Xinying Shi 001819931
Wendi Yu 001825852

README

In this assignment, we populate our database with tag data to the database on male tennis topic. This domain have entities that represent people (players, coaches), places(courts) and things(tournaments, match scores, tournaments and match statistics).

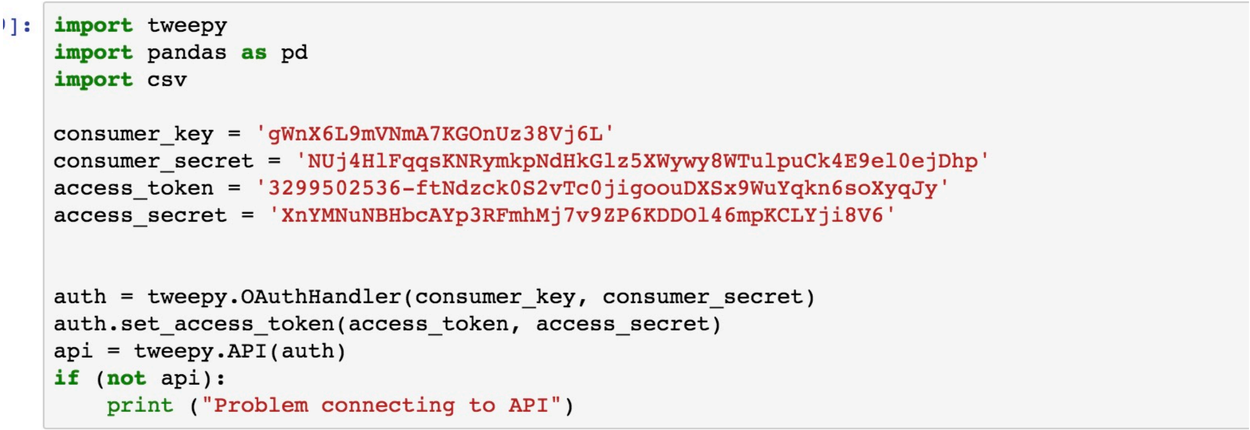
1. Firstly, we scrape some data from http://www.atpworldtour.com and some data from github.

2. Secondly, we cleansed ATP tennis dataset.

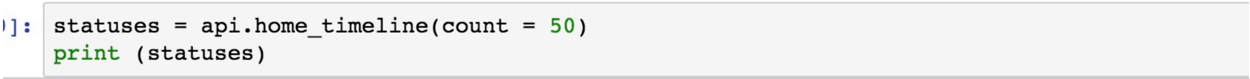
3. Then, we set up database and create tables.

4. The forth step is to scrape tags data from Twitter

   1> Authorizing an application to access Twitter account data

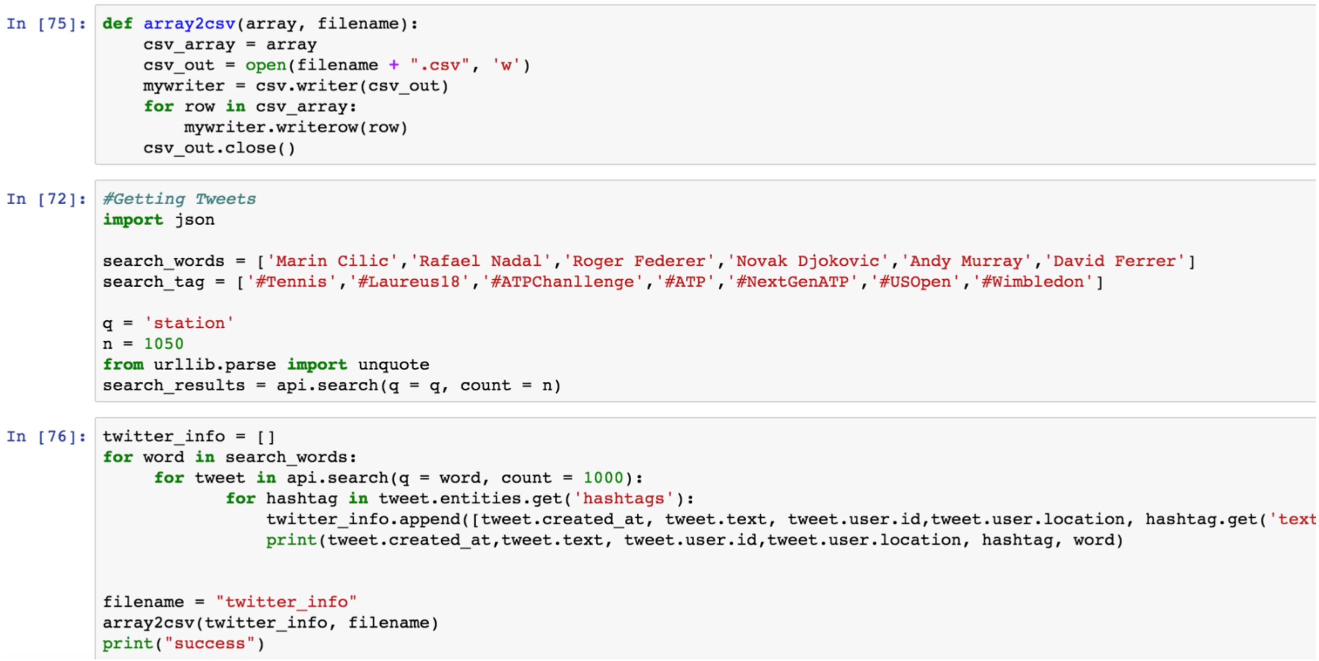


```
import tweepy
import pandas as pd
import csv

consumer_key = 'gWnX6L9mVNmA7KGOnUz38Vj6L'
consumer_secret = 'NUj4HlFqqsKNRymkpNdHkGlz5XWywy8WTulpuCk4E9el0ejDhp'
access_token = '3299502536-ftNdzck0S2vTc0jigoouDXSx9WuYqkn6soXyqJy'
access_secret = 'XnYMNuNBHbcAYp3RFmhMj7v9ZP6KDDOl46mpKCLYji8V6'


auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
api = tweepy.API(auth)
if (not api):
    print ("Problem connecting to API")
```



```
statuses = api.home_timeline(count = 50)
print (statuses)
```

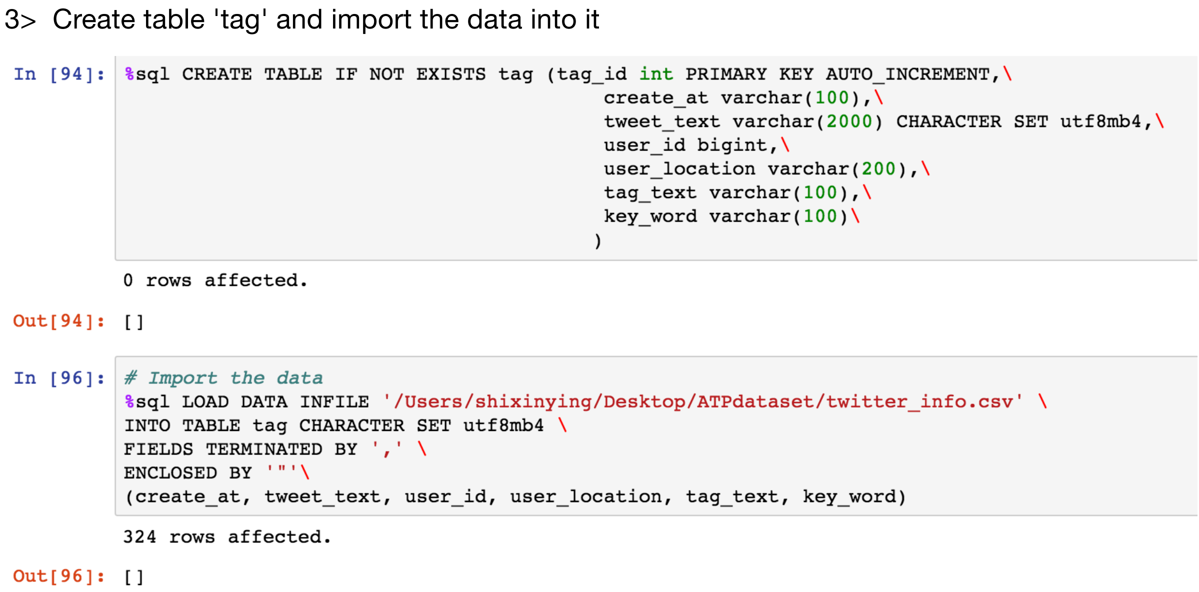
   2> Extracting create time, text, user ids, user location, hashtags and key words from tweets



```
In [75]: def array2csv(array, filename):
             csv_array = array
             csv_out = open(filename + ".csv", 'w')
             mywriter = csv.writer(csv_out)
             for row in csv_array:
                 mywriter.writerow(row)
             csv_out.close()
```

```
In [72]: #Getting Tweets
         import json

         search_words = ['Marin Cilic','Rafael Nadal','Roger Federer','Novak Djokovic','Andy Murray','David Ferrer']
         search_tag = ['#Tennis','#Laureus18','#ATPChanllenge','#ATP','#NextGenATP','#USOpen','#Wimbledon']

         q = 'station'
         n = 1050
         from urllib.parse import unquote
         search_results = api.search(q = q, count = n)
```

```
In [76]: twitter_info = []
         for word in search_words:
              for tweet in api.search(q = word, count = 1000):
                     for hashtag in tweet.entities.get('hashtags'):
                         twitter_info.append([tweet.created_at, tweet.text, tweet.user.id,tweet.user.location, hashtag.get('text
                         print(tweet.created_at,tweet.text, tweet.user.id,tweet.user.location, hashtag, word)


         filename = "twitter_info"
         array2csv(twitter_info, filename)
         print("success")
```

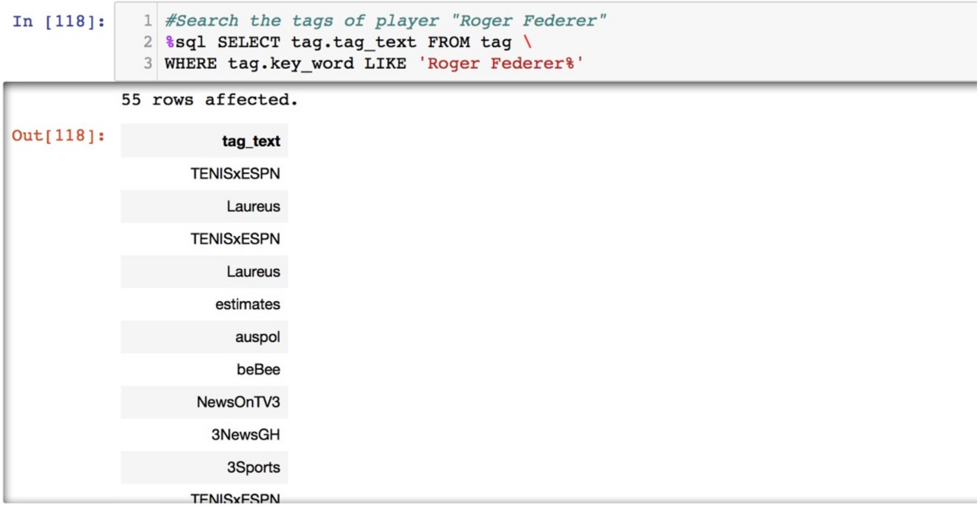


3> Create table 'tag' and import the data into it

```
In [94]: %sql CREATE TABLE IF NOT EXISTS tag (tag_id int PRIMARY KEY AUTO_INCREMENT,\
                                             create_at varchar(100),\
                                             tweet_text varchar(2000) CHARACTER SET utf8mb4,\
                                             user_id bigint,\
                                             user_location varchar(200),\
                                             tag_text varchar(100),\
                                             key_word varchar(100)\
                                            )

         0 rows affected.

Out[94]: []
```

```
In [96]: # Import the data
         %sql LOAD DATA INFILE '/Users/shixinying/Desktop/ATPdataset/twitter_info.csv' \
         INTO TABLE tag CHARACTER SET utf8mb4 \
         FIELDS TERMINATED BY ',' \
         ENCLOSED BY '"'\
         (create_at, tweet_text, user_id, user_location, tag_text, key_word)

         324 rows affected.

Out[96]: []
```

5. What are tags are associated with player:

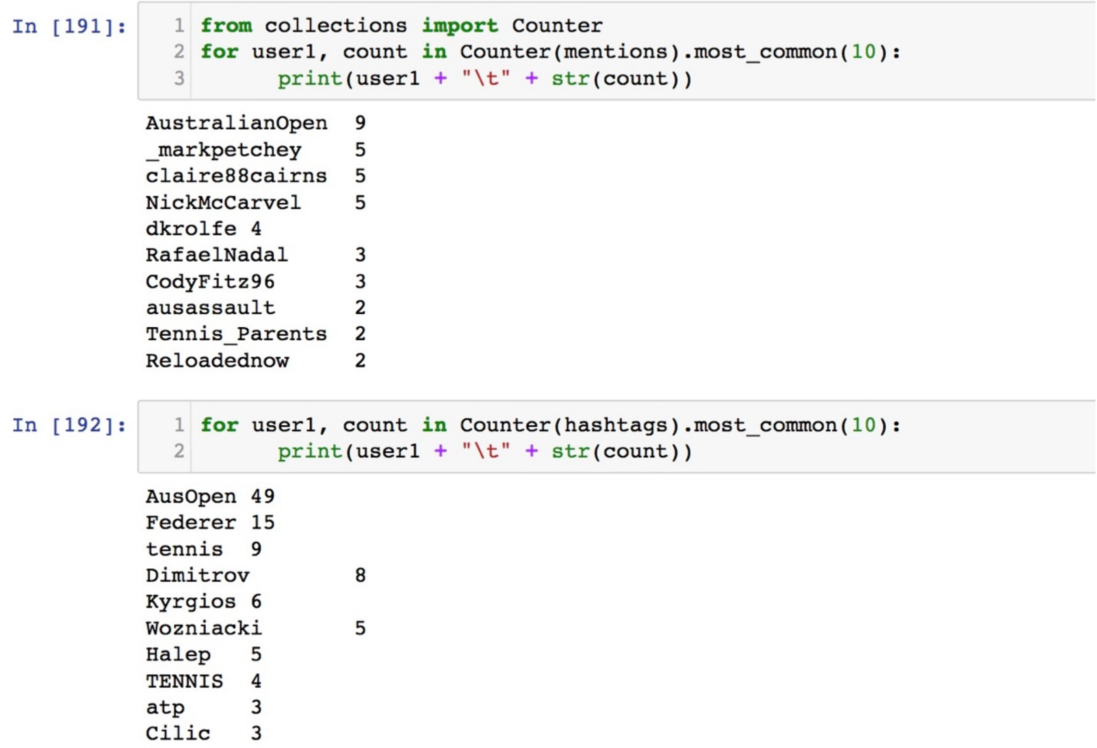


```
In [118]: #Search the tags of player "Roger Federer"
          %sql SELECT tag.tag_text FROM tag \
          WHERE tag.key_word LIKE 'Roger Federer%'

          55 rows affected.
```

Out[118]:

tag_text
TENISxESPN
Laureus
TENISxESPN
Laureus
estimates
auspol
beBee
NewsOnTV3
3NewsGH
3Sports
TENISxESPN

6. What social media users are like other social media users in your domain:
   We search people who come from same country (Australia) and also posted some tweets about tennis in the past, then we compare tags in their tweets in past 100 days, we find the most common tags for two of them are both about Australian Open and Roger Federer, so they are similar user.

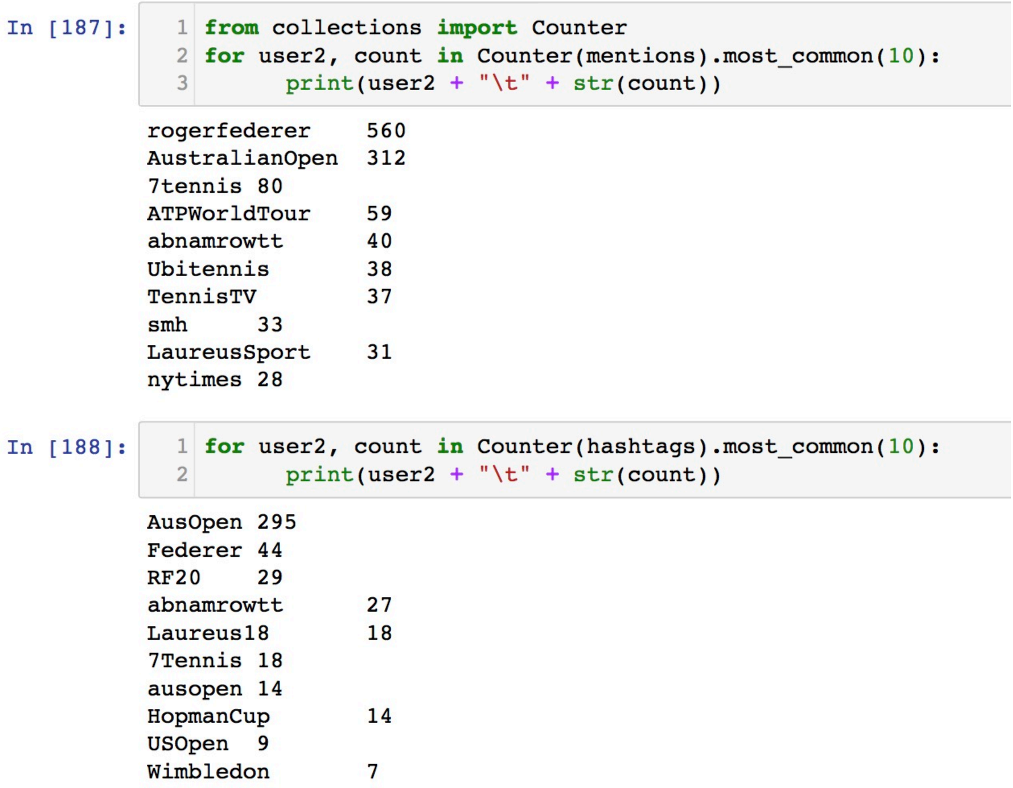


```
In [191]:
from collections import Counter
for user1, count in Counter(mentions).most_common(10):
        print(user1 + "\t" + str(count))
```

```
AustralianOpen	9
_markpetchey	5
claire88cairns	5
NickMcCarvel	5
dkrolfe	4
RafaelNadal	3
CodyFitz96	3
ausassault	2
Tennis_Parents	2
Reloadednow	2
```

```
In [192]:
for user1, count in Counter(hashtags).most_common(10):
        print(user1 + "\t" + str(count))
```

```
AusOpen	49
Federer	15
tennis	9
Dimitrov	8
Kyrgios	6
Wozniacki	5
Halep	5
TENNIS	4
atp	3
Cilic	3
```



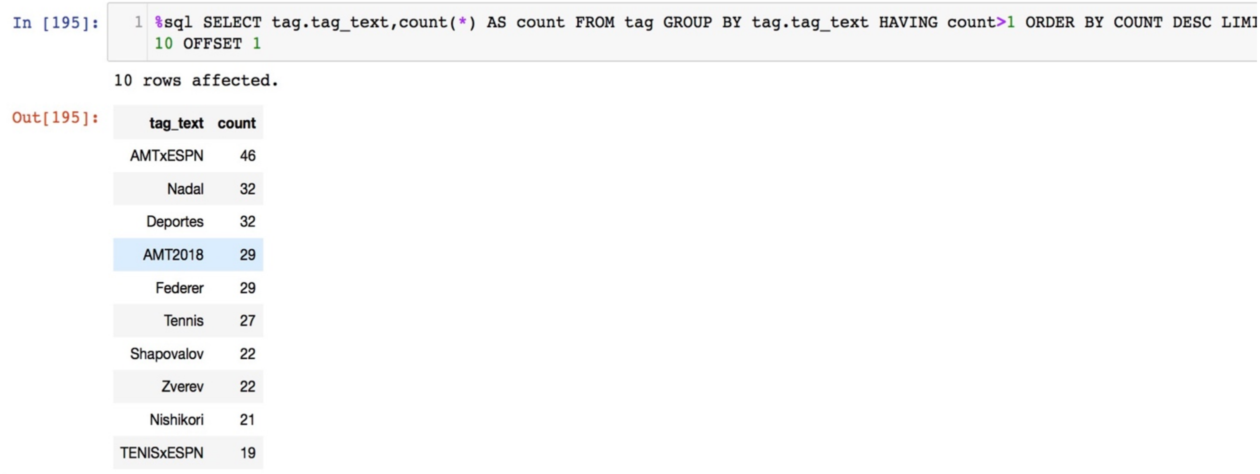
```
In [187]:
from collections import Counter
for user2, count in Counter(mentions).most_common(10):
        print(user2 + "\t" + str(count))
```

```
rogerfederer	560
AustralianOpen	312
7tennis	80
ATPWorldTour	59
abnamrowtt	40
Ubitennis	38
TennisTV	37
smh	33
LaureusSport	31
nytimes	28
```

```
In [188]:
for user2, count in Counter(hashtags).most_common(10):
        print(user2 + "\t" + str(count))
```

```
AusOpen	295
Federer	44
RF20	29
abnamrowtt	27
Laureus18	18
7Tennis	18
ausopen	14
HopmanCup	14
USOpen	9
Wimbledon	7
```

7. What people, places or things are popular in your domain:
   We order the top 10 popular tags when users post tweets about some tennis players.



In [195]:

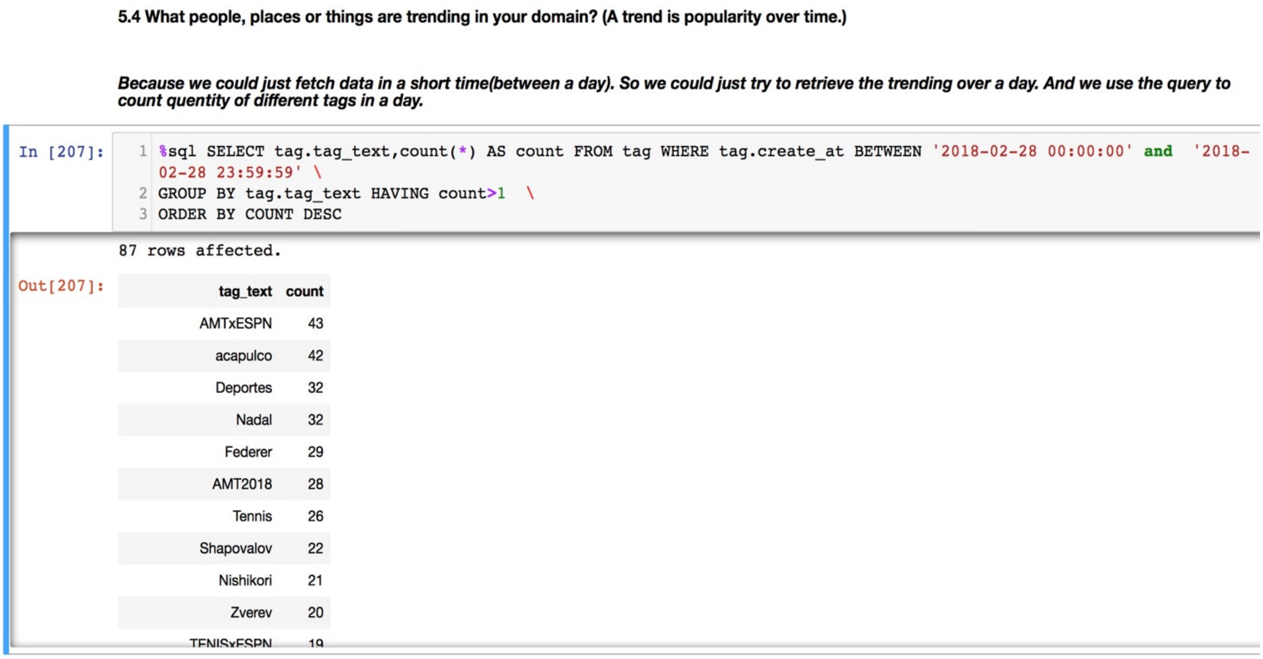
```
%sql SELECT tag.tag_text,count(*) AS count FROM tag GROUP BY tag.tag_text HAVING count>1 ORDER BY COUNT DESC LIMI
10 OFFSET 1
```

```
10 rows affected.
```

Out[195]:

tag_text	count
AMTxESPN	46
Nadal	32
Deportes	32
AMT2018	29
Federer	29
Tennis	27
Shapovalov	22
Zverev	22
Nishikori	21
TENISxESPN	19

8. What topics about tennis players are trending in your domain? (A trend is popularity over a day.)



5.4 What people, places or things are trending in your domain? (A trend is popularity over time.)

Because we could just fetch data in a short time(between a day). So we could just try to retrieve the trending over a day. And we use the query to count quentity of different tags in a day.

In [207]:

```
%sql SELECT tag.tag_text,count(*) AS count FROM tag WHERE tag.create_at BETWEEN '2018-02-28 00:00:00' and  '2018-
02-28 23:59:59' \
GROUP BY tag.tag_text HAVING count>1  \
ORDER BY COUNT DESC
```

```
87 rows affected.
```

Out[207]:

tag_text	count
AMTxESPN	43
acapulco	42
Deportes	32
Nadal	32
Federer	29
AMT2018	28
Tennis	26
Shapovalov	22
Nishikori	21
Zverev	20
TENISxESPN	19