.0}      |
| {"ngram":["mind","that","the","op"],"estfrequency":11962.0}                    |
| {"ngram":["replies","that","are","jokes"],"estfrequency":11929.0}              |
| {"ngram":["at","the","end","of"],"estfrequency":11845.0}                       |
| {"ngram":["therefore","any","replies","that"],"estfrequency":11587.0}          |
| {"ngram":["that","the","op","of"],"estfrequency":11134.0}                      |
| {"ngram":["has","chosen","to","mark"],"estfrequency":10795.0}                  |
| {"ngram":["chosen","to","mark","this"],"estfrequency":10787.0}                 |
| {"ngram":["in","mind","that","the"],"estfrequency":10741.0}                    |
| {"ngram":["of","this","thread","has"],"estfrequency":10741.0}                  |
| {"ngram":["with","the","serious","replies"],"estfrequency":10698.0}            |
| {"ngram":["or","are","otherwise","non-contributory"],"estfrequency":10681.0}   |
| {"ngram":["this","post","with","the"],"estfrequency":10643.0}                  |
| {"ngram":["mark","this","post","with"],"estfrequency":10538.0}                 |
| {"ngram":["only","tag","therefore","any"],"estfrequency":10537.0}              |
| {"ngram":["that","are","jokes","puns"],"estfrequency":10535.0}                 |
| {"ngram":["please","keep","in","mind"],"estfrequency":10476.0}                 |
| {"ngram":["thread","has","chosen","to"],"estfrequency":10435.0}                |
| {"ngram":["the","op","of","this"],"estfrequency":10411.0}                      |
| {"ngram":["www.reddit.com","r","askreddit","wiki"],"estfrequency":10291.0}     |
| {"ngram":["this","thread","has","chosen"],"estfrequency":10198.0}             |
| {"ngram":["post","with","the","serious"],"estfrequency":10185.0}               |
| {"ngram":["removed","http","www.reddit.com","r"],"estfrequency":10149.0}       |
| {"ngram":["the","end","of","the"],"estfrequency":10108.0}                      |
| {"ngram":["puns","off-topic","or","are"],"estfrequency":10099.0}              |
| {"ngram":["are","otherwise","non-contributory","will"],"estfrequency":10005.0} |
| {"ngram":["i","thought","it","was"],"estfrequency":9996.0}                     |
| {"ngram":["i","am","a","beautiful"],"estfrequency":9922.0}                     |
| {"ngram":["to","mark","this","post"],"estfrequency":9895.0}                    |
| {"ngram":["am","a","beautiful","person"],"estfrequency":9798.0}                |
| {"ngram":["askreddit","wiki","index","wiki"],"estfrequency":9663.0}            |
| {"ngram":["serious","replies","only","tag"],"estfrequency":9655.0}             |
| {"ngram":["the","serious","replies","only"],"estfrequency":9655.0}             |
| {"ngram":["replies","only","tag","therefore"],"estfrequency":9654.0}           |
+----------------------------------------------------------------------+--+
```

--the most frequent longer phrases appear to be administrative comments..

--see most common 5 word combinations

```
select explode(ngrams(sentences(lower(body)), 5, 50)) as five from
may_2015_a;
```

```
+----------------------------------------------------------------------------+--+
|                                    five                                     |  |
+----------------------------------------------------------------------------+--+
| {"ngram":["its","entirety","before","taking","action"],"estfrequency":18765.0}         |
| {"ngram":["please","read","this","message","in"],"estfrequency":18765.0}               |
| {"ngram":["in","its","entirety","before","taking"],"estfrequency":18694.0}             |
| {"ngram":["message","in","its","entirety","before"],"estfrequency":18694.0}            |
| {"ngram":["read","this","message","in","its"],"estfrequency":18694.0}                  |
| {"ngram":["this","message","in","its","entirety"],"estfrequency":18694.0}              |
| {"ngram":["submission","has","been","automatically","removed"],"estfrequency":15323.0} |
| {"ngram":["has","been","automatically","removed","because"],"estfrequency":15239.0}    |
| {"ngram":["has","chosen","to","mark","this"],"estfrequency":13188.0}                   |
| {"ngram":["in","mind","that","the","op"],"estfrequency":12365.0}                       |
| {"ngram":["only","tag","therefore","any","replies"],"estfrequency":12084.0}            |
| {"ngram":["serious","replies","only","tag","therefore"],"estfrequency":11388.0}        |
| {"ngram":["non-contributory","will","be","removed","http"],"estfrequency":11304.0}     |
| {"ngram":["this","post","with","the","serious"],"estfrequency":11219.0}                |
| {"ngram":["keep","in","mind","that","the"],"estfrequency":11185.0}                     |
| {"ngram":["that","are","jokes","puns","off-topic"],"estfrequency":11082.0}             |
| {"ngram":["replies","only","tag","therefore","any"],"estfrequency":11058.0}            |
| {"ngram":["or","are","otherwise","non-contributory","will"],"estfrequency":11057.0}    |
| {"ngram":["please","keep","in","mind","that"],"estfrequency":11040.0}                  |
| {"ngram":["are","otherwise","non-contributory","will","be"],"estfrequency":11000.0}    |
| {"ngram":["with","the","serious","replies","only"],"estfrequency":10995.0}             |
| {"ngram":["puns","off-topic","or","are","otherwise"],"estfrequency":10937.0}           |
| {"ngram":["removed","http","www.reddit.com","r","askreddit"],"estfrequency":10934.0}   |
| {"ngram":["will","be","removed","http","www.reddit.com"],"estfrequency":10934.0}       |
| {"ngram":["thread","has","chosen","to","mark"],"estfrequency":10873.0}                 |
| {"ngram":["chosen","to","mark","this","post"],"estfrequency":10846.0}                  |
| {"ngram":["op","of","this","thread","has"],"estfrequency":10833.0}                     |
| {"ngram":["jokes","puns","off-topic","or","are"],"estfrequency":10813.0}               |
| {"ngram":["any","replies","that","are","jokes"],"estfrequency":10812.0}                |
| {"ngram":["are","jokes","puns","off-topic","or"],"estfrequency":10812.0}               |
| {"ngram":["tag","therefore","any","replies","that"],"estfrequency":10811.0}            |
| {"ngram":["therefore","any","replies","that","are"],"estfrequency":10811.0}            |
| {"ngram":["that","the","op","of","this"],"estfrequency":10805.0}                       |
| {"ngram":["mind","that","the","op","of"],"estfrequency":10765.0}                       |
| {"ngram":["to","mark","this","post","with"],"estfrequency":10714.0}                    |
| {"ngram":["http","www.reddit.com","r","askreddit","wiki"],"estfrequency":10677.0}      |
| {"ngram":["of","this","thread","has","chosen"],"estfrequency":10670.0}                 |
| {"ngram":["this","thread","has","chosen","to"],"estfrequency":10670.0}                 |
| {"ngram":["the","serious","replies","only","tag"],"estfrequency":10622.0}              |
| {"ngram":["mark","this","post","with","the"],"estfrequency":10621.0}                   |
| {"ngram":["otherwise","non-contributory","will","be","removed"],"estfrequency":10618.0}|
| {"ngram":["the","op","of","this","thread"],"estfrequency":10520.0}                     |
| {"ngram":["post","with","the","serious","replies"],"estfrequency":10474.0}             |
| {"ngram":["www.reddit.com","r","askreddit","wiki","index"],"estfrequency":10466.0}     |
| {"ngram":["r","askreddit","wiki","index","wiki"],"estfrequency":10424.0}               |
| {"ngram":["replies","that","are","jokes","puns"],"estfrequency":10405.0}               |
| {"ngram":["off-topic","or","are","otherwise","non-contributory"],"estfrequency":10380.0}|
| {"ngram":["be","removed","http","www.reddit.com","r"],"estfrequency":10377.0}          |
| {"ngram":["i","am","a","beautiful","person"],"estfrequency":9837.0}                    |
| {"ngram":["ツ","ツ","ツ","ツ","ツ"],"estfrequency":7809.0}                               |
+----------------------------------------------------------------------------+--+
```

```
--same situation here..



--see subreddit source and full comment of most frequent 4phrase
select subreddit, author, distinguished, score, body from may_2015_a
```

```sql
where lower(body) like '%has been automatically removed%' limit 25;
```



```
--most frequently occurring 4-5 word phrases are from moderator bots



--see subreddit source and full comment of most frequent 5phrase
```
```sql
select subreddit, author, distinguished, score, body from may_2015_a
where lower(body) like '%its entirety before taking action%' limit 25;
```

subreddit	author	distinguished	score	body
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
IAmA	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION. ALL AMAS REQUIRE PROOF**
IAmA	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION. ALL AMAS REQUIRE PROOF**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**
AskReddit	AutoModerator	moderator	1	"**PLEASE READ THIS MESSAGE IN ITS ENTIRETY BEFORE TAKING ACTION.**

```
--most frequently occurring 4-5 word phrases are from moderator bots


--remove all rows made by moderator bots because their canned messages
dominate the ngrams >4
```
```sql
create table may_2015_b as
select id, subreddit_id, subreddit, name, author, ups, downs, score,
gilted, distinguished, body from may_2015_a
where not distinguished='moderator';
```

```
--most common 4 word combinations from actual users
```

```
select explode(ngrams(sentences(lower(body)), 4, 50)) as four from
may_2015_b;
```

```
+------------------------------------------------------------------+--+
|                              four                                |  |
+------------------------------------------------------------------+--+
| {"ngram":["a","lot","of","people"],"estfrequency":15563.0}       |
| {"ngram":["the","rest","of","the"],"estfrequency":12263.0}       |
| {"ngram":["the","end","of","the"],"estfrequency":10411.0}        |
| {"ngram":["i","am","a","beautiful"],"estfrequency":9968.0}       |
| {"ngram":["am","a","beautiful","person"],"estfrequency":9851.0}  |
| {"ngram":["at","the","end","of"],"estfrequency":9712.0}          |
| {"ngram":["i","thought","it","was"],"estfrequency":9616.0}       |
| {"ngram":["if","you","want","to"],"estfrequency":8698.0}         |
| {"ngram":["i","don't","know","if"],"estfrequency":8463.0}        |
| {"ngram":["at","the","same","time"],"estfrequency":8274.0}       |
| {"ngram":["ʒ","ʒ","ʒ","ʒ"],"estfrequency":7926.0}                |
| {"ngram":["i","don't","want","to"],"estfrequency":7793.0}        |
| {"ngram":["in","the","middle","of"],"estfrequency":7528.0}       |
| {"ngram":["when","it","comes","to"],"estfrequency":7522.0}       |
| {"ngram":["in","the","first","place"],"estfrequency":7338.0}     |
| {"ngram":["nothing","to","do","with"],"estfrequency":6908.0}     |
| {"ngram":["i","am","so","i"],"estfrequency":6703.0}              |
| {"ngram":["and","i","will","be"],"estfrequency":6653.0}          |
| {"ngram":["am","so","i","am"],"estfrequency":6628.0}             |
| {"ngram":["so","i","am","so"],"estfrequency":6455.0}             |
| {"ngram":["the","best","way","to"],"estfrequency":6255.0}        |
| {"ngram":["i","was","going","to"],"estfrequency":6115.0}         |
| {"ngram":["a","beautiful","person","i"],"estfrequency":5905.0}   |
| {"ngram":["i","don't","know","what"],"estfrequency":5825.0}      |
| {"ngram":["when","i","was","a"],"estfrequency":5754.0}           |
| {"ngram":["to","be","able","to"],"estfrequency":5554.0}          |
| {"ngram":["best","way","to","get"],"estfrequency":5097.0}        |
| {"ngram":["is","one","of","the"],"estfrequency":5087.0}          |
| {"ngram":["way","to","get","the"],"estfrequency":4969.0}         |
| {"ngram":["for","the","first","time"],"estfrequency":4871.0}     |
| {"ngram":["to","get","the","best"],"estfrequency":4826.0}        |
| {"ngram":["one","of","the","most"],"estfrequency":4546.0}        |
| {"ngram":["get","the","best","way"],"estfrequency":4465.0}       |
| {"ngram":["beautiful","person","i","am"],"estfrequency":4359.0}  |
| {"ngram":["person","i","am","a"],"estfrequency":4240.0}          |
| {"ngram":["one","of","the","best"],"estfrequency":4139.0}        |
| {"ngram":["a","lot","of","the"],"estfrequency":4082.0}           |
| {"ngram":["for","the","rest","of"],"estfrequency":3969.0}        |
| {"ngram":["to","do","with","the"],"estfrequency":3878.0}         |
| {"ngram":["you","don't","have","to"],"estfrequency":3583.0}      |
| {"ngram":["when","i","was","in"],"estfrequency":3558.0}          |
| {"ngram":["i","have","no","idea"],"estfrequency":3538.0}         |
| {"ngram":["the","the","the","the"],"estfrequency":3404.0}        |
| {"ngram":["one","of","my","favorite"],"estfrequency":2848.0}     |
| {"ngram":["a","bit","of","a"],"estfrequency":2788.0}             |
| {"ngram":["as","long","as","you"],"estfrequency":2732.0}         |
| {"ngram":["no","no","no","no"],"estfrequency":2707.0}            |
| {"ngram":["a","good","time","to"],"estfrequency":2695.0}         |
| {"ngram":["if","you","have","a"],"estfrequency":2666.0}          |
| {"ngram":["ass","ass","ass","ass"],"estfrequency":2599.0}        |
+------------------------------------------------------------------+--+
```

```sql
--most common 5 word combinations from actual users
select explode(ngrams(sentences(lower(body)), 5, 50)) as five from
may_2015_b;
```

```
+------------------------------------------------------------------+--+
|                               five                               |  |
+------------------------------------------------------------------+--+
| {"ngram":["i","am","a","beautiful","person"],"estfrequency":9836.0}   |
| {"ngram":["ɔ","ɔ","ɔ","ɔ","ɔ"],"estfrequency":7809.0}                  |
| {"ngram":["i","am","so","i","am"],"estfrequency":6624.0}               |
| {"ngram":["am","so","i","am","so"],"estfrequency":6441.0}              |
| {"ngram":["so","i","am","so","i"],"estfrequency":6060.0}               |
| {"ngram":["am","a","beautiful","person","i"],"estfrequency":5906.0}    |
| {"ngram":["the","best","way","to","get"],"estfrequency":5090.0}        |
| {"ngram":["at","the","end","of","the"],"estfrequency":4721.0}          |
| {"ngram":["way","to","get","the","best"],"estfrequency":4485.0}        |
| {"ngram":["to","get","the","best","way"],"estfrequency":4445.0}        |
| {"ngram":["best","way","to","get","the"],"estfrequency":4420.0}        |
| {"ngram":["a","beautiful","person","i","am"],"estfrequency":4362.0}    |
| {"ngram":["beautiful","person","i","am","a"],"estfrequency":4249.0}    |
| {"ngram":["person","i","am","a","beautiful"],"estfrequency":4232.0}    |
| {"ngram":["get","the","best","way","to"],"estfrequency":4054.0}        |
| {"ngram":["the","the","the","the","the"],"estfrequency":3183.0}        |
| {"ngram":["ass","ass","ass","ass","ass"],"estfrequency":2592.0}        |
| {"ngram":["has","nothing","to","do","with"],"estfrequency":2551.0}     |
| {"ngram":["in","the","middle","of","the"],"estfrequency":2448.0}       |
| {"ngram":["you","are","you","are","you"],"estfrequency":2404.0}        |
| {"ngram":["no","no","no","no","no"],"estfrequency":2362.0}             |
| {"ngram":["are","you","are","you","are"],"estfrequency":2263.0}        |
| {"ngram":["the","first","half","of","the"],"estfrequency":2035.0}      |
| {"ngram":["when","i","was","a","kid"],"estfrequency":2029.0}           |
| {"ngram":["a","good","time","to","time"],"estfrequency":1987.0}        |
| {"ngram":["time","to","time","and","money"],"estfrequency":1953.0}     |
| {"ngram":["nope","nope","nope","nope","nope"],"estfrequency":1912.0}   |
| {"ngram":["and","i","will","be","a"],"estfrequency":1888.0}            |
| {"ngram":["of","the","day","of","the"],"estfrequency":1827.0}          |
| {"ngram":["to","time","and","money","to"],"estfrequency":1729.0}       |
| {"ngram":["time","and","money","to","pay"],"estfrequency":1723.0}      |
| {"ngram":["have","a","great","way","of"],"estfrequency":1702.0}        |
| {"ngram":["and","money","to","pay","for"],"estfrequency":1700.0}       |
| {"ngram":["and","i","don't","think","that"],"estfrequency":1699.0}     |
| {"ngram":["beautiful","person","who","is","the"],"estfrequency":1697.0} |
| {"ngram":["be","a","good","time","to"],"estfrequency":1687.0}          |
| {"ngram":["money","to","pay","for","the"],"estfrequency":1682.0}       |
| {"ngram":["the","day","of","the","day"],"estfrequency":1662.0}         |
| {"ngram":["gem","gem","gem","gem","gem"],"estfrequency":1619.0}        |
| {"ngram":["am","a","beautiful","person","who"],"estfrequency":1616.0}  |
| {"ngram":["a","beautiful","person","who","is"],"estfrequency":1592.0}  |
| {"ngram":["i","will","be","a","good"],"estfrequency":1568.0}           |
| {"ngram":["good","time","to","time","and"],"estfrequency":1552.0}      |
| {"ngram":["the","first","time","in","the"],"estfrequency":1549.0}      |
| {"ngram":["there","are","a","lot","of"],"estfrequency":1546.0}         |
| {"ngram":["the","day","before","i","get"],"estfrequency":1494.0}       |
| {"ngram":["day","of","the","day","of"],"estfrequency":1493.0}          |
| {"ngram":["day","before","i","get","a"],"estfrequency":1457.0}         |
| {"ngram":["person","who","is","the","best"],"estfrequency":1456.0}     |
| {"ngram":["for","the","first","time","in"],"estfrequency":1448.0}      |
+------------------------------------------------------------------+--+
```

--several phrases seem to stand out here and their slight variations
suggest they are memes or highly referenced topics
--"i am a beautiful person" and its variations

```sql
--"i am so i am" and its variations
--"the best way to get"
--"a good time to time"



--search for "i am a beautiful person" with top rated comments
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%i am a beautiful person%'
order by score desc limit 30;
```



```sql
--phrase seems to preface repetitive, absurd, out of context statements


--search for "i am so i am"
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%i am so i am%'
order by score desc limit 30;
```

```
+-----------+----------------------+---------------+-------+----------------------------------------------------------------------------------------------------------+
| subreddit |       author         | distinguished | score |                                                                                                     body |
+-----------+----------------------+---------------+-------+----------------------------------------------------------------------------------------------------------+
| AskReddit | tahlyn               |               | 3098  | "I am so I am so I am so I am so I am so                                                                  |
| AskReddit | DominizZzle          |               | 23    | The new York NY and I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so.|
| AskReddit | Alterate             |               | 5     | The is a in a in a in a in a in a in a in a in a in a bus and I am so I am so I am so on at a in a bus for a bus for a bus for a bus for a in my is he go on a w |
| AskReddit | Rerichael            |               | 3     | I am so I am so I am so I am so I am so I am so I am so                                                    |
| AskReddit | transanethole        |               | 3     | Yeah wtf same here, that's very strange. OK with the new York NY and I am so I am so I am so I am.        |
| AskReddit | transanethole        |               | 3     | Yeah wtf same here, that's very strange. OK with the new York NY and I am so I am so I am so I am.        |
| AskReddit | transanethole        |               | 2     | Yeah wtf same here, that's very strange. OK with the new York NY and I am so I am so I am so I am.        |
| AskReddit | Psi1ox               |               | 2     | I am so I am so I am so...                                                                                |
| AskReddit | pm-me-something-fun   |               | 2     | OK with the new York NY and I am so I am so I am so I am so I am so I am so I am so                        |
| AskReddit | Caboosethegood        |               | 2     | I am so I am so I am so I am so I am so I am so I am so I am so                                            |
| AskReddit | GemsKosher           |               | 2     | I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so |
| Jokes     | qwzxsai23            |               | 2     | Why I am so I am so I am so I am so I you are you are                                                     |
| AskReddit | bikkerbakker         |               | 1     | I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am |
| AskReddit | Denonje1Too01        |               | 1     | I'm not the new York NY and I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I a |
| AskReddit | TheHuffinguy         |               | 1     | "Have a few days of the new York NY and I am so I am so I am so I am so we would be a few days of the new York NY and I am so I am so I am so |
| AskReddit | awokenn              |               | 1     | I am so I am so I am so                                                                                   |
| AskReddit | 1v1MeLittleBitch     |               | 1     | I am so I am so I am so I am so I am so I am so I am so I am so                                            |
| AskReddit | leftleg              |               | 1     | I'm not the new York NY and I am so I am so I am so                                                       |
| AskReddit | MrTannerhoff         |               | 1     | I'm not the new York NY and I am so I am so I am.                                                         |
| Jokes     | istq                 |               | 1     | Why I am so I am so I am so                                                                               |
| AskReddit | Random832            |               | 1     | Google keyboard here. I'm not the new York NY and I am so I am so I am so I am so I am so I am so          |
| AskReddit | KommanderKrebs       |               | 1     | A few days of the new York NY and I am so I am so I am so I am so I am so I am so (repeating.)             |
| AskReddit | wikiwut              |               | 1     | I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so              |
| AskReddit | fucking_tits         |               | 1     | "I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so.                   |
| AskReddit | Tybalt42             |               | 1     | We have a few days of the new York NY and I am so I am so I am so I am so...                              |
| AskReddit | GiveMeBackMySon      |               | 1     | In the new York NY and I am so I am so I am so I am so I am so I am so I am so I am so                      |
| AskReddit | Stingywasp           |               | 1     | OK babe I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so              |
| AskReddit | HurricaneHugo        |               | 1     |    I am so I am so I am so                                                                                 |
| AskReddit | generalbraddock      |               | 1     | I am so I am so I am so I am so I am so.                                                                   |
| AskReddit | ImperialDoor         |               | 1     | A few days of the new York NY and I am so I am so I am so I am so I am so...                              |
+-----------+----------------------+---------------+-------+----------------------------------------------------------------------------------------------------------+
```

```sql
--phrase seems to be used repetitively "i am so" and related to new york


--search for "the best way to get"
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%the best way to get%'
order by score desc limit 30;
```

```
+------------+------------------+---------------+-------+----------------------------------------------------------------------------------------------------------+
| subreddit  |      author      | distinguished | score |                                                                                                     body |
+------------+------------------+---------------+-------+----------------------------------------------------------------------------------------------------------+
| AskReddit  | Prageo           |               | 1381  | "I am a a great time to time to to to get the best way to get a fire going to be getting the best way to get a fire going |
| AskReddit  | kanezelley        |               | 908   | I am a great dinosaur the best way to get there are a few days ago and I will be in town May be a good idea to have a truck for storage of my tools and whatnot because i feel like I said I was wondering if you are sexy af. |
| AskReddit  | Bob_the_Hamstar  |               | 431   | I found a trail of ants in my kitchen this morning. Anybody know the best way to get rid of them? |
| books      | [deleted]        |               | 330   | "What the best way to get teenagers interested in a book? |
| gifs       | mattheotheninja  |               | 283   | The best way to get then to not do that is to have banked turns so you weren't getting extreme lateral g's whenever they go around a corner. |
| AskReddit  | Cronus           |               | 260   | "I feel like the best way to get around someone being home is opening the door and saying ""Hey! Robby you home?"" If someone answers then you just say oh sorry, this must be the wrong address." |
| news       | Captain_Excellent|               | 235   | I just don't understand how either of the Tsarnaev brothers, or any terrorist, comes to the conclusion that the best way to get your point across is to kill innocent civilians. |
| movies     | exitno           |               | 218   | Honestly, the best way to get a better perspective on the matrix is by watching Animatrix. |
| todayilearned | TheoryS        |               | 192   | I've found the best way to get my cat to come to me is to ensure he sees me when I call his name, and then I move out of his line of sight. He comes running in! |
| aww        | Defenestratio    |               | 192   | "People really underestimate how good cats are at controlling rodent populations. Even with modern technology, cats are the best way to get rid of mice. |
| books      | Guy_Buttersnaps  |               | 183   | "&gt; What the best way to get human beings to want anything? |
| gifs       | Furderhur        |               | 96    | The best way to get women, hitting them. |
| AskReddit  | jesternonchalant |               | 89    | I am so #blessed to call you mine even though you scratch your butt at dinner and a half hour or so and so on the weekends and holidays and I will be dead to the world of beer Reston the best way to get the stuff from the front desk. |
| AskReddit  | iceharvester      |               | 76    | "Sadly, in many cases, it's ""tough love."" They somehow think that the best way to increase self esteem is to insult them and talk bad about them in front of your friends, the best way to get your child's attention is by calling them names, the best way to enforce rules is by screaming and guilt tripping and such. |
| personalFinance | jodythebad   |               | 17    | "To add to the job advice - the best way to get a job is to find out what you can make it to most easily, then start walking around that area, applying to anything, but heavily favoring anything that might pay tips - car wash, busing tables, waiting tables in a perfect world. |
| videos     | rscott09         |               | 55    | "This reminds me of one time when I was growing up we had a big brush pile that built up in our back yard after we did some yard work. Me and my brother decided the best way to get rid of it was to burn it. We figured we should douse the whole thing in gasoline because...well why not? So we poured a couple of gallons on the pile and then drew a line in gas to about 15 feet away and lit it. The fire raced to the pile and damn near blew up, the flames were at least even or above our second story windows. Of course my parents saw and my mom was yelling so my dad raced out to us and we were sure we were about to get our asses beat. When he got out to us he yelled ""DON'T YOU EVER! Is t your mother see you doing something like that again"" and then he stood around with us feeding brush onto the fire for another hour trying to get the flames hot enough to melt glass bottles. " |
| IAmA       | HehPsh           |               | 47    | What is the best way to get on the show? |
| worldnews  | InfamousRLT      |               | 40    | In an ironic twist, the visa they denied was for someone who is demonstrably not an extremist. I guess the best way to get into Sweden is to be an extremist! |
| pics       | pcurve           |               | 40    | "Because FIFA is corrupt organization and don't care about public sharing.  But their sponsors do care.  So going after FIFA Partners is the best way to get FIFA to do anything. |
| explainlikeimfive | idontlol    |               | 40    | Why do mob hits leave the gun on the scene of the crime on TV (e.g., Sopranos...)? Isn't that the best way to get caught? |
+------------+------------------+---------------+-------+----------------------------------------------------------------------------------------------------------+
```

```sql
--just a really common phrase used everywhere


--search for "a good time to time"
```

```sql
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%a good time to time%'
order by score desc limit 30;
```

```
+-----------+----------------------+---------------+-------+----+
|-----------|----------------------|---------------|-------|----|
| subreddit |        author        | distinguished | score |    |
|           |                      |               |       | body
+-----------+----------------------+---------------+-------+----+
|           |                      |               |       |
| AskReddit | jlucasfb             |               |   4   | "Hello my name is a good time to time and money to pay for the shooting range of the applicable liability limits at check-in or ensure that your baggage is fully insured prior to travel to the towers and I d
idn't find Brazil in the list of shipping countries.
| AskReddit | deventio7            |               |   3   | "Sure, sounds good to be a bit late though I could not be able to make it to the company is a good time to time in the general vicinity the same time as the party question of the bonus stats system.
|           |                      |               |       |
| AskReddit | Evil__Jon            |               |   2   | I am a beautiful person. I have a static machine that makes me feel like a good time to time. I speed for fun.
|           |                      |               |       |
| AskReddit | complicatedape       |               |   2   | Hi to everyone who has been a bit of a comedy of errors with our previous sparkie the world of the world is a good time to time and money when you get out of the five hours of meetings and events in the worl
d to me that the idea of videogaming as a televised sport in the world to me that the idea of videogaming as a televised sport
| AskReddit | TheChebert           |               |   2   | I'm not sure if you are not the intended recipient, you can also be used to be a good time to time. I am a beautiful person.
|           |                      |               |       |
| AskReddit | sterling_mallory     |               |   2   | I am a quiet and I will be a good time to time and money to pay for the first time the other day on the right shows that there's a good amount of alcohol that isn't cooking out of that is a nice butt hole.
|           |                      |               |       |
| AskReddit | taispen              |               |   2   | I am too. I have a feeling of being a good time to time.
|           |                      |               |       |
| AskReddit | jackanus             |               |   2   | I am feom south Carolina South Dakota West Virginia Beach Resort and Casino in Las Vegas and I will be a good time to time and money to pay for the first time in the water and the other hand I am a chef.
|           |                      |               |       |
| AskReddit | CallOfCorgithulhu    |               |   2   | I am a very good at it and I will be a good time to time and money to pay for the first time is an occasion to the world u can get a chance to win the game is at the dudes feet so you can lick themselves.
|           |                      |               |       |
| gifs      | Aryion               |               |   2   | """I am not sure if you are not the intended use of the verse search is a good time to time in the morning of the verse search is a good time to time in the morning to make it to the rest of the verse search
 is a good time to time in the morning of the verse search is a good time to time in the morning of the verse..."""
| AskReddit | DirtyUnicorn         |               |   2   | I am a very good at it and I will be a good time to time and money to pay for the first time in the morning and I will be a good time to time and money to pay for the first
|           |                      |               |       |
| AskReddit | migz562              |               |   2   | I am a very good at it and I will be a good time to time and money to pay for the directions on the phone with a few days ago and I will be a good time.
|           |                      |               |       |
| AskReddit | Datashdoe            |               |   2   | I love you too. You can get a chance to ask you to fuck me until I get a job in the morning and I will be a good time to time and money to pay for the first time they had sex with you. Your cock in my ass is
 yours to the University of Mississippi which was October 1962 the baby is born in the morning.
| AskReddit | Oknogo               |               |   2   | I am a beautiful person who is the best of luck to you by the way to get a chance to win the game is a good time to time and money to buy a new one of the year play-off the same time as a result of the year
 play-off the same time as a result of the year.
| AskReddit | booty_flexx          |               |   2   | I love you too baby is born in the morning with screenshots the kids are here to help you with the wedding coming the next few days ago and I was like a good time to time and money to pay for the first time
 ever since I was wondering if you are the most beautiful person to spend my life with you and your family and friends and family members and friends and family members and friends and family members and friend
s and family members and...  Well I guess that's that
| AskReddit | 2Cuil4School         |               |   2   | I am a last-minute tax-filing bastard the party is a good time to time and money to pay for the first time in the near future lol
|           |                      |               |       |
| AskReddit | GlobalThreat777      |               |   2   | I am so glad you are doing well. I didnt go @ the same time. I have a picture? I have no idea how to make sure you have to be a good time to time. I have a picture? I have no idea how to make sure you have
 to be a good time to time.
| AskReddit | DoubleBot            |               |   2   | I am a very good at it and I will be a good time to time and money to pay for the first time in the world of the game is at the end of the day.
|           |                      |               |       |
| AskReddit | balderm              |               |   1   | "I am a very good at it and I will be a good time to time and money to pay for the time you use those extra machines.
|           |                      |               |       |
| Jokes     | NCR0MNCR0916         |               |   1   | Why do you have to follow up with a few days. the only thing that I have a final decision. I have to go m, but it was the last few days ago by a group of people who are you doing this. it is a good time to t
ime. I am a beautiful person.
| AskReddit | Widget_Slap          |               |   1   | I am will of course I will be a good time to time and money to pay for the day of the entire year old man I am a tractor in every way except physically
|           |                      |               |       |
| AskReddit | HellOrHeaven         |               |   1   | The only thing I can do it for the first time in the morning and I will be a good time to time and money to pay for the first time in the morning.
|           |                      |               |       |
| AskReddit | Lemoncholy           |               |   1   | I love you too can be used to be a good time to time and money to pay for the rest of the day.
```

```
--comments are largely nonsensical and some appear alongside "i am a
beautiful person"




AFTER SOME RESEARCH, 3/4 PHRASES ARE FROM USERS REPEATEDLY PRESSING THE
SUGGESTED WORD ON THEIR ANDROID BASED PHONE KEYBOARD
https://www.reddit.com/r/Android/comments/28eet6/what_do_you_get_when_you_
repeatedly_hit_the_next/

Screenshot from link
```

212

What do you get when you repeatedly hit the next suggested word on your phone's keyboard?

Question

> I am a beautiful person who is the best of luck to you by the way to get a chance to win the game is a pre-built it might not say on the board itself is a pre-built it might not say on the board itself is a pre-built it might not say on the board itself...

I wonder how long it takes for others to start looping.

Edit: I'm using SwiftKey, post your keyboard of choice as well.

Edit 2: Front page on a big subreddit, woo!

895 Comments   Give Award   Share   Save   Hide   Report   82% Upvoted

Screenshot from link

Rektangular ⬟ 83 points · 6 years ago

I am not the first time to the first time to the first time

Damn you google keyboard.

Give Award   Share   Report   Save

⊕ xsvpollux 31 points · 6 years ago 8 children

⊕ Pandabol 4 points · 6 years ago 1 child

⊕ m1ss1ontomars2k4 2 points · 6 years ago 0 children

9 more replies

Hoogyme **Razer Phone | Freedom Mobile** ⬟ 76 points · 6 years ago

I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so...

Give Award   Share   Report   Save

⊕ [deleted] 22 points · 6 years ago 3 children

⊕ narangutang 4 points · 6 years ago 0 children

5 more replies

xole 41 points · 6 years ago

I am a beautiful person who is the weather.

SwiftKey

Give Award   Share   Report   Save

⊕ Alexithymia 16 points · 6 years ago 1 child

1 more reply