收集

导入数据集

In [1]:

```
import json
import requests
import pandas as pd
import numpy as np
import os
import re
```

In [2]:

```
folder_name = 'udacity_project2'
if not os.path.exists(folder_name):
    os.makedirs(folder_name)

url = 'https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/image-predictions.tsv'
response = requests.get(url)
```

In [3]:

```
response
```

Out[3]:

```
<Response [200]>
```

In [4]:

```
with open(os.path.join(folder_name,
                       url.split('/')[-1]), mode='wb') as file:
    file.write(response.content)
```

In [5]:

```
image_predictions = pd.read_csv('udacity_project2/image-predictions.tsv',sep='\t')
```

In [6]:

```
image_predictions.head()
```

Out[6]:

	tweet_id	jpg_url	img_num
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1

In [7]:

```
image_predictions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

In [8]:

```
url_2 = 'https://raw.githubusercontent.com/udacity/new-dand-advanced-china/maste
r/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/twitter-arch
ive-enhanced.csv'
response2 = requests.get(url_2)
```

In [9]:

```
with open(os.path.join(folder_name,
                       url_2.split('/')[-1]), mode='wb') as file:
    file.write(response2.content)
```

In [10]:

```
twitter_archive_enhanced = pd.read_csv('udacity_project2/twitter-archive-enhance
d.csv')
```

In [11]:

```
twitter_archive_enhanced.head()
```

Out[11]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="ht r...
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="ht r...
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="ht r...
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="ht r...
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="ht r...

In [12]:

```
twitter_archive_enhanced['retweeted_status_timestamp'].count()
```

Out[12]:

```
181
```

In [13]:

```
twitter_archive_enhanced.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
rating_denominator            2356 non-null int64
name                          2356 non-null object
doggo                         2356 non-null object
floofer                       2356 non-null object
pupper                        2356 non-null object
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

质量

twitter_archive_enhanced 表格

- 不包括转发数据(retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp，删除是转发的数据行)
- timestamp 错误的数据类型
- 本实验中source不具有分析意义
- 本实验中不具有分析意义，清理列（retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp

  in_reply_to_status_id、in_reply_to_user_id、expanded_urls,expanded_urls）

- 缺少狗的名称、评分和评级（无法清理）
- name 数据有误
- new_rating_numerator、new_rating_denominator数据有误
- 评级列数据有误
- 统一评分模式，分母统一为10
- 错误的数据类型，status(地位）应该是catogary

清洁度

- twitter_archive_enhanced 表格缺少喜爱和转发数两列
- twitter_archive_enhanced 表格中表示地位的四列（doggo、floofer、pupper、puppo）表示同一个变量”地位“
- 对于狗狗品种的猜测和喜爱数，转发数应该都属于twitter_archive_enhanced表格的一部分

清理

In [14]:

```
image_predictions_clean = image_predictions.copy()
twitter_archive_enhanced_clean = twitter_archive_enhanced.copy()
```

缺失数据

twitter_archive_enhanced：缺少记录 (喜爱和转发数)

代码

In [15]:

```
# 从text中提取缺少的转发和喜爱数，保留tweet_id之后和原始数据集合并
df_tweet = []
with open('udacity_project2/tweet_json.txt', encoding='utf-8') as file:
    for line in file.readlines():
        dic = json.loads(line)
        tweet_id = dic['id']
        retweet_count = dic['retweet_count']
        favorite_count = dic['favorite_count']
        text = dic['full_text']
        df_tweet.append({'tweet_id': tweet_id,
                         'retweet_count': retweet_count,
                         'favorite_count': favorite_count})


df = pd.DataFrame(df_tweet, columns = ['tweet_id','retweet_count','favorite_coun
t'])
```

测试

In [16]:

```
df.head()
```

Out[16]:

	tweet_id	retweet_count	favorite_count
0	892420643555336193	8842	39492
1	892177421306343426	6480	33786
2	891815181378084864	4301	25445
3	891689557279858688	8925	42863
4	891327558926688256	9721	41016

In [17]:

```
df.tail()
```

Out[17]:

	tweet_id	retweet_count	favorite_count
2347	666049248165822465	41	111
2348	666044226329800704	147	309
2349	666033412701032449	47	128
2350	666029285002620928	48	132
2351	666020888022790149	530	2528

In [18]:

```
df.isnull().sum()
```

Out[18]:

```
tweet_id          0
retweet_count     0
favorite_count    0
dtype: int64
```

In [19]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 3 columns):
tweet_id          2352 non-null int64
retweet_count     2352 non-null int64
favorite_count    2352 non-null int64
dtypes: int64(3)
memory usage: 55.2 KB
```

删除列 in_reply_to_status_id、in_reply_to_user_id、source

In [20]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['in_reply_
to_status_id',
                                                                      'in_reply_
to_user_id','source','expanded_urls'],axis=1)
```

In [21]:

```
twitter_archive_enhanced_clean.head()
```

Out[21]:

	tweet_id	timestamp	text	retweeted_status_id	retweeted_stat
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	NaN	NaN
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	NaN	NaN
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	NaN	NaN
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	NaN	NaN
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	NaN	NaN

清理转发数据

In [22]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean[twitter_archive_
enhanced_clean.retweeted_status_id.isnull()]
```

测试 确保数据集不包含转发的数据

In [23]:

```
twitter_archive_enhanced_clean.retweeted_status_id.count()
```

Out[23]:

0

In [24]:

```
twitter_archive_enhanced_clean.retweeted_status_user_id.count()
```

Out[24]:

0

In [25]:

```
twitter_archive_enhanced_clean.retweeted_status_timestamp.count()
```

Out[25]:

0

In [26]:

```
twitter_archive_enhanced_clean.shape
```

Out[26]:

(2175, 13)

删除列 retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp

In [27]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['retweeted
_status_id',
                                                                      'retweeted
_status_user_id','retweeted_status_timestamp'],axis=1)
```

In [28]:

```
twitter_archive_enhanced_clean.head()
```

Out[28]:

	tweet_id	timestamp	text	rating_numerator	rating_denominato
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

twitter_archive_enhanced 表格四列中的一个变量 (doggo、floofer、pupper、puppo)

In [29]:

```
twitter_archive_enhanced_clean.columns
```

Out[29]:

```
Index(['tweet_id', 'timestamp', 'text', 'rating_numerator',
       'rating_denominator', 'name', 'doggo', 'floofer', 'pupper',
'puppo'],
      dtype='object')
```

In [30]:

```
twitter_archive_enhanced_clean = pd.melt(twitter_archive_enhanced_clean,id_vars
= ['tweet_id', 'timestamp', 'text', 'rating_numerator','rating_denominator', 'na
me'],
             value_vars = ['doggo', 'floofer', 'pupper', 'puppo'],value_name =
'type')
```

In [31]:

```
twitter_archive_enhanced_clean.head()
```

Out[31]:

	tweet_id	timestamp	text	rating_numerator	rating_denomin
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

In [32]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['variable'
],axis=1)
```

In [33]:

```
twitter_archive_enhanced_clean.head()
```

Out[33]:

	tweet_id	timestamp	text	rating_numerator	rating_denominato
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

In [34]:

```
twitter_archive_enhanced_clean.duplicated().sum()
```

Out[34]:

```
6169
```

In [35]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop_duplicates
()
```

In [36]:

```
#此时的twitter_archive_enhanced_clean包含了相同tweet_id，但是type是None或者四种地位的其中一种，有重复
twitter_archive_enhanced_clean.head()
```

Out[36]:

	tweet_id	timestamp	text	rating_numerator	rating_denominato
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

In [37]:

```
#提取出tweet_id没有重复的值
da = twitter_archive_enhanced_clean[~(twitter_archive_enhanced_clean.tweet_id.duplicated(keep=False))]
```

In [38]:

```
#提取出type不是None的值
db = twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.type!='None']
```

In [39]:

```
#拼接
twitter_archive_enhanced_clean = da.append(db)
```

In [40]:

```
#得出数据集包含狗狗地位，且tweet_id不重复
twitter_archive_enhanced_clean.head()
```

Out[40]:

	tweet_id	timestamp	text	rating_numerator	rating_denomin
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

喜爱数和转发数应是 `twitter_archive_enhanced` 表格的一部分

定义

将 *转发 喜爱* 两列合并到 `twitter_archive_enhanced` 表格中，按照 *tweet_id* 进行合并。

代码

In [41]:

```
twitter_archive_enhanced_clean = pd.merge(twitter_archive_enhanced_clean,df,
                                          on=['tweet_id'],how='inner')
```

In [42]:

```
twitter_archive_enhanced_clean.head()
```

Out[42]:

	tweet_id	timestamp	text	rating_numerator	rating_denomina
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

将 图片预测数据 合并到 `twitter_archive_enhanced` 表格中，按照 *tweet_id* 进行合并。 得到只含有图片的原始评级（不包括转发）

In [43]:

```
twitter_archive_enhanced_clean = pd.merge(twitter_archive_enhanced_clean,image_predictions,
                                          on=['tweet_id'],how='inner')
```

In [44]:

```
twitter_archive_enhanced_clean.head()
```

Out[44]:

	tweet_id	timestamp	text	rating_numerator	rating_denominato
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

`twitter_archive_enhanced` 表格四列中的一个变量 (doggo、floofer、pupper、puppo)直接用text中提取出来的玛位信息替代

In [45]:

```
#从text中提取宠物评级
twitter_archive_enhanced_clean['status'] = 'None'
dog_lists = ['doggo','floofer','pupper','puppo']
for i in range(0,len(twitter_archive_enhanced_clean)):
    text = twitter_archive_enhanced_clean.text[i].lower()
    for dog_status in dog_lists:
        if dog_status in text:
            twitter_archive_enhanced_clean.status[i] = dog_status
```

```
/Users/caicai/anaconda/envs/python3.6/lib/python3.6/site-packages/ip
ykernel_launcher.py:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
```

In [46]:

```
twitter_archive_enhanced_clean.head()
```

Out[46]:

	tweet_id	timestamp	text	rating_numerator	rating_denominato
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

5 rows × 21 columns

In [47]:

```
twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.type != twitter_ar
chive_enhanced_clean.status]
```

Out[47]:

	tweet_id	timestamp	text	rating_numerator	rating_d
41	881666595344535552	2017-07-03 00:11:11 +0000	This is Gary. He couldn't miss this puppertuni...	13	10
61	876120275196170240	2017-06-17 16:52:05 +0000	Meet Venti, a seemingly caffeinated puppoccino...	13	10
73	871879754684805121	2017-06-06 00:01:46 +0000	Say hello to Lassie. She's celebrating #PrideM...	13	10
89	866868824827068416	2017-05-22 16:06:55 +0000	This is Lili. She can't believe you betrayed h...	12	10
178	841439858740625411	2017-03-14 00:04:30 +0000	Here we have some incredible doggos for #K9Vet...	14	10
194	837366284874571778	2017-03-02 18:17:34 +0000	This is Lucy. She has a portrait of herself on...	13	10
291	815990720817401858	2017-01-02 18:38:42 +0000	Meet Jack. He's one of the rare doggos that do...	11	10
330	807010152071229440	2016-12-08 23:53:08 +0000	This is Lennon. He's a Boopershnoop Pupperdoop...	12	10
335	805826884734976000	2016-12-05 17:31:15 +0000	This is Duke. He is not a fan of the pupporazz...	12	10
478	772877495989305348	2016-09-05 19:22:09 +0000	You need to watch these two doggos argue throu...	11	10

	tweet_id	timestamp	text	rating_numerator	rating_d
575	753420520834629632	2016-07-14 02:47:04 +0000	Here we are witnessing an isolated squad of bo...	11	10
604	749036806121881602	2016-07-02 00:27:45 +0000	This is Dietrich. He hops at random. Other dog...	8	10
608	748575535303884801	2016-06-30 17:54:50 +0000	This is one of the most reckless puppers I've ...	6	10
632	746056683365994496	2016-06-23 19:05:49 +0000	This is Arlen and Thumpelina. They are best pa...	11	10
677	737310737551491075	2016-05-30 15:52:33 +0000	Everybody stop what you're doing and watch the...	13	10
695	731156023742988288	2016-05-13 16:15:54 +0000	Say hello to this unbelievably well behaved sq...	204	170
770	714606013974974464	2016-03-29 00:12:05 +0000	Here are two lil cuddly puppers. Both 12/10 wo...	12	10
776	713900603437621249	2016-03-27 01:29:02 +0000	Happy Saturday here's 9 puppers on a bench. 99...	99	90
798	710658690886586372	2016-03-18 02:46:49 +0000	Here's a brigade of puppers. All look very pre...	80	80
806	709901256215666688	2016-03-16 00:37:03 +0000	WeRateDogs stickers are here and they're 12/10...	12	10
871	704054845121142784	2016-02-28 21:25:30 +0000	Here is a whole flock of puppers. 60/50 I'll ...	60	50

	tweet_id	timestamp	text	rating_numerator	rating_d
989	690959652130045952	2016-01-23 18:09:53 +0000	This golden is happy to refute the soft mouth ...	11	10
1067	684225744407494656	2016-01-05 04:11:44 +0000	Two sneaky puppers were not initially seen, mo...	143	130
1068	684222868335505415	2016-01-05 04:00:18 +0000	Someone help the girl is being mugged. Several...	121	110
1069	684200372118904832	2016-01-05 02:30:55 +0000	Gang of fearless hoofed puppers here. Straight...	6	10
1075	683857920510050305	2016-01-04 03:50:08 +0000	Meet Sadie. She fell asleep on the beach and h...	10	10
1122	680583894916304897	2015-12-26 03:00:19 +0000	This is Penny. Her tennis ball slowly rolled d...	8	10
1124	680494726643068929	2015-12-25 21:06:00 +0000	Here we have uncovered an entire battalion of ...	26	10
1172	677716515794329600	2015-12-18 05:06:23 +0000	IT'S PUPPERGEDDON. Total of 144/120 ...I think...	144	120
1207	676440007570247681	2015-12-14 16:34:00 +0000	Hope your Monday isn't too awful. Here's two b...	11	10
1221	675853064436391936	2015-12-13 01:41:41 +0000	Here we have an entire platoon of puppers. Tot...	88	80
1223	675820929667219457	2015-12-12 23:34:00 +0000	Here's a handful of sleepy puppers. All look u...	11	10

	tweet_id	timestamp	text	rating_numerator	rating_d
1237	675432746517426176	2015-12-11 21:51:30 +0000	Happy Friday. Here's some golden puppers. 12/1...	12	10
1269	674664755118911488	2015-12-09 18:59:46 +0000	This is Rodman. He's getting destroyed by the ...	10	10
1290	674045139690631169	2015-12-08 01:57:39 +0000	Herd of wild dogs here. Not sure what they're ...	3	10
1479	669993076832759809	2015-11-26 21:36:12 +0000	This is Zoey. Her dreams of becoming a hippo b...	9	10
1693	858843525470990336	2017-05-01 00:40:27 +0000	I have stumbled puppon a doggo painting party....	13	10
1694	855851453814013952	2017-04-22 18:31:02 +0000	Here's a puppo participating in the #ScienceMa...	13	10
1696	854010172552949760	2017-04-17 16:34:26 +0000	At first I thought this was a shy doggo, but i...	11	10
1720	817777686764523521	2017-01-07 16:59:28 +0000	This is Dido. She's playing the lead role in "...	13	10
1726	808106460588765185	2016-12-12 00:29:28 +0000	Here we have Burke (pupper) and Dexter (doggo)...	12	10
1728	802265048156610565	2016-11-25 21:37:47 +0000	Like doggo, like pupper version 2. Both 11/10 ...	11	10
1730	801115127852503040	2016-11-22 17:28:25 +0000	This is Bones. He's being haunted by another d...	12	10

	tweet_id	timestamp	text	rating_numerator	rating_d
1737	785639753186217984	2016-10-11 00:34:48 +0000	This is Pinot. He's a sophisticated doggo. You...	10	10
1746	759793422261743616	2016-07-31 16:50:42 +0000	Meet Maggie & Lila. Maggie is the doggo, L...	12	10
1755	751583847268179968	2016-07-09 01:08:47 +0000	Please stop sending it pictures that don't eve...	5	10
1762	741067306818797568	2016-06-10 00:39:48 +0000	This is just downright precious af. 12/10 for ...	12	10
1766	733109485275860992	2016-05-19 01:38:16 +0000	Like father (doggo), like son (pupper). Both 1...	12	10

48 rows × 21 columns

In [48]:

```
twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.tweet_id==806826884734976000].text
```

Out[48]:

```
335    This is Duke. He is not a fan of the pupporazz...
Name: text, dtype: object
```

对比type和status有差异的数据，发现一般是doggo，floofer，pupper，puppo的延伸词，例如：pupporazzi，puppers等。决定保留这些地位信息并归于相应类别。所以用status列表示狗狗的地位，删除type列。

In [49]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['type'],axis=1)
```

In [50]:

```
twitter_archive_enhanced_clean.head()
```

Out[50]:

	tweet_id	timestamp	text	rating_numerator	rating_denominato
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	12	10
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	13	10
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	12	10

从text列中提取宠物评级分子和分母

In [51]:

```
#从text中获取宠物评级分子
twitter_archive_enhanced_clean['new_rating_numerator'] = twitter_archive_enhance
d_clean.text.str.extract('((?:\d+\.)?\d+)/\d+',expand=True)
```

In [52]:

```
#从text中获取宠物评级分母
twitter_archive_enhanced_clean['new_rating_denominator'] = twitter_archive_enhan
ced_clean.text.str.extract('\d+/(\d+)',expand=True)
```

In [53]:

```
twitter_archive_enhanced_clean.new_rating_numerator = twitter_archive_enhanced_c
lean.new_rating_numerator.astype(float)
```

In [54]:

```
twitter_archive_enhanced_clean.new_rating_denominator = twitter_archive_enhanced
_clean.new_rating_denominator.astype(int)
```

查看rating_numerator、rating_denominator和new_rating_numerator、new_rating_denominator存在差异的行

In [55]:

```
twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.rating_numerator!=
twitter_archive_enhanced_clean.new_rating_numerator]
```

Out[55]:

	tweet_id	timestamp	text	rating_numerator	rating_denomi
34	883482846933004288	2017-07-08 00:28:19 +0000	This is Bella. She hopes her smile made you sm...	5	10
416	786709082849828864	2016-10-13 23:23:56 +0000	This is Logan, the Chow who lived. He solemnly...	75	10
1124	680494726643068929	2015-12-25 21:06:00 +0000	Here we have uncovered an entire battalion of ...	26	10
1810	778027034220126208	2016-09-20 00:24:34 +0000	This is Sophie. She's a Jubilant Bush Pupper. ...	27	10

4 rows × 22 columns

In [56]:

```
twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.rating_denominator
!=twitter_archive_enhanced_clean.new_rating_denominator]
```

Out[56]:

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	retweet_cou

0 rows × 22 columns

通过查看text知道，rating_numerator取值有错误是因为分子是小数的时候没有正确取值

删除rating_numerator、rating_denominator列

In [57]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['rating_nu
merator','rating_denominator'],axis=1)
```

In [58]:

```
twitter_archive_enhanced_clean.head()
```

Out[58]:

	tweet_id	timestamp	text	name	retweet_count	favorite_cou
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	Phineas	8842	39492
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	Tilly	6480	33786
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	Archie	4301	25445
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	Darla	8925	42863
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	Franklin	9721	41016

从text列中提取宠物名字

In [59]:

```
#从text中提取宠物名字
twitter_archive_enhanced_clean['dog_name'] = twitter_archive_enhanced_clean.text
.str.extract('(?:This is|Here is|Meet|name is|Say hello to|named) ([A-Z][a-z]+)'
,expand=True)
```

In [60]:

```
twitter_archive_enhanced_clean.head()
```

Out[60]:

	tweet_id	timestamp	text	name	retweet_count	favorite_cou
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	Phineas	8842	39492
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	Tilly	6480	33786
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	Archie	4301	25445
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	Darla	8925	42863
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	Franklin	9721	41016

5 rows × 21 columns

In [61]:

```
#查看name列和dog_name列不同的数据
twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.name != twitter_ar
chive_enhanced_clean.dog_name].sample(20)
```

Out[61]:

	tweet_id	timestamp	text	name	retweet_count	favo
1788	854120357044912130	2017-04-17 23:52:16 +0000	Sometimes you guys remind me just how impactfu...	None	8245	3379
1172	677716515794329600	2015-12-18 05:06:23 +0000	IT'S PUPPERGEDDON. Total of 144/120 ...I think...	None	1101	3310
1037	687102708889812993	2016-01-13 02:43:46 +0000	Army of water dogs here. None of them know whe...	None	1115	2560
565	756303284449767430	2016-07-22 01:42:09 +0000	Pwease accept dis rose on behalf of dog. 11/10...	None	1224	4345
1321	673317986296586240	2015-12-06 01:48:12 +0000	Take a moment and appreciate how these two dog...	None	292	920
1862	703407252292673536	2016-02-27 02:32:12 +0000	This pupper doesn't understand gates. 10/10 so...	None	779	2677
1025	688116655151435777	2016-01-15 21:52:49 +0000	Please send dogs. I'm tired of seeing other st...	None	885	3083
1344	672482722825261057	2015-12-03 18:29:09 +0000	This is light saber pup. Ready to fight off ev...	light	662	1212
1338	672604026190569472	2015-12-04 02:31:10 +0000	This is a baby Rand Paul. Curls for days. 11/1...	a	441	1177
1150	679047485189439488	2015-12-21 21:15:11 +0000	This dog doesn't know how to stairs. Quite tra...	None	773	2457

	tweet_id	timestamp	text	name	retweet_count	favo
1723	813202720496779264	2016-12-26 02:00:11 +0000	Here's a doggo who has concluded that Christma...	None	2079	1016
1952	675334060156301312	2015-12-11 15:19:21 +0000	Good morning here's a grass pupper. 12/10 http...	None	1428	2994
1558	668297328638447616	2015-11-22 05:17:54 +0000	2 rare dogs. They waddle (v inefficient). Some...	None	318	653
1957	674638615994089473	2015-12-09 17:15:54 +0000	This pupper is fed up with being tickled. 12/1...	None	649	1795
575	753420520834629632	2016-07-14 02:47:04 +0000	Here we are witnessing an isolated squad of bo...	None	4049	8690
1044	686050296934563840	2016-01-10 05:01:51 +0000	This is Flávio. He's a Macedonian Poppycock. 9...	Flávio	832	2411
1310	673636718965334016	2015-12-06 22:54:44 +0000	This is a Lofted Aphrodisiac Terrier named Kip...	a	403	1191
1988	859607811541651456	2017-05-03 03:17:27 +0000	Sorry for the lack of posts today. I came home...	None	1695	1939
1031	687480748861947905	2016-01-14 03:45:57 +0000	Another magnificent photo. 12/10 https://t.co/...	None	281	1753
1751	755206590534418437	2016-07-19 01:04:16 +0000	This is one of the most inspirational stories ...	one	6120	1813

20 rows × 21 columns

对比看出name中确实存在很多名字提取错误，例如：a、an、the等，我们删除name列，保留从text中提取出的dog_name

In [62]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.drop(['name'],ax
is=1)
```

In [63]:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced_clean.rename(columns={
'dog_name':'name',
                                                    'new_rating_numerator':'rating_nu
merator',
                                                    'new_rating_denominator':'rating_
denominator'})
```

In [64]:

```
twitter_archive_enhanced_clean.head()
```

Out[64]:

	tweet_id	timestamp	text	retweet_count	favorite_count	
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	8842	39492	http
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	6480	33786	http
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	4301	25445	http
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	8925	42863	http
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	9721	41016	http

In [65]:

```
#不按照正常评分分母是10来评分的数据
twitter_archive_enhanced_clean[twitter_archive_enhanced_clean.rating_denominator
 != 10]
```

Out[65]:

	tweet_id	timestamp	text	retweet_count	favorite_c
267	820690176645140481	2017-01-15 17:52:40 +0000	The floofs have been released I repeat the flo...	3699	13476
317	810984652412424192	2016-12-19 23:06:23 +0000	Meet Sam. She smiles 24/7 & secretly aspir...	1647	5904
553	758467244762497024	2016-07-28 01:00:57 +0000	Why does this never happen at my front door.....	2516	5297
658	740373189193256964	2016-06-08 02:41:38 +0000	After so many requests, this is Bretagne. She ...	15029	37704
695	731156023742988288	2016-05-13 16:15:54 +0000	Say hello to this unbelievably well behaved sq...	1427	4172
728	722974582966214656	2016-04-21 02:25:47 +0000	Happy 4/20 from the squad! 13/10 for all https...	1754	4473
756	716439118184652801	2016-04-03 01:36:11 +0000	This is Bluebert. He just saw that both #Final...	246	2562
776	713900603437621249	2016-03-27 01:29:02 +0000	Happy Saturday here's 9 puppers on a bench. 99...	827	3049
798	710658690886586372	2016-03-18 02:46:49 +0000	Here's a brigade of puppers. All look very pre...	633	2513
815	709198395643068416	2016-03-14 02:04:08 +0000	From left to right:\nCletus, Jerome, Alejandro...	716	2623

	tweet_id	timestamp	text	retweet_count	favorite_c
871	704054845121142784	2016-02-28 21:25:30 +0000	Here is a whole flock of puppers. 60/50 I'll ...	1023	3193
934	697463031882764288	2016-02-10 16:51:59 +0000	Happy Wednesday here's a bucket of pups. 44/40...	1547	3730
1067	684225744407494656	2016-01-05 04:11:44 +0000	Two sneaky puppers were not initially seen, mo...	239	1361
1068	684222868335505415	2016-01-05 04:00:18 +0000	Someone help the girl is being mugged. Several...	1554	4198
1086	682962037429899265	2016-01-01 16:30:13 +0000	This is Darrel. He just robbed a 7/11 and is i...	18393	39005
1172	677716515794329600	2015-12-18 05:06:23 +0000	IT'S PUPPERGEDDON. Total of 144/120 ...I think...	1101	3310
1221	675853064436391936	2015-12-13 01:41:41 +0000	Here we have an entire platoon of puppers. Tot...	1447	2903
1667	666287406224695296	2015-11-16 16:11:11	This is an Albanian 3 1/2 legged Episcopalian	71	152

In [66]:

```
#新增一列分子除以分母
twitter_archive_enhanced_clean['rating_score'] = twitter_archive_enhanced_clean.
rating_numerator/twitter_archive_enhanced_clean.rating_denominator
```

In [67]:

```
twitter_archive_enhanced_clean.head()
```

Out[67]:

	tweet_id	timestamp	text	retweet_count	favorite_count	
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	8842	39492	https:
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	6480	33786	https:
2	891815181378084864	2017-07-31 00:18:03 +0000	This is Archie. He is a rare Norwegian Pouncin...	4301	25445	https:
3	891689557279858688	2017-07-30 15:58:51 +0000	This is Darla. She commenced a snooze mid meal...	8925	42863	https:
4	891327558926688256	2017-07-29 16:00:24 +0000	This is Franklin. He would like you to stop ca...	9721	41016	https:

5 rows × 21 columns

timestamp的格式错误

In [68]:

```
twitter_archive_enhanced_clean.timestamp = pd.to_datetime(twitter_archive_enhanced_clean.timestamp)
```

In [69]:

```
twitter_archive_enhanced_clean.timestamp.head()
```

Out[69]:

```
0   2017-08-01 16:23:56
1   2017-08-01 00:17:27
2   2017-07-31 00:18:03
3   2017-07-30 15:58:51
4   2017-07-29 16:00:24
Name: timestamp, dtype: datetime64[ns]
```

In [70]:

```
twitter_archive_enhanced_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2005 entries, 0 to 2004
Data columns (total 21 columns):
tweet_id              2005 non-null int64
timestamp             2005 non-null datetime64[ns]
text                  2005 non-null object
retweet_count         2005 non-null int64
favorite_count        2005 non-null int64
jpg_url               2005 non-null object
img_num               2005 non-null int64
p1                    2005 non-null object
p1_conf               2005 non-null float64
p1_dog                2005 non-null bool
p2                    2005 non-null object
p2_conf               2005 non-null float64
p2_dog                2005 non-null bool
p3                    2005 non-null object
p3_conf               2005 non-null float64
p3_dog                2005 non-null bool
status                2005 non-null object
rating_numerator      2005 non-null float64
rating_denominator    2005 non-null int64
name                  1380 non-null object
rating_score          2005 non-null float64
dtypes: bool(3), datetime64[ns](1), float64(5), int64(5), object(7)
memory usage: 303.5+ KB
```

修改status列数据类型

In [71]:

```
twitter_archive_enhanced_clean.status = twitter_archive_enhanced_clean.status.as
type('category')
```

In [72]:

```
twitter_archive_enhanced_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2005 entries, 0 to 2004
Data columns (total 21 columns):
tweet_id              2005 non-null int64
timestamp             2005 non-null datetime64[ns]
text                  2005 non-null object
retweet_count         2005 non-null int64
favorite_count        2005 non-null int64
jpg_url               2005 non-null object
img_num               2005 non-null int64
p1                    2005 non-null object
p1_conf               2005 non-null float64
p1_dog                2005 non-null bool
p2                    2005 non-null object
p2_conf               2005 non-null float64
p2_dog                2005 non-null bool
p3                    2005 non-null object
p3_conf               2005 non-null float64
p3_dog                2005 non-null bool
status                2005 non-null category
rating_numerator      2005 non-null float64
rating_denominator    2005 non-null int64
name                  1380 non-null object
rating_score          2005 non-null float64
dtypes: bool(3), category(1), datetime64[ns](1), float64(5), int64
(5), object(6)
memory usage: 290.0+ KB
```

In [73]:

```
twitter_archive_enhanced_clean.to_csv('udacity_project2/twitter_archive_master.c
sv', index=False)
```