Warm-up

How many unique subreddits are there?

For this question I used: /sampled_reddit/*

```
df.select("subreddit").distinct().count()
```

```
253336
```

Answer: 253336

Pick a subreddit. What user wrote the most comments in January of 2012? What was the user's top three most-upvoted comments? Filter out bots or other types of automated posts.

For this question I used: /reddit/2012/RC_2012-01.bz2

Here is the code for the first part with results:

Answer: ('Corrupted_Planet', 287)

```
# January 1st 2012 -> 1325376000
#January 31st 2012 -> 1327968000

df.createOrReplaceTempView("TEMP_DF")

sample_pd = spark.sql("""select * from TEMP_DF where temp_df.subreddit = 'runescape'
and temp_df.created_utc > 1325376000
and temp_df.created_utc < 1327968000
and temp_df.author != '[deleted]'""").toPandas()

from collections import Counter
counter = Counter(sample_pd.author)
counter.most_common()[:5] # get the five most common elements
```

```
[('Corrupted_Planet', 287), ('darkhackspal', 284), ('AlmostNPC', 206), ('trimmmy', 137), ('tomblifter', 134)]
```

User with most comments: Corrupted_Planet

Here is the code for the second part with results:

```
sample_pd_2 = spark.sql("""select * from TEMP_DF
where temp_df.subreddit = 'runescape'
and temp_df.created_utc > 1325376000
and temp_df.created_utc < 1327968000
and temp_df.author != '[deleted]'
order by temp_df.score desc""").toPandas()
sample_pd_2.iloc[1:4]
```

flair_css_class	author_flair_text	body	controversiality	created_utc	disti
corruptedplanet	Nova Science	GIVE IT TO ME RIGHT NOW I WANT IT SO BAD	0	1327731513	
corruptedplanet	Nova Science	The new one is too bright, and it doesn't flic...	0	1326838371	
None	None	Funniy enough, that new user button was always...	0	1325785433	

User's top three most-upvoted comments:

1. Corrupted_Planet -> GIVE IT TO ME RIGHT NOW I WANT IT SO BAD
2. Corrupted_Planet -> The new one is too bright, and it doesn't flic...
3. Lawls255 -> Funniy enough, that new user button was always...

Choose a day of significance to you (e.g., your birthday), and retrieve a 5% sample of the comments posted on this particular day across all 5 years of the dataset.

For this question I used: /sampled_reddit/*

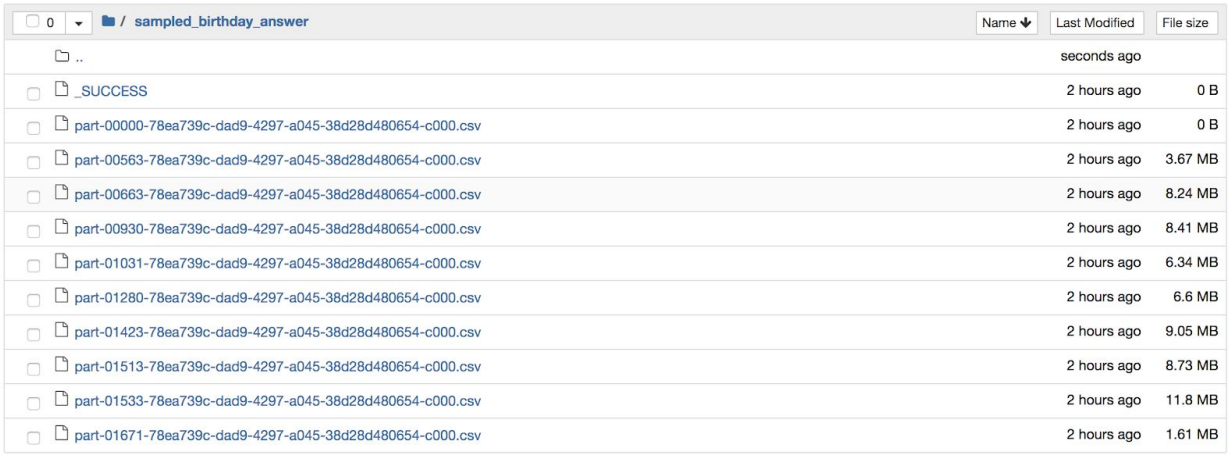
I decided to take the 11th of December.

```
#12.11.20XX from 00:00:00 to 23:59:59

#1512950400 1513036799    2017
#1481414400 1481504399    2016
#1449792000 1449878399    2015
#1418256000 1418342399    2014
#1386720000 1386806399    2013
```

The code is:

```
df.createOrReplaceTempView("TEMP_DF")
pd_3 = spark.sql("""select temp_df.body from TEMP_DF
where (temp_df.created_utc > 1512950400 and temp_df.created_utc < 1513036799)
or (temp_df.created_utc > 1481414400 and temp_df.created_utc < 1481504399)
or (temp_df.created_utc > 1449792000 and temp_df.created_utc < 1449878399)
or (temp_df.created_utc > 1418256000 and temp_df.created_utc < 1418342399)
or (temp_df.created_utc > 1386720000 and temp_df.created_utc < 1386806399)""")

samp = pd_3.sample(False, .5)
samp.write.format('csv').save('hdfs://orion11:32001/sampled_birthday_answer')
```

And here is the sample folder:



And the part of sample screen shot:

	part-00563-78ea739c-d
1	GO HABS GO! Pretty excited. I've watched all but two games for both teams. Good luck, Kings! Just not enough to beat us.
2	Oh yes!!!! Are u going to use that toy?
3	Go Knicks! :)
4	You can sign up and see more about Section II [here](http://signup.sectionii.com/).
5	Like Knocksteady said, use more muted colors. I have a pair of salmon chinos and they're best worn with a white or light-blue OCBD.
6	thank you so much :)
7	Stupid bot... But it's rather cool actually, even though i'm not a *you know* (i didn't say it because then the bot would probably return.)
8	The week? I think you're better off renting a room for cheap and maybe offer to drop them off and pick them up at the stadium. I don't think anyone will be there a week unless they're loaded and if that's th As long as you're not looking to make an ass-load of money off of some guy looking for a bed you should make some nice walking around money. Or you could rent out the entire house but that seems like a hassle and a way to create some problems.
9	It's gonna sound crazy but suggest a 3some to your current girl involving the other girl. She will think it sounds crazy and immediately dismiss it. She will then tell the other girl who is involved and the other
10	Box 3, Eevee #12 matches my SV (179)! Can I have her? FC is 1289-9431-0930 IGN: Roll

The number of comments posted per year will likely trend upward over time as more users join Reddit. However, the popularity of some subreddits may increase or decrease over time. Find An example of both.

For this question I used: /reddit/2016/*

```
In [55]: df.createOrReplaceTempView("TEMP_DF")
pd_4 = spark.sql("""select temp_df.subreddit, MONTH(FROM_UNIXTIME(temp_df.created_utc)) month,
count(temp_df.body) comments
from TEMP_DF
GROUP BY
MONTH(FROM_UNIXTIME(temp_df.created_utc)), temp_df.subreddit""").toPandas()

pd_5 = pd_4.pivot_table(index=['subreddit'],
                        columns='month',
                        values='comments')
pd_5.iloc[:100]
```

Out[55]:

month	1	2	3	4	5	6	7	8	9	10	11	12
subreddit												
000000000000000000	1.0	5.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

So that I will show trends in this year. (Doing through all years will take too much time)

(columns format)

month	1	2	3	4	5	6	7	8	9	10	11	12
subreddit												

<u>Decreasing trends:</u>

Here we can see that subreddits “007” has a decreasing trend, it started with 10 in January and went down to 1 to December.

007	10.0	9.0	6.0	NaN	3.0	1.0	4.0	2.0	2.0	NaN	NaN	1.0

Here we can see that subreddits “0to60” has a decreasing trend, it started with 12 in January, went up to 28 in March but decreased to 7.0 in December.

0to60	12.0	15.0	28.0	2.0	NaN	10.0	NaN	NaN	NaN	NaN	6.0	7.0

<u>Increasing trend:</u>

Here we can see that subreddits “0x10c” has an increasing trend, it started with 2 in February and went up to 45 to December.

0x10c	NaN	2.0	NaN	1.0	2.0	13.0	19.0	2.0	2.0	2.0	NaN	45.0