Baseball Database

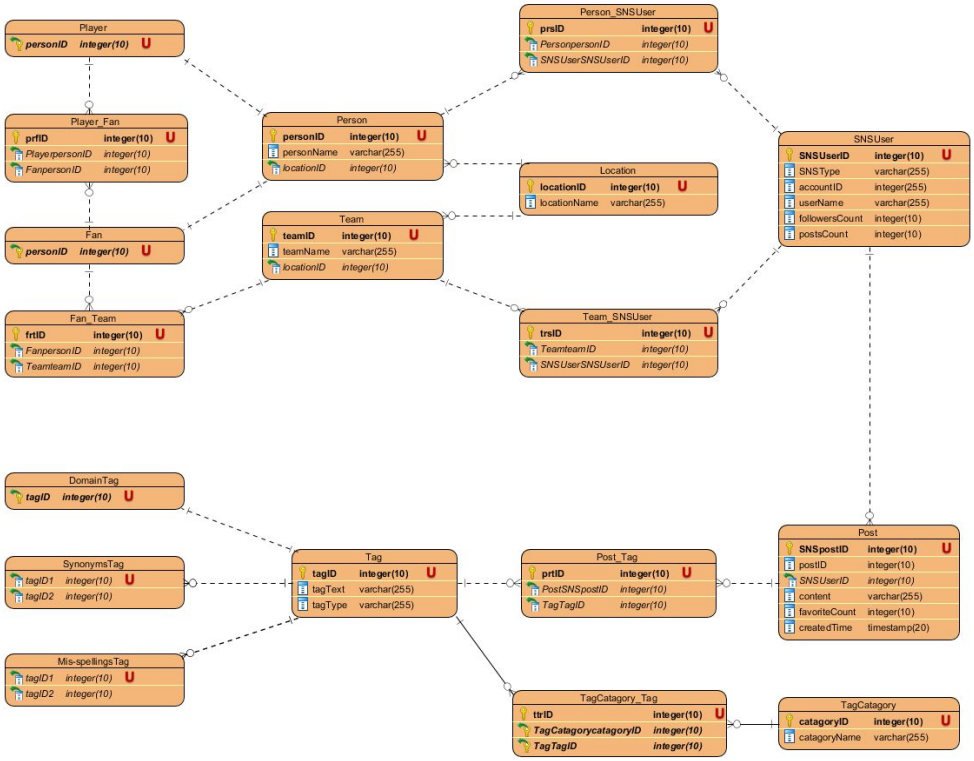
Jing Wang/Siwei Zeng

Background

In this project, we are assumed to be working for a company called Nerd Analytics. Our job is creating a social news site. We choose to build a database about baseball. Each domain must have entities that represent people, places and things. In our project, baseball player and fan represent people, team represents thing and location represents place. The whole project includes building conceptual model, acquiring data from Twitter and Instagram, Tagging and writing use case according to our database.

Procedure Diagram

Entity Relationship Diagram



Data Acquisition and Tagging

We get the data from Instagram and Twitter by Instaloader and Tweepy. For the convenience, we define many functions first to use the API to get the data. We use hashtag and mention to do the data tagging. Then we convert these tables to csv files.

1) Instagram:

import libraries and setting instaloader

```
In [1]: import instaloader
        import datetime
        import warnings
        import time
        warnings.filterwarnings("ignore")
        import time
        from instaloader import Instaloader,Profile
        L=Instaloader()
```

define API functions

```
]: def anyname2user(anyname):
       if type(anyname)==int:
           profile=Profile.from_id(L.context,anyname)
       else:
           profile=Profile.from_username(L.context,anyname)
       return profile
```

Make DataFrame

```
import pandas as pd
Player = pd.DataFrame(columns=["personID"])
Player_Fan = pd.DataFrame(columns=["prfID", "PlayerpersonID", "FanpersonID"])
Fan = pd.DataFrame(columns=["personID"])
Fan_Team = pd.DataFrame(columns=["frtID", "FanpersonID", "TeamteamID"])
Person = pd.DataFrame(columns=["personID", "personName", "locationID"])
Team = pd.DataFrame(columns=["teamID", "teamName", "locationID"])
Person_SNSUser = pd.DataFrame(columns=["prsID", "PersonpersonID", "SNSUserSNSUserID"])
Location = pd.DataFrame(columns=["locationID", "locationName"])
Team_SNSUser = pd.DataFrame(columns=["trsID", "TeamteamID", "SNSUserSNSUserID"])
SNSUser = pd.DataFrame(columns=["SNSUserID", "SNSType", "accountID", "userName", "followersCount", "postsC
Post = pd.DataFrame(columns=["SNSpostID", "postID", "SNSUserID", "content", "favoriteCount", "createdTime"
Post_Tag = pd.DataFrame(columns=["prtID", "PostSNSpostID", "TagTagID"])
Tag = pd.DataFrame(columns=["tagID", "tagText", "tagType"])
DomainTag = pd.DataFrame(columns=["tagID"])
SynonymsTag = pd.DataFrame(columns=["tagID1", "tagID2"])
MisspellingsTag = pd.DataFrame(columns=["tagID1", "tagID2"])
TagCatagory_Tag = pd.DataFrame(columns=["ttrID", "TagCatagorycatagoryID", "TagTagID"])
TagCatagory = pd.DataFrame(columns=["catagoryID", "catagoryName"])
```

Data Retrieving Function

```
def savesnsuser(user):
    if SNSUser[SNSUser["accountID"].isin([user2accountid(user)])].shape[0] == 0:
        SNSUser.loc[SNSUser.shape[0] + 1] = {"SNSUserID": SNSUser.shape[0] + 1,
                                             "SNSType": "Twitter",
                                             "accountID": user2accountid(user),
                                             "userName": user2username(user),
                                             "followersCount": user2followerscount(user),
                                             "postsCount":user2postscount(user)}
    else:
        SNSUser.set_value(SNSUser[SNSUser["accountID"].isin([user2accountid(user)])].loc[:, ["SNSUserID"]
        SNSUser.set_value(SNSUser[SNSUser["accountID"].isin([user2accountid(user)])].loc[:, ["SNSUserID"]
```

domain tag

```
teamSNS=[["New York Yankees", "yankees"],
         ["Los Angeles Dodgers", "dodgers"],
         ["Tampa Bay Rays", "raysbaseball"]]
playerSNS=[["Greg Bird", "_gregbird33"],
           ["DJ LeMahieu", "dj_lemahieu"],
           ["Zack Britton", "zbritton"],
           ["Max Muncy", "maxmuncy"],
           ["Julio Urías", "theteenager7"],
           ["Cody Bellinger", "cody_bellinger"],
           ["Christian Arroyo", "arroyoc22"],
           ["Austin Meadows", "austinmeadows13"],
           ["Guillermo Heredia", "heredia54"]]
```

```
for tag in TagList:
    if tag[2]=="hashtag":
        for player in playerSNS:
            if tag[1] in player[1]:
                DomainTagList.append([tag[0]])
        for team in teamSNS:
            if tag[1] in team[1]:
                DomainTagList.append([tag[0]])
```

synonyms tag

```
def tagcount1(tagid):
    counter = 0
    for pt in Post_TagList:
        if pt[2]==tagid:
            counter=counter+1
    return counter
tagcount1(5)
```

Misspelling Tag

```
# Levenshtein distance calculation function
# Reference: https://rosettacode.org/wiki/Levenshtein_distance#Python
def ldistance(s1, s2):
    if len(s1) > len(s2):
        s1, s2 = s2, s1
    distances = range(len(s1) + 1)
    for index2, char2 in enumerate(s2):
        newDistances = [index2 + 1]
        for index1,char1 in enumerate(s1):
            if char1 == char2:
                newDistances.append(distances[index1])
            else:
                newDistances.append(1 + min((distances[index1],
                                             distances[index1 + 1],
                                             newDistances[-1])))
        distances = newDistances
    return distances[-1]
```

Tagging dataframe

```
def list2tuplelist(list):
    tuplelist=[]
    for row in list:
        tuplelist.append(tuple(row))
    return tuplelist
```

2) Twitter:

Import Libraries and Setting API

```
import tweepy
import json
import datetime
import warnings
import time
warnings.filterwarnings('ignore')
consumer_key = 'nRO8b2R7tnipKtrFTJVRUru7V'
consumer_secret = 'bK2v3PVVZGktAT2BbviC96Tz9vzjq0ZDj2WLKZ1pmm5aGgu4eW'
access_token = '1098973181468205057-xXfZG8xbtYbXjE6T2s0Ok5J0bKRpsN'
access_token_secret = 'W7mu0M4ABrbSH59nHLH43YMeR13GZCfN1v8FOh1p8tPHh'
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit = True)
```

Make DataFrame

```
import pandas as pd
Player = pd.DataFrame(columns = ["personID"])
Player_Fan = pd.DataFrame(columns = ["prfID", "PlayerpersonID", "FanpersonID"])
Fan = pd.DataFrame(columns = ["personID"])
Fan_Team = pd.DataFrame(columns = ["frtID", "FanpersonID", "TeamteamID"])
Person = pd.DataFrame(columns = ["personID", "personName", "locationID"])
Team = pd.DataFrame(columns = ["teamID", "teamName", "locationID"])
Person_SNSUser = pd.DataFrame(columns = ["prsID", "PersonpersonID", "SNSUserSNSUserID"])
Location = pd.DataFrame(columns = ["locationID", "locationName"])
Team_SNSUser = pd.DataFrame(columns = ["trsID", "TeamteamID", "SNSUserSNSUserID"])
SNSUser = pd.DataFrame(columns = ["SNSUserID", "SNSType", "accountID", "userName", "followersCount", "postsCo
Post = pd.DataFrame(columns = ["SNSpostID", "postID", "SNSUserID", "content", "favoriteCount", "createdTime"]
Post_Tag = pd.DataFrame(columns = ["prtID", "PostSNSpostID", "TagTagID"])
Tag = pd.DataFrame(columns = ["tagID", "tagText", "tagType"])
DomainTag = pd.DataFrame(columns = ["tagID"])
SynonymsTag = pd.DataFrame(columns = ["tagID1", "tagID2"])
MisspellingsTag = pd.DataFrame(columns = ["tagID1", "tagID2"])
TagCatagory_Tag = pd.DataFrame(columns = ["trtID", "TagCatagorycatagoryID", "TagTagID"])
TagCatagory = pd.DataFrame(columns = ["catagoryID", "catagoryName"])
```

DomainTag

```
for tag in TagList:
    if tag[2]=="hashtag":
        for player in playerSNS:
            if tag[1] in player[1]:
                DomainTagList.append([tag[0]])
        for team in teamSNS:
            if tag[1] in team[1]:
                DomainTagList.append([tag[0]])
DomainTagList
```

SynonymsTag

```
ratioleast = 0.75
countleast = 10
for tag1 in TagList:
    if tag1[2] == "hashtag" and tagcount1(tag1[0])>countleast:
        for tag2 in TagList:
            if tag2[2] == "hashtag" and tagcount1(tag2[0]>countleast):
                if tag1 != tag2:
                    tag1id = tag1[0]
                    tag2id = tag2[0]
                    f1 = tagcount1(tag1id)
                    f2 = tagcount1(tag2id)
                    f3 = tagcount2(tag1id, tag2id)
                    if f3/f1 > ratioleast or f3/f2 >ratioleast:
                        if tag1id > tag2id:
                            temptagid = tag1id
                            tag1id = tag2id
                            tag2id = temptagid
                        if [tag1id, tag2id] not in SynonymsTagList:
                            SynonymsTagList.append([tag1id, tag2id])
```

Levenshtein distance calculation function

```
# Levenshtein distance calculation function
def ldistance(s1, s2):
    if len(s1) > len(s2):
        s1, s2 = s2, s1
    distances = range(len(s1) + 1)
    for index2, char2 in enumerate(s2):
        newDistances = [index2 + 1]
        for index1,char1 in enumerate(s1):
            if char1 == char2:
                newDistances.append(distances[index1])
            else:
                newDistances.append(1 + min((distances[index1],
                                             distances[index1 + 1],
                                             newDistances[-1])))
        distances = newDistances
    return distances[-1]
```

Tagging dataframe

```
def list2tuplelist(list):
    tuplelist=[]
    for row in list:
        tuplelist.append(tuple(row))
    return tuplelist
```

```
labels=["tagID"]
DomainTag=pd.DataFrame.from_records(list2tuplelist(DomainTagList),columns=labels)
```

3) Sample data:

```
Player.head()
```

	personID
1	1
2	2
3	3
4	4
5	5

```
Player_Fan.head()
```

	prfID	PlayerpersonID	FanpersonID
1	1	3	684
2	2	5	758
3	3	5	759
4	4	5	760
5	5	5	761

```
Fan.head()
```

	personID
1	10
2	11
3	12
4	13
5	14

```
Fan_Team.head()
```

	frtID	FanpersonID	TeamteamID
1	1	10	1
2	2	11	1
3	3	12	1
4	4	13	1
5	5	14	1

```
Person.head()
```

	personID	personName	locationID
1	1	Alexander David Bregman	4
2	2	Carlos Javier Correa	5
3	3	José Carlos Altuve	5
4	4	Nicholas A. Castellanos	6
5	5	Jose Miguel Cabrera	7

```
Team.head()
```

	teamID	teamName	locationID
1	1	Houston Astros	1
2	2	Detroit Tigers	2
3	3	Texas Rangers	3

```
Person_SNSUser.head()
```

	prsID	PersonpersonID	SNSUserSNSUserID
1	1	1	4
2	2	2	5
3	3	3	6
4	4	4	7
5	5	5	8

```
1 Location.head()
```

	locationID	locationName
1	1	Minute Maid Park
2	2	Comerica Park
3	3	Deep in the Heart of Texas
4	4	Albuquerque, NM
5	5	

```
1 Team_SNSUser.head()
```

	trsID	TeamteamID	SNSUserSNSUserID
1	1	1	1
2	2	2	2
3	3	3	3

```
1 SNSUser.head()
```

	SNSUserID	SNSType	accountID	userName	followersCount	postsCount
1	1	Twitter	52803520	astros	1455653	72000
2	2	Twitter	30008146	tigers	1453071	49251
3	3	Twitter	40931019	Rangers	1426982	43627
4	4	Twitter	382708289	ABREG_1	242213	7895
5	5	Twitter	2811330143	TeamCJCorrea	221551	740

```
1 Post.head()
```

	SNSpostID	postID	SNSUserID	content	favoriteCount	createdTime
1	1	1121607683507810304	1	Final: Indians 2, #Astros 1	111	2019-04-26 02:51:24
2	2	1121598945145147392	1	We know you'd like this in your living room. \...	1068	2019-04-26 02:16:41
3	3	1121582532955230208	1	Hang it, then bang it. 💥\n\nWe're all square! ...	2829	2019-04-26 01:11:28
4	4	1121579802551435264	1	George doesn't look too frustrated anymore.\n\...	1453	2019-04-26 01:00:37
5	5	1121575278755471360	1	Special message from the @Indians to Uncle Mik...	623	2019-04-26 00:42:38

```
Post_Tag.head()
```

	prtID	PostSNSpostID	TagTagID
1	1	1	1
2	2	2	2
3	3	3	2
4	4	4	2
5	5	5	2

```
Tag.head()
```

	tagID	tagText	tagType
1	1	Astros	hashtag
2	2	TakeItBack	hashtag
3	3	Indians	mention
4	4	ABREG_1	mention
5	5	ATTSportsNetSW	mention

```
DomainTag.head()
```

	tagID
0	2553
1	2571
2	2571
3	2613
4	2656

```
SynonymsTag.head()
```

	tagID1	tagID2
0	1	55
1	1	102
2	1	2649
3	1	2650
4	1	2651

```
MisspellingsTag.head()
```

	tagID1	tagID2
0	37	1
1	2613	1
2	2689	1
3	16	2
4	2626	2

```
1 TagCatagory_Tag.head()
```

	trtID	TagCatagorycatagoryID	TagTagID
0	1	1	1
1	2	1	2
2	3	2	3
3	4	2	4
4	5	2	5

```
1 TagCatagory.head()
```

	catagoryID	catagoryName
0	1	hashtag
1	2	mention
2	3	domaintag
3	4	synonymstag
4	5	misspellingstag

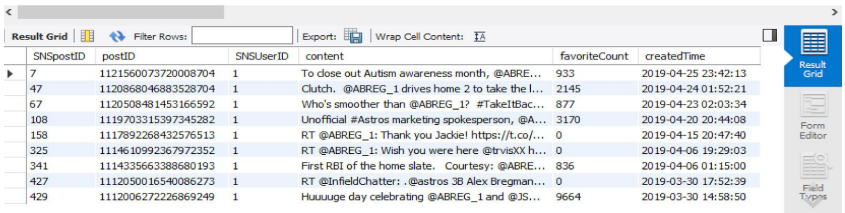
Check the Table

1) 1NF check: Each table has a primary key; The values in each column of a table are atomic; There are no repeating groups.
2) 2 NF check: No partial dependencies; No calculated data.
3) 3NF check: Eliminate fields that do not directly depend on the primary key, that is no transitive dependencies.

Newsfeed (popular hashtags)

1) What are people saying about me (somebody)?

```
1 SELECT * FROM post WHERE SNSpostID IN (
2 SELECT PostSNSpostID FROM post_tag WHERE TagTagID IN (
3 SELECT tagID FROM tag WHERE tagText = "ABREG_1" AND tagType = "mention"))
```



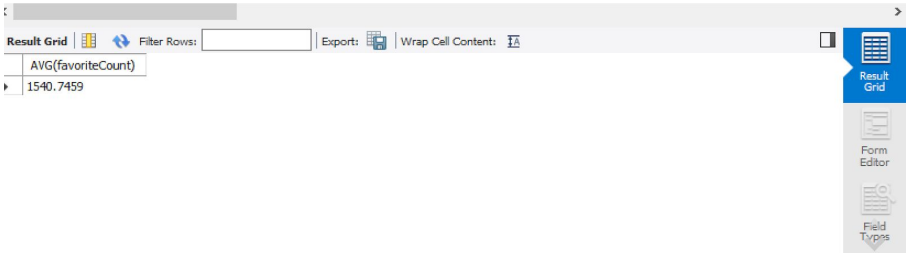
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

SNSpostID	postID	SNSUserID	content	favoriteCount	createdTime
7	1121560073720008704	1	To close out Autism awareness month, @ABRE...	933	2019-04-25 23:42:13
47	1120868046883528704	1	Clutch. @ABREG_1 drives home 2 to take the l...	2145	2019-04-24 01:52:21
67	1120508481453166592	1	Who's smoother than @ABREG_1? #TakeItBac...	877	2019-04-23 02:03:34
108	1119703315397345282	1	Unofficial #Astros marketing spokesperson, @A...	3170	2019-04-20 20:44:08
158	1117892268432576513	1	RT @ABREG_1: Thank you Jackie! https://t.co/...	0	2019-04-15 20:47:40
325	1114610992367972352	1	RT @ABREG_1: Wish you were here @trvisXX h...	0	2019-04-06 19:29:03
341	1114335663388680193	1	First RBI of the home slate. Courtesy: @ABRE...	836	2019-04-06 01:15:00
427	1112050016540086273	1	RT @InfieldChatter: .@astros 3B Alex Bregman...	0	2019-03-30 17:52:39
429	1112006272226869249	1	Huuuuge day celebrating @ABREG_1 and @JS...	9664	2019-03-30 14:58:50

Result Grid | Form Editor | Field Types

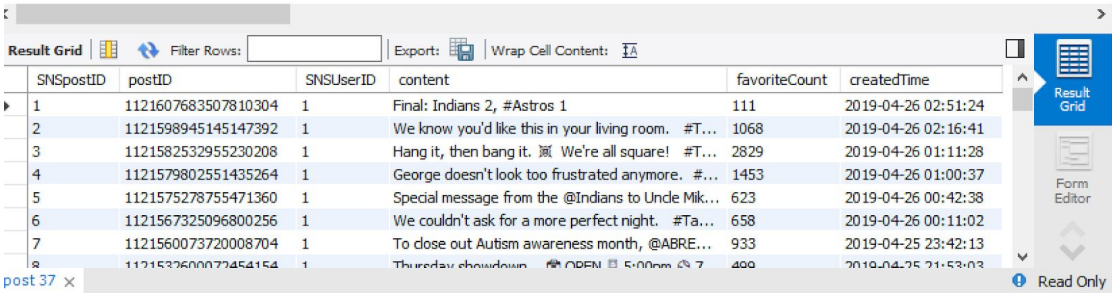
2) How viral are my posts?

```
SELECT AVG(favoriteCount) FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros")
```



3) What posts are likely to be interesting to me?

```
SELECT * FROM post WHERE SNSpostID IN(
SELECT DISTINCT PostSNSpostID FROM post_tag WHERE TagTagID IN(
SELECT DISTINCT TagTagID FROM post_tag WHERE PostSNSpostID IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros"))))
```



4) What posts are like mine?

```
SELECT * FROM post WHERE SNSpostID IN(
SELECT DISTINCT PostSNSpostID FROM post_tag WHERE TagTagID IN(
SELECT DISTINCT TagTagID FROM post_tag WHERE PostSNSpostID IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros")))
AND PostSNSpostID NOT IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros")))
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

SNSpostID	postID	SNSUserID	content	favoriteCount	createdTime
668	1120879397039411200	2	@ltwhale @mams33 @DetroitRedWings @Lions ...	9	2019-04-24 02:37:27
763	1120833350200119297	2	@lwoodjr_ @MLB This is an elite tweet	1	2019-04-23 23:34:29
823	1120684273424580615	2	@espn @ us next time https://t.co/BuewTHwMH2	2767	2019-04-23 13:42:06
899	1120048092013891585	2	@Indians You're punnier than this, AtIndians.	26	2019-04-21 19:34:09
1110	1118984141889052672	2	@MLB https://t.co/lQx7tWilVL	380	2019-04-18 21:06:23
1113	1118983138120732676	2	RT @MLB: It's Miggy's birthday, so let's take a l...	0	2019-04-18 21:02:24
1204	1118900597548355585	2	@Indians Luv u, mean it.	37	2019-04-18 15:34:25

Result Grid | Form Editor

5) What users post like me?

```
SELECT DISTINCT SNSUserID FROM post WHERE SNSpostID IN(
SELECT DISTINCT PostSNSpostID FROM post_tag WHERE TagTagID IN(
SELECT DISTINCT TagTagID FROM post_tag WHERE PostSNSpostID IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros")))
AND PostSNSpostID NOT IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros")))
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

SNSUserID
2

Result Grid | Form Editor

6) Who should I be following?

```
SELECT * FROM snsuser WHERE SNSUserID IN (
SELECT DISTINCT SNSUserID FROM post WHERE SNSpostID IN(
SELECT DISTINCT PostSNSpostID FROM post_tag WHERE TagTagID IN(
SELECT DISTINCT TagTagID FROM post_tag WHERE PostSNSpostID IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros")))
AND PostSNSpostID NOT IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros"))))
```

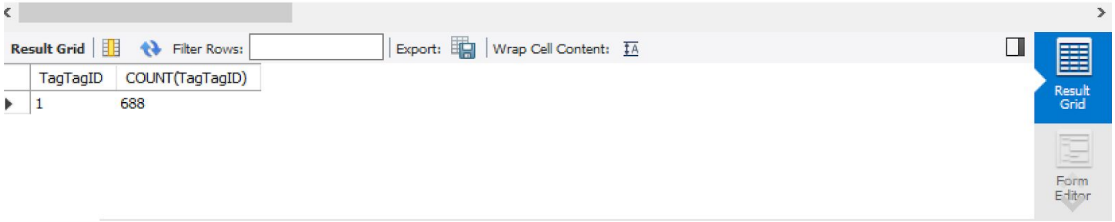
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

SNSUserID	SNSType	accountID	userName	followersCount	postsCount
2	Twitter	30008146	tigers	1453071	49251

Result Grid | Form Editor

7) What topics are trending in my domain?

```
SELECT DISTINCT TagTagID, COUNT(TagTagID) FROM post_tag WHERE PostSNSpostID IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros"))
ORDER BY COUNT(TagTagID)
LIMIT 10
```



8) What keywords/ hashtags should I add to my post?

```
SELECT DISTINCT TagTagID, ztag.tagText, SUM(zpost.favoriteCount) AS favoriteCount
FROM zpost_tag
INNER JOIN zpost
ON zpost_tag.PostSNSpostID = zpost.SNSpostID
INNER JOIN ztag
ON zpost_tag.TagTagID = ztag.tagID
GROUP BY zpost_tag.TagTagID
ORDER BY favoriteCount DESC
LIMIT 5
```

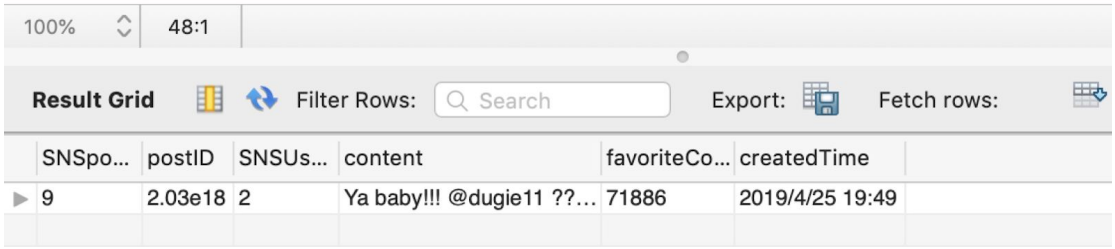
9) Should I follow somebody back?

```
SELECT * FROM zsnsuser
WHERE accountID IN (SELECT DISTINCT SNSUserID FROM zpost
WHERE SNSpostID IN (SELECT DISTINCT PostSNSpostID FROM zpost_tag
WHERE TagTagID IN (SELECT DISTINCT TagTagID FROM zpost_tag
WHERE PostSNSpostID IN (SELECT SNSpostID FROM zpost
WHERE SNSUserID IN (SELECT SNSUserID FROM zsnsuser
WHERE userName = """ + "'" + me + "'" + """)))))
```

10 Use Cases

1.Which ins post is the most popular?

```
1 •  select * from wpost order by favoriteCount DESC limit 1
```



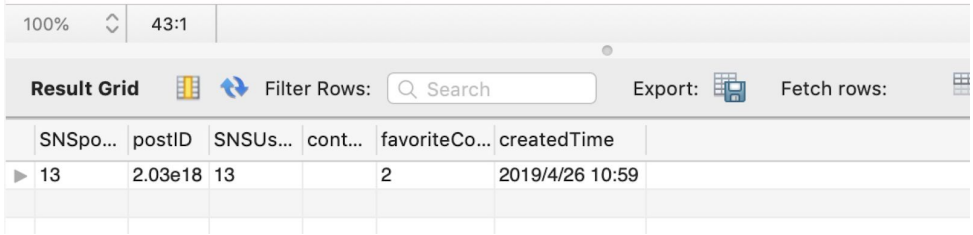
100% 48:1

Result Grid Filter Rows: Search Export: Fetch rows:

SNSpo...	postID	SNSUs...	content	favoriteCo...	createdTime
9	2.03e18	2	Ya baby!!! @dugie11 ??...	71886	2019/4/25 19:49

2.Which ins post is the least popular?

```
1 •  select * from wpost order by favoriteCount limit 1
```



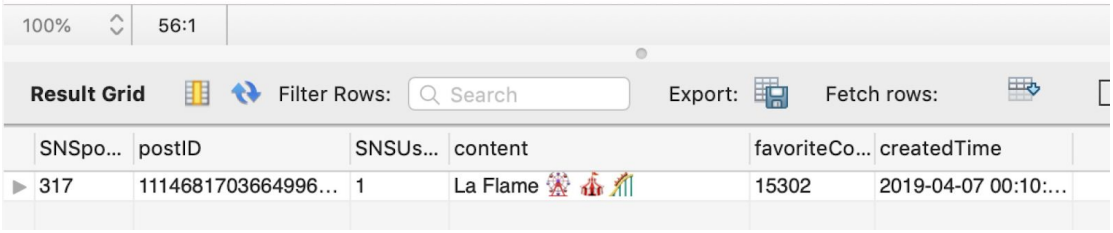
100% 43:1

Result Grid Filter Rows: Search Export: Fetch rows:

SNSpo...	postID	SNSUs...	cont...	favoriteCo...	createdTime
13	2.03e18	13		2	2019/4/26 10:59

3.Which twitter post is the most popular?

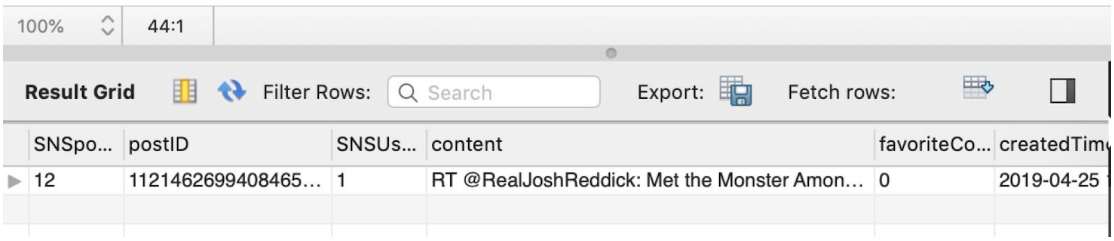
```
1 •  select * from zpost order by favoriteCount DESC limit 1
```



SNSpo...	postID	SNSUs...	content	favoriteCo...	createdTime
317	1114681703664996...	1	La Flame	15302	2019-04-07 00:10:...

4.Which twitter post is the least popular?

```
1 •  select * from zpost order by favoriteCount limit 1
```



SNSpo...	postID	SNSUs...	content	favoriteCo...	createdTim
12	1121462699408465...	1	RT @RealJoshReddick: Met the Monster Amon...	0	2019-04-25

5. How many favorites did the latest twitter post get?

```
1 •  select postID, content, favoriteCount from zpost
2    order by createdTime limit 1
```

100% 1:2

Result Grid Filter Rows: Search Export: Fetch rows:

postID	content	favoriteCo...
1110874191790829...	Tomorrow.	1511

6. How many favorites did the latest ins post get?

```
1 •  select postID, content, favoriteCount from wpost
2    order by createdTime limit 1
```

100% 29:2

Result Grid Filter Rows: Search Export: Fetch

postID	cont...	favoriteCo...
1.63e18		12

7. Rank the ins posts according to the favorite count.

```
1 •  select * from wpost order by favoriteCount
```

100% 16:1

Result Grid Filter Rows: Search Export:

SNSpo...	postID	SNSUs...	content	favoriteCo...	createdTin
16	1.63e18	16		12	2017/10/18
15	2.03e18	15		48	2019/4/26 6
11	2.03e18	3	Holey biceps! ??	4975	2019/4/25 1
2	2.03e18	2	Amazing pic????	12088	2019/4/25 2
1	2.03e18	2	Can we a do a friends night ???	12119	2019/4/26 0
4	2.03e18	2	Bullpen was solid... they deserve credit, they we…	23631	2019/4/25 2
10	2.03e18	2	Listening at work @dodgers ??	24277	2019/4/25 1

8. Rank the twitter posts according to the favorite count.

```
1 •  select * from zpost order by favoriteCount
```

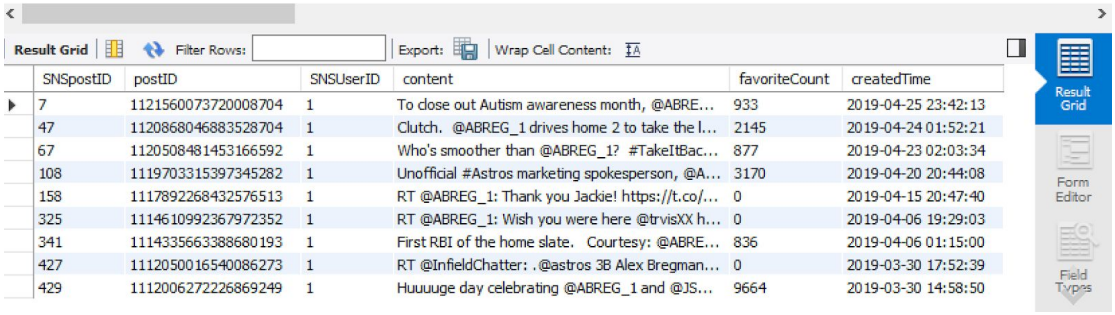
100% 43:1

Result Grid Filter Rows: Search Export: Fetch rows:

SNSpo...	postID	SNSUs...	content	favoriteC
12	1121462699408465…	1	RT @RealJoshReddick: Met the Monster Amon…	0
29	1121143886779494…	1	RT @ATTSportsNetSW: Tonight, @astros vs Tw…	0
32	1121123419448332…	1	RT @BraunStrowman: Pre first pitch pump eng…	0
33	1121105565911060…	1	RT @MLB: These infielders are crushing.	0
40	1120890826945511…	1	RT @AstrosTrainGuy: You know who can hit a b…	0
59	1120732323874471…	1	RT @HouOpenGolf: Updates to the 19th Hole!	0
98	1119970808389148…	1	RT @RealJoshReddick: Happy Easter everyon…	0

9. What posts did people mention about ABREG_1?

```
SELECT * FROM post WHERE SNSpostID IN (
SELECT PostSNSpostID FROM post_tag WHERE TagTagID IN (
SELECT tagID FROM tag WHERE tagText = "ABREG_1" AND tagType = "mention"))
```



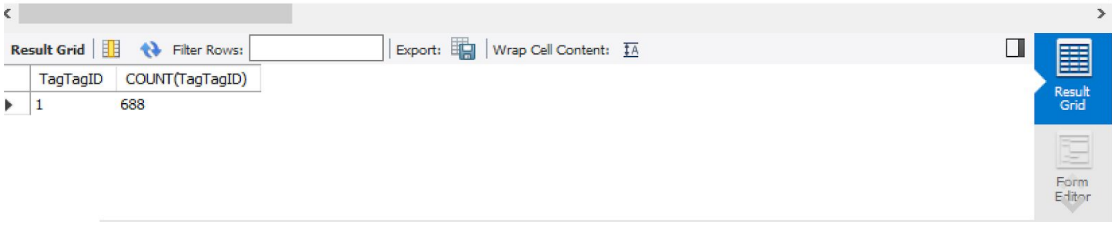
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

SNSpostID	postID	SNSUserID	content	favoriteCount	createdTime
7	1121560073720008704	1	To close out Autism awareness month, @ABRE...	933	2019-04-25 23:42:13
47	1120868046883528704	1	Clutch. @ABREG_1 drives home 2 to take the l...	2145	2019-04-24 01:52:21
67	1120508481453166592	1	Who's smoother than @ABREG_1? #TakeItBac...	877	2019-04-23 02:03:34
108	1119703315397345282	1	Unofficial #Astros marketing spokesperson, @A...	3170	2019-04-20 20:44:08
158	1117892268432576513	1	RT @ABREG_1: Thank you Jackie! https://t.co/...	0	2019-04-15 20:47:40
325	1114610992367972352	1	RT @ABREG_1: Wish you were here @trvisXX h...	0	2019-04-06 19:29:03
341	1114335663388680193	1	First RBI of the home slate. Courtesy: @ABRE...	836	2019-04-06 01:15:00
427	1112050016540086273	1	RT @InfieldChatter: .@astros 3B Alex Bregman...	0	2019-03-30 17:52:39
429	1112006272226869249	1	Huuuuge day celebrating @ABREG_1 and @JS...	9664	2019-03-30 14:58:50

Result Grid | Form Editor | Field Types

10. How many tags are related to “astros”?

```
SELECT DISTINCT TagTagID, COUNT(TagTagID) FROM post_tag WHERE PostSNSpostID IN(
SELECT SNSpostID FROM post WHERE SNSUserID IN(
SELECT SNSUserID FROM snsuser WHERE userName = "astros"))
ORDER BY COUNT(TagTagID)
LIMIT 10
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

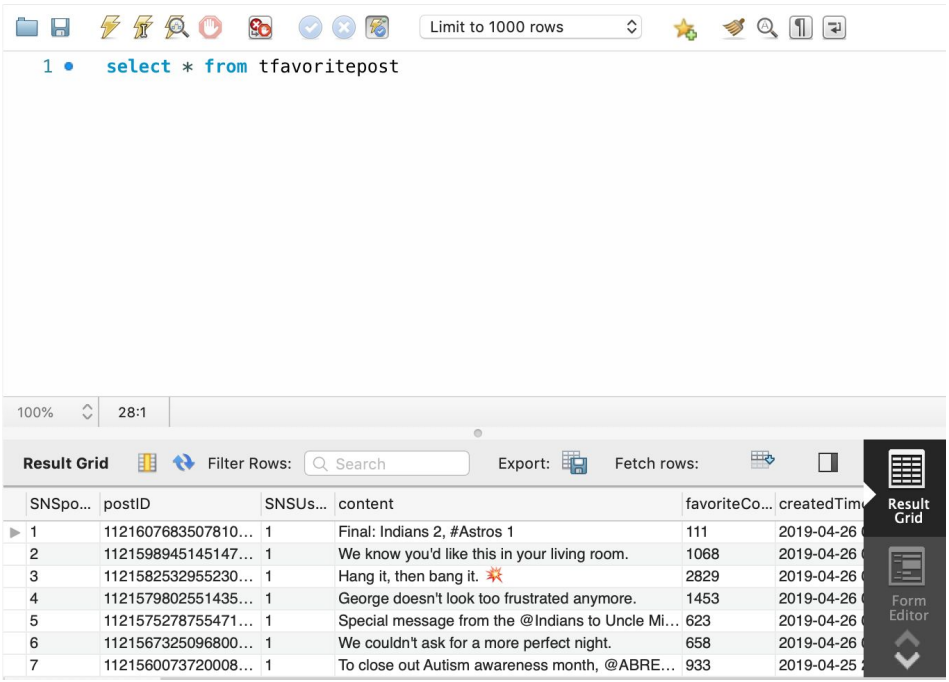
TagTagID	COUNT(TagTagID)
1	688

Result Grid | Form Editor

4 VIEWS

1) what twitter posts get favorites more than 3?

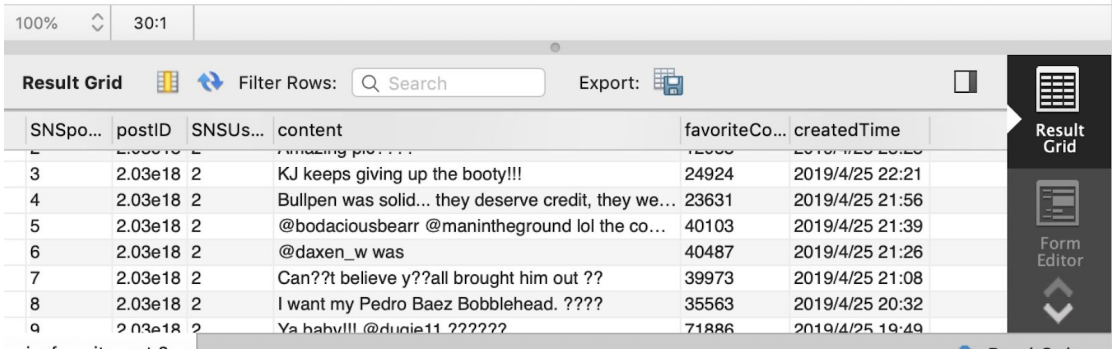
```
CREATE
    ALGORITHM = UNDEFINED
    DEFINER = `root`@`localhost`
    SQL SECURITY DEFINER
VIEW `tfavoritepost` AS
    SELECT
        `zpost`.`SNSpostID` AS `SNSpostID`,
        `zpost`.`postID` AS `postID`,
        `zpost`.`SNSUserID` AS `SNSUserID`,
        `zpost`.`content` AS `content`,
        `zpost`.`favoriteCount` AS `favoriteCount`,
        `zpost`.`createdTime` AS `createdTime`
    FROM
        `zpost`
    WHERE
        (`zpost`.`favoriteCount` > 3)
```



2) what ins posts get favorites more than 3?

```
1  CREATE
2      ALGORITHM = UNDEFINED
3      DEFINER = `root`@`localhost`
4      SQL SECURITY DEFINER
5  VIEW `insfavoritepost` AS
6      SELECT
7          `wpost`.`SNSpostID` AS `SNSpostID`,
8          `wpost`.`postID` AS `postID`,
9          `wpost`.`SNSUserID` AS `SNSUserID`,
10         `wpost`.`content` AS `content`,
11         `wpost`.`favoriteCount` AS `favoriteCount`,
12         `wpost`.`createdTime` AS `createdTime`
13     FROM
14         `wpost`
15     WHERE
16         (`wpost`.`favoriteCount` > 3)
```

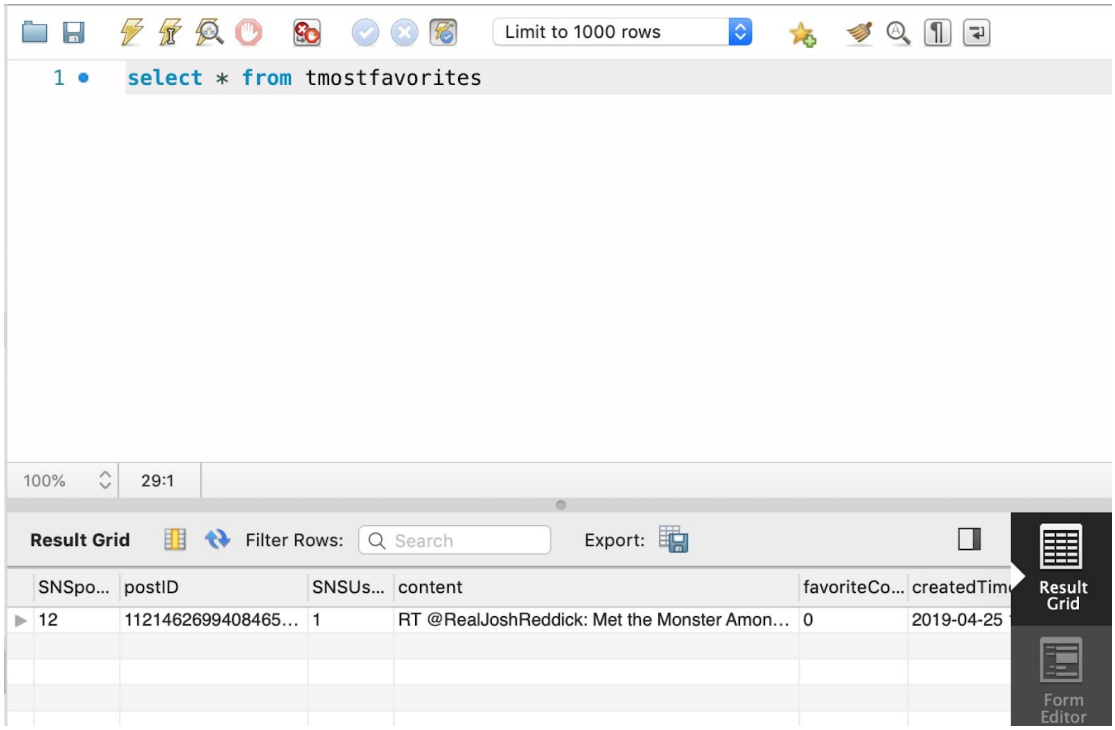
```
1 •  select * from insfavoritepost
```



SNSpo...	postID	SNSUs...	content	favoriteCo...	createdTime
3	2.03e18	2	KJ keeps giving up the booty!!!	24924	2019/4/25 22:21
4	2.03e18	2	Bullpen was solid... they deserve credit, they we…	23631	2019/4/25 21:56
5	2.03e18	2	@bodaciousbearr @manintheground lol the co…	40103	2019/4/25 21:39
6	2.03e18	2	@daxen_w was	40487	2019/4/25 21:26
7	2.03e18	2	Can??t believe y??all brought him out ??	39973	2019/4/25 21:08
8	2.03e18	2	I want my Pedro Baez Bobblehead. ????	35563	2019/4/25 20:32
9	2.03e18	2	Ya baby!!! @dugie11 ??????	71886	2019/4/25 19:49

3)which twitter post get the most favorites?

```
1 •  CREATE
2        ALGORITHM = UNDEFINED
3        DEFINER = `root`@`localhost`
4        SQL SECURITY DEFINER
5    VIEW `tmostfavorites` AS
6        SELECT
7            `zpost`.`SNSpostID` AS `SNSpostID`,
8            `zpost`.`postID` AS `postID`,
9            `zpost`.`SNSUserID` AS `SNSUserID`,
10           `zpost`.`content` AS `content`,
11           `zpost`.`favoriteCount` AS `favoriteCount`,
12           `zpost`.`createdTime` AS `createdTime`
13       FROM
14           `zpost`
15       ORDER BY `zpost`.`favoriteCount`
16       LIMIT 1
```



4)which ins post get the most favorites?

```
1  CREATE
2      ALGORITHM = UNDEFINED
3      DEFINER = `root`@`localhost`
4      SQL SECURITY DEFINER
5  VIEW `insmostfavorites` AS
6      SELECT
7          `wpost`.`SNSpostID` AS `SNSpostID`,
8          `wpost`.`postID` AS `postID`,
9          `wpost`.`SNSUserID` AS `SNSUserID`,
10         `wpost`.`content` AS `content`,
11         `wpost`.`favoriteCount` AS `favoriteCount`,
12         `wpost`.`createdTime` AS `createdTime`
13     FROM
14         `wpost`
15     ORDER BY `wpost`.`favoriteCount`
16     LIMIT 1
```

```
1 •  select * from insmostfavorites
```

100% 31:1

Result Grid | Filter Rows: Search | Export:

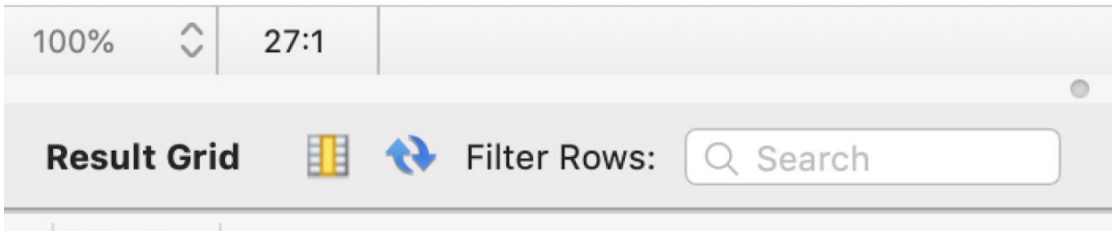
SNSpo...	postID	SNSUs...	cont...	favoriteCo...	createdTime
13	2.03e18	13		2	2019/4/26 10:59

4 FUNCTIONS

1)get the twitter SNSuserid whose number of favorite is more than the given number.

```
CREATE FUNCTION `twitter_favorite` (favorite INT)
RETURNS INTEGER
BEGIN
SELECT SNSUserID FROM zpost where favoriteCount>=favorite
RETURN SNSUserID;
END
```

```
1 •   select twitter_favorite(8)
```



2)get the ins SNSuserid whose number of favorite is more than the given number.

```
CREATE FUNCTION `ins_favorite` (favorite INT)
RETURNS INTEGER
BEGIN
SELECT SNSUserID FROM wpost where favoriteCount>=favorite
RETURN SNSUserID;
END
```

```
1 •   select ins_favorite(8)
```

100% 23:1

Result Grid Filter Rows: Search

SNSUs...
2
2
2
2
2
2
2

3)get the twitter SNSUserID which concludes target keywords

```
CREATE FUNCTION `tcontent_target` (keywords VARCHAR(255))
RETURNS VARCHAR(255)
BEGIN
SELECT SNSUserID FROM wpost
WHERE content like "%keywords%"
RETURN SNSUserID;
```

```
1    select tcontent_target(I)
```

100% 26:1

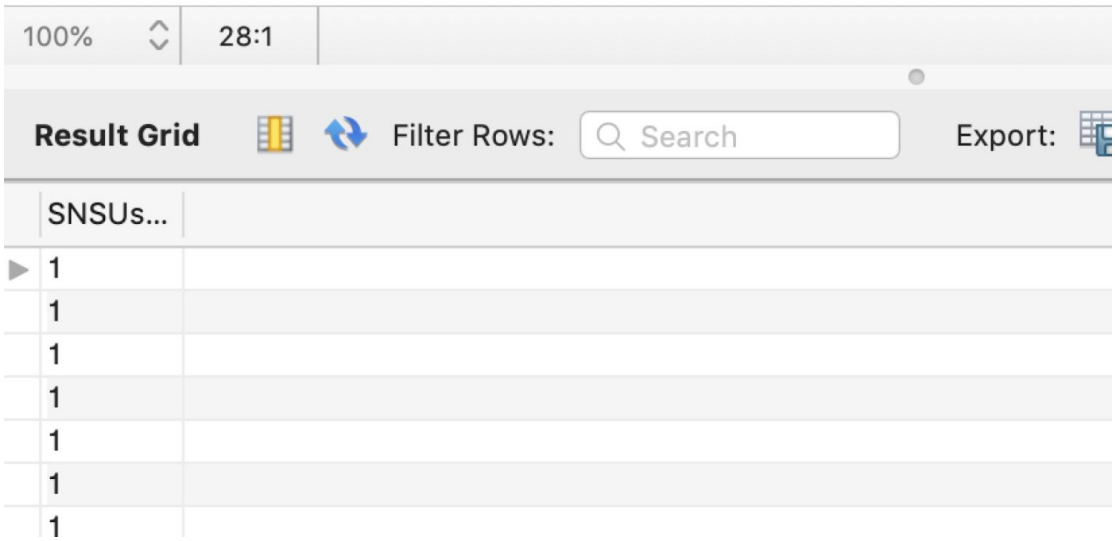
Result Grid Filter Rows: Search Export:

SNSUs...
2
2
2
2
2
2
2

4) get the ins SNSUserID which concludes target keywords

```
CREATE FUNCTION `inscontent_target` (keywords VARCHAR(255))
RETURNS VARCHAR(255)
BEGIN
SELECT SNSUserID FROM zpost
WHERE content like "%keywords%"
RETURN SNSUserID;
END
```

```
1 •   select inscontent_target(I)
```



Conclusion

In this project, we build a database about baseball which includes the name、location and post information about baseball player, team and their fans. Our database is able to tag the social media data that we collect which can help us with the mess of social media tagging. We create some tables to define the tags which are synonyms, mis-spellings and their categories to better do the tagging work. Thus, we can know some questions like “what topics are trending in my domain” and “should I follow somebody back” after our tagging work.

References

https://tweepy.readthedocs.io/en/v3.5.0
https://developer.twitter.com/en/docs
https://instaloader.github.io/
https://rosettacode.org/wiki/Levenshtein_distance#Python