收集数据

- tweet无法访问 :)
- 从github仓库https://github.com/udacity/new-dand-advanced-china (https://github.com/udacity/new-dand-advanced-china)中下载文件

In [76]:

```
# -*- coding=utf-8 -*-
import requests

def download(url):
    with open(url.split('/')[-1], mode="wb") as f:
        respone = requests.get(url)
        f.write(respone.content)

# 下载 image-predictions.tsv
image_prediction_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"
download(image_prediction_url)
# 下载 twitter_archive_enhanced.csv
twitter_archive_enhanced_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/twitter-archive-enhanced.csv"
download(twitter_archive_enhanced_url)
# 下载 tweet_json.txt
tweet_json_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/tweet_json.txt"
download(tweet_json_url)
```

评估数据

In [77]:

```
# 使用pandas读取数据文件
import pandas as pd
```

In [78]:

```
image_predictions = pd.read_csv("image-predictions.tsv", sep='\t')
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: 152.0+ KB
```

In [79]:

```
image_predictions.head(3)
```

Out[79]:

	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

In [80]:

```
# 处理和评估tweet_json.txt数据
import json
with open("tweet_json.txt", 'r') as json_file:
    for data in json_file.readlines():
        print(json.dumps(json.loads(data), indent=4, sort_keys=False, ensure_ascii=True))
        break
```

```
{
    "contributors": null,
    "truncated": false,
    "is_quote_status": false,
    "in_reply_to_status_id": null,
    "id": 892420643555336193,
    "favorite_count": 39492,
    "full_text": "This is Phineas. He's a mystical boy. Only ever appears in the h
ole of a donut. 13/10 https://t.co/MgUWQ76dJU",
    "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twi
tter for iPhone</a>",
    "retweeted": false,
    "coordinates": null,
    "entities": {
        "symbols": [],
        "user_mentions": [],
        "hashtags": [],
        "urls": [],
        "media": [
            {
                "expanded_url": "https://twitter.com/dog_rates/status/892420643555
336193/photo/1",
                "display_url": "pic.twitter.com/MgUWQ76dJU",
                "url": "https://t.co/MgUWQ76dJU",
                "media_url_https": "https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jp
g",
                "id_str": "892420639486877696",
                "sizes": {
                    "large": {
                        "h": 528,
                        "resize": "fit",
                        "w": 540
                    },
                    "small": {
                        "h": 528,
                        "resize": "fit",
                        "w": 540
                    },
                    "medium": {
                        "h": 528,
                        "resize": "fit",
                        "w": 540
                    },
                    "thumb": {
                        "h": 150,
                        "resize": "crop",
                        "w": 150
                    }
                },
                "indices": [
                    86,
                    109
                ],
                "type": "photo",
                "id": 892420639486877696,
                "media_url": "http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg"
            }
        ]
    },
    "in_reply_to_screen_name": null,
    "in_reply_to_user_id": null,
```

```
        "display_text_range": [
            0,
            85
        ],
        "retweet_count": 8842,
        "id_str": "892420643555336193",
        "favorited": false,
        "user": {
            "follow_request_sent": false,
            "has_extended_profile": true,
            "profile_use_background_image": false,
            "default_profile_image": false,
            "id": 4196983835,
            "profile_background_image_url_https": "https://abs.twimg.com/images/theme
    s/theme1/bg.png",
            "verified": true,
            "translator_type": "none",
            "profile_text_color": "000000",
            "profile_image_url_https": "https://pbs.twimg.com/profile_images/914581071
    265755136/2h5uFpwU_normal.jpg",
            "profile_sidebar_fill_color": "000000",
            "entities": {
                "url": {
                    "urls": [
                        {
                            "url": "https://t.co/N7sNNHAEXS",
                            "indices": [
                                0,
                                23
                            ],
                            "expanded_url": "http://weratedogs.com",
                            "display_url": "weratedogs.com"
                        }
                    ]
                },
                "description": {
                    "urls": []
                }
            },
            "followers_count": 3768791,
            "profile_sidebar_border_color": "000000",
            "id_str": "4196983835",
            "profile_background_color": "000000",
            "listed_count": 3169,
            "is_translation_enabled": false,
            "utc_offset": null,
            "statuses_count": 5749,
            "description": "Only Legit Source for Professional Dog Ratings STORE: @Sho
    pWeRateDogs | IG, FB & SC: WeRateDogs | MOBILE APP: @GoodDogsGame Business: dograt
    ingtwitter@gmail.com",
            "friends_count": 107,
            "location": "MERCH\u21b4  DM DOGS. WE WILL RATE",
            "profile_link_color": "F5ABB5",
            "profile_image_url": "http://pbs.twimg.com/profile_images/9145810712657551
    36/2h5uFpwU_normal.jpg",
            "following": false,
            "geo_enabled": true,
            "profile_banner_url": "https://pbs.twimg.com/profile_banners/4196983835/15
    06888628",
            "profile_background_image_url": "http://abs.twimg.com/images/themes/theme
    1/bg.png",
```

```
        "screen_name": "dog_rates",
        "lang": "en",
        "profile_background_tile": false,
        "favourites_count": 120162,
        "name": "SpookyWeRateDogs\u2122",
        "notifications": false,
        "url": "https://t.co/N7sNNHAEXS",
        "created_at": "Sun Nov 15 21:41:29 +0000 2015",
        "contributors_enabled": false,
        "time_zone": null,
        "protected": false,
        "default_profile": false,
        "is_translator": false
    },
    "geo": null,
    "in_reply_to_user_id_str": null,
    "possibly_sensitive": false,
    "possibly_sensitive_appealable": false,
    "lang": "en",
    "created_at": "Tue Aug 01 16:23:56 +0000 2017",
    "in_reply_to_status_id_str": null,
    "place": null,
    "extended_entities": {
        "media": [
            {
                "expanded_url": "https://twitter.com/dog_rates/status/892420643555
336193/photo/1",
                "display_url": "pic.twitter.com/MgUWQ76dJU",
                "url": "https://t.co/MgUWQ76dJU",
                "media_url_https": "https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jp
g",
                "id_str": "892420639486877696",
                "sizes": {
                    "large": {
                        "h": 528,
                        "resize": "fit",
                        "w": 540
                    },
                    "small": {
                        "h": 528,
                        "resize": "fit",
                        "w": 540
                    },
                    "medium": {
                        "h": 528,
                        "resize": "fit",
                        "w": 540
                    },
                    "thumb": {
                        "h": 150,
                        "resize": "crop",
                        "w": 150
                    }
                },
                "indices": [
                    86,
                    109
                ],
                "type": "photo",
                "id": 892420639486877696,
                "media_url": "http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg"
```

```
            }
        ]
    }
}
```

In [81]:

```
# 逐行读取tweet_json.txt文件并添加到pandas DataFrame中，（至少）包含 tweet ID、retweet_count和favorite_count字段...
# 提出问题，放到list中，然后通过 'id_str': list1, 'fav_count': list2的方式创建df的效率是否更高?
df_tweet = pd.DataFrame()

# 观察tweet_json.txt后，对以下数据感兴趣：
# retweet_count favorite_count full_text retweeted source favorited
with open("tweet_json.txt", 'r') as json_file:
    index = ['id', 'retweet_count', 'favorite_count', 'full_text', 'retweeted', 'source', 'favorited']
    for data in json_file.readlines():
        dict_data = json.loads(data)

        data = []
        for idx in index:
            data.append(dict_data[idx])

        s_tweet = pd.Series(data, index=index)
        retweet_count = dict_data[u'retweet_count']
        df_tweet = df_tweet.append(s_tweet, ignore_index=True)

df_tweet.head(3)
```

Out[81]:

	favorite_count	favorited	full_text	id	retweet_count	retweeted	
0	39492.0	0.0	This is Phineas. He's a mystical boy. Only eve...	8.924206e+17	8842.0	0.0	<a href=r...
1	33786.0	0.0	This is Tilly. She's just checking pup on you....	8.921774e+17	6480.0	0.0	<a href=r...
2	25445.0	0.0	This is Archie. He is a rare Norwegian Pouncin...	8.918152e+17	4301.0	0.0	<a href=r...

In [82]:

```
df_tweet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 7 columns):
favorite_count    2352 non-null float64
favorited         2352 non-null float64
full_text         2352 non-null object
id                2352 non-null float64
retweet_count     2352 non-null float64
retweeted         2352 non-null float64
source            2352 non-null object
dtypes: float64(5), object(2)
memory usage: 128.7+ KB
```

In [83]:

```
df_tweet.describe()
```

Out[83]:

	favorite_count	favorited	id	retweet_count	retweeted
count	2352.000000	2352.0	2.352000e+03	2352.000000	2352.0
mean	8109.198980	0.0	7.425913e+17	3134.932398	0.0
std	11980.795669	0.0	6.846210e+16	5237.846296	0.0
min	0.000000	0.0	6.660209e+17	0.000000	0.0
25%	1417.000000	0.0	6.783949e+17	618.000000	0.0
50%	3596.500000	0.0	7.193536e+17	1456.500000	0.0
75%	10118.000000	0.0	7.991219e+17	3628.750000	0.0
max	132318.000000	0.0	8.924206e+17	79116.000000	0.0

In [84]:

```
df_tweet
```

Out[84]:

	favorite_count	favorited	full_text	id	retweet_coun
0	39492.0	0.0	This is Phineas. He's a mystical boy. Only eve...	8.924206e+17	8842.0
1	33786.0	0.0	This is Tilly. She's just checking pup on you....	8.921774e+17	6480.0
2	25445.0	0.0	This is Archie. He is a rare Norwegian Pouncin...	8.918152e+17	4301.0
3	42863.0	0.0	This is Darla. She commenced a snooze mid meal...	8.916896e+17	8925.0
4	41016.0	0.0	This is Franklin. He would like you to stop ca...	8.913276e+17	9721.0
5	20548.0	0.0	Here we have a majestic great white breaching ...	8.910880e+17	3240.0
6	12053.0	0.0	Meet Jax. He enjoys ice cream so much he gets ...	8.909719e+17	2142.0
7	66596.0	0.0	When you watch your owner call another dog a g...	8.907292e+17	19548.0
8	28187.0	0.0	This is Zoey. She doesn't want to be one of th...	8.906092e+17	4403.0
9	32467.0	0.0	This is Cassie. She is a college pup. Studying...	8.902403e+17	7684.0
10	31127.0	0.0	This is Koda. He is a South Australian decksha...	8.900066e+17	7584.0
11	28208.0	0.0	This is Bruno. He is a service shark. Only get...	8.898809e+17	5116.0
12	38745.0	0.0	Here's a puppo that seems to be on the fence a...	8.896654e+17	8502.0
13	27633.0	0.0	This is Ted. He does his best. Sometimes that'...	8.896388e+17	4705.0

	favorite_count	favorited	full_text	id	retweet_coun
14	15329.0	0.0	This is Stuart. He's sporting his favorite fan...	8.895311e+17	2309.0
15	25712.0	0.0	This is Oliver. You're witnessing one of his m...	8.892788e+17	5635.0
16	29555.0	0.0	This is Jim. He found a fren. Taught him how t...	8.889172e+17	4681.0
17	26021.0	0.0	This is Zeke. He has a new stick. Very proud o...	8.888050e+17	4535.0
18	20267.0	0.0	This is Ralphus. He's powering up. Attempting ...	8.885550e+17	3722.0
19	22144.0	0.0	This is Gerald. He was just told he didn't get...	8.880784e+17	3637.0
20	30690.0	0.0	This is Jeffrey. He has a monopoly on the pool...	8.877053e+17	5584.0
21	46940.0	0.0	I've yet to rate a Venezuelan Hover Wiener. Th...	8.875171e+17	12053.0
22	70007.0	0.0	This is Canela. She attempted some fancy porch...	8.874740e+17	18813.0
23	34223.0	0.0	You may not have known you needed to see this ...	8.873432e+17	10713.0
24	31045.0	0.0	This... is a Jubilant Antarctic House Bear. We...	8.871014e+17	6147.0
25	35786.0	0.0	This is Maya. She's very shy. Rarely leaves he...	8.869832e+17	8045.0
26	12286.0	0.0	This is Mingus. He's a wonderful father to his...	8.867369e+17	3420.0
27	22802.0	0.0	This is Derek. He's late for a dog meeting. 13...	8.866803e+17	4597.0

	favorite_count	favorited	full_text	id	retweet_coun
28	21488.0	0.0	This is Roscoe. Another pupper fallen victim t...	8.863661e+17	3297.0
29	117.0	0.0	@NonWhiteHat @MayhewMayhem omg hello tanner yo...	8.862670e+17	4.0
...	...	...	...	...	...
2322	457.0	0.0	This is quite the dog. Gets really excited whe...	6.664115e+17	337.0
2323	113.0	0.0	This is a southern Vesuvius bumblegruff. Can d...	6.664071e+17	43.0
2324	171.0	0.0	Oh goodness. A super rare northeast Qdoba kang...	6.663962e+17	91.0
2325	194.0	0.0	Those are sunglasses and a jean jacket. 11/10 ...	6.663738e+17	99.0
2326	801.0	0.0	Unique dog here. Very small. Lives in containe...	6.663628e+17	590.0
2327	228.0	0.0	Here we have a mixed Asiago from the Galápagos...	6.663533e+17	76.0
2328	308.0	0.0	Look at this jokester thinking seat belt laws ...	6.663454e+17	146.0
2329	203.0	0.0	This is an extremely rare horned Parthenon. No...	6.663379e+17	96.0
2330	519.0	0.0	This is a funny dog. Weird toes. Won't come do...	6.662939e+17	365.0
2331	152.0	0.0	This is an Albanian 3 1/2 legged Episcopalian...	6.662874e+17	71.0
2332	183.0	0.0	Can take selfies 11/10 https://t.co/ws2AMaNwPW	6.662731e+17	81.0
2333	108.0	0.0	Very concerned about fellow dog trapped in com...	6.662689e+17	37.0

	favorite_count	favorited	full_text	id	retweet_coun
2334	14703.0	0.0	Not familiar with this breed. No tail (weird)....	6.661041e+17	6835.0
2335	81.0	0.0	Oh my. Here you are seeing an Adobe Setter giv...	6.661022e+17	15.0
2336	160.0	0.0	Can stand on stump for what seems like a while...	6.660995e+17	73.0
2337	168.0	0.0	This appears to be a Mongolian Presbyterian mi...	6.660940e+17	78.0
2338	121.0	0.0	Here we have a well-established sunblockerspan...	6.660829e+17	47.0
2339	334.0	0.0	Let's hope this flight isn't Malaysian (lol). ...	6.660731e+17	173.0
2340	154.0	0.0	Here we have a northern speckled Rhododendron....	6.660712e+17	67.0
2341	494.0	0.0	This is the happiest dog you will ever see. Ve...	6.660638e+17	230.0
2342	117.0	0.0	Here is the Rand Paul of retrievers folks! He'...	6.660586e+17	61.0
2343	304.0	0.0	My oh my. This is a rare blond Canadian terrie...	6.660571e+17	146.0
2344	449.0	0.0	Here is a Siberian heavily armored polar bear ...	6.660555e+17	261.0
2345	1250.0	0.0	This is an odd dog. Hard on the outside but lo...	6.660519e+17	877.0
2346	136.0	0.0	This is a truly beautiful English Wilson Staff...	6.660508e+17	60.0
2347	111.0	0.0	Here we have a 1949 1st generation vulpix. Enj...	6.660492e+17	41.0

	favorite_count	favorited	full_text	id	retweet_coun
2348	309.0	0.0	This is a purebred Piers Morgan. Loves to Netf...	6.660442e+17	147.0
2349	128.0	0.0	Here is a very happy pup. Big fan of well-main...	6.660334e+17	47.0
2350	132.0	0.0	This is a western brown Mitsubishi terrier. Up...	6.660293e+17	48.0
2351	2528.0	0.0	Here we have a Japanese Irish Setter. Lost eye...	6.660209e+17	530.0

2352 rows × 7 columns

In [85]:

```
witter_archive_enhanced = pd.read_csv("twitter-archive-enhanced.csv")
witter_archive_enhanced.info()
witter_archive_enhanced.describe()
```

```
<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
```

Out[85]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	retweeted_status_id
count	2.356000e+03	7.800000e+01	7.800000e+01	1.810000e+02
mean	7.427716e+17	7.455079e+17	2.014171e+16	7.720400e+17
std	6.856705e+16	7.582492e+16	1.252797e+17	6.236928e+16
min	6.660209e+17	6.658147e+17	1.185634e+07	6.661041e+17
25%	6.783989e+17	6.757419e+17	3.086374e+08	7.186315e+17
50%	7.196279e+17	7.038708e+17	4.196984e+09	7.804657e+17
75%	7.993373e+17	8.257804e+17	4.196984e+09	8.203146e+17
max	8.924206e+17	8.862664e+17	8.405479e+17	8.874740e+17

In [86]:

```
witter_archive_enhanced
```

Out[86]:

	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 hr r..
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a hr r..
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a hr r..
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a hr r..
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a hr r..
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a hr r..
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a hr r..
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40 +0000	<a hr r..
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51 +0000	<a hr r..
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51 +0000	<a hr r..
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53 +0000	<a hr r..
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32 +0000	<a hr r..
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02 +0000	<a hr r..
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04 +0000	<a hr r..
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32 +0000	<a hr r..
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39 +0000	<a hr r..
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37 +0000	<a hr r..
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06 +0000	<a hr r..
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36 +0000	<a hr r..
20	888078434458587136	NaN	NaN	2017-07-20 16:49:33 +0000	<a hr r..
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09 +0000	<a hr r..
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34 +0000	<a hr r..
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03 +0000	<a hr r..
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08 +0000	<a hr r..
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36 +0000	<a hr r..
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41 +0000	<a hr r..
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00 +0000	<a hr r..
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31 +0000	<a hr r..
...	...	...	...	...	...
2326	666411507551481857	NaN	NaN	2015-11-17 00:24:19 +0000	<a hr r..
2327	666407126856765440	NaN	NaN	2015-11-17 00:06:54 +0000	<a hr r..
2328	666396247373291520	NaN	NaN	2015-11-16 23:23:41 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18 +0000	<a hr r..
2330	666362758909284353	NaN	NaN	2015-11-16 21:10:36 +0000	<a hr r..
2331	666353288456101888	NaN	NaN	2015-11-16 20:32:58 +0000	<a hr r..
2332	666345417576210432	NaN	NaN	2015-11-16 20:01:42 +0000	<a hr r..
2333	666337882303524864	NaN	NaN	2015-11-16 19:31:45 +0000	<a hr r..
2334	666293911632134144	NaN	NaN	2015-11-16 16:37:02 +0000	<a hr r..
2335	666287406224695296	NaN	NaN	2015-11-16 16:11:11 +0000	<a hr r..
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19 +0000	<a hr r..
2337	666268910803644416	NaN	NaN	2015-11-16 14:57:41 +0000	<a hr r..
2338	666104133288665088	NaN	NaN	2015-11-16 04:02:55 +0000	<a hr r..
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04 +0000	<a hr r..
2340	666099513787052032	NaN	NaN	2015-11-16 03:44:34 +0000	<a hr r..
2341	666094000022159362	NaN	NaN	2015-11-16 03:22:39 +0000	<a hr r..
2342	666082916733198337	NaN	NaN	2015-11-16 02:38:37 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2343	666073100786774016	NaN	NaN	2015-11-16 01:59:36 +0000	<a hr r..
2344	666071193221509120	NaN	NaN	2015-11-16 01:52:02 +0000	<a hr r..
2345	666063827256086533	NaN	NaN	2015-11-16 01:22:45 +0000	<a hr r..
2346	666058600524156928	NaN	NaN	2015-11-16 01:01:59 +0000	<a hr r..
2347	666057090499244032	NaN	NaN	2015-11-16 00:55:59 +0000	<a hr r..
2348	666055525042405380	NaN	NaN	2015-11-16 00:49:46 +0000	<a hr r..
2349	666051853826850816	NaN	NaN	2015-11-16 00:35:11 +0000	<a hr r..
2350	666050758794694657	NaN	NaN	2015-11-16 00:30:50 +0000	<a hr r..
2351	666049248165822465	NaN	NaN	2015-11-16 00:24:50 +0000	<a hr r..
2352	666044226329800704	NaN	NaN	2015-11-16 00:04:52 +0000	<a hr r..
2353	666033412701032449	NaN	NaN	2015-11-15 23:21:54 +0000	<a hr r..
2354	666029285002620928	NaN	NaN	2015-11-15 23:05:30 +0000	<a hr r..
2355	666020888022790149	NaN	NaN	2015-11-15 22:32:08 +0000	<a hr r..

2356 rows × 17 columns

质量

df_tweet ***表格***

- favorite_count、id、retweet_count应该是int64类型，而不是浮点浮点类型
- favorited 和 retweeted 的值都是零，需要删除这两列

twitter_archive_enhanced ***表格***

- doggo floofer pupper puppo字段有"None"这是python关键字，在csv体现为字符串，这应该在表示空
- in_reply_to_status_id和in_reply_to_user_id，只有78个值，这两个列的数据很可能没用
- timestamp、retweeted_status_timestamp现在是object类型，要转换为时间类型
- rating_denominator不是10的需要清理掉,比如170是最大值是不对的
- rating_numerator分子小于rating_denominator的需要清理掉
- expanded_urls的单条记录中有重复的url地址，清理清理只保留一个
- name字段有字符串"None"，有的跟text中的描述不符，比如"a"

清洁度

df_tweet ***表格***

- source字段内容是html文本，需要更加简洁，只需要获取该html的text部分即可
- 通过witter_archive_enhanced.tweet_id和df_tweet.id和image_predictions.tweet_id关联创建新的表格 twitter_archive_master.csv

清理

In [87]:

```
# 清理前的备份
df_tweet_clean = df_tweet.copy()
witter_archive_enhanced_clean = witter_archive_enhanced.copy()
```

In [88]:

```
# favorite_count id retweet_count应该是int64类型，现在做格式转换 long
for i in ['favorite_count', 'id', 'retweet_count']:
    df_tweet_clean[i] = df_tweet_clean[i].astype(long)
df_tweet_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 7 columns):
favorite_count    2352 non-null int64
favorited         2352 non-null float64
full_text         2352 non-null object
id                2352 non-null int64
retweet_count     2352 non-null int64
retweeted         2352 non-null float64
source            2352 non-null object
dtypes: float64(2), int64(3), object(2)
memory usage: 128.7+ KB
```

In [89]:

```
# 顺便调整列的顺序, 排除favorited和retweeted列
index = ['id', 'retweet_count', 'favorite_count', 'full_text', 'source']
df_tweet_clean = df_tweet_clean.reindex_axis(index, axis=1)
df_tweet_clean
```

Out[89]:

	id	retweet_count	favorite_count	full_text
0	892420643555336192	8842	39492	This is Phineas. He's a mystical boy. Only eve...
1	892177421306343424	6480	33786	This is Tilly. She's just checking pup on you....
2	891815181378084864	4301	25445	This is Archie. He is a rare Norwegian Pouncin...
3	891689557279858688	8925	42863	This is Darla. She commenced a snooze mid meal...
4	891327558926688256	9721	41016	This is Franklin. He would like you to stop ca...
5	891087950875897856	3240	20548	Here we have a majestic great white breaching ...
6	890971913173991424	2142	12053	Meet Jax. He enjoys ice cream so much he gets ...
7	890729181411237888	19548	66596	When you watch your owner call another dog a g...
8	890609185150312448	4403	28187	This is Zoey. She doesn't want to be one of th...
9	890240255349198848	7684	32467	This is Cassie. She is a college pup. Studying...
10	890006608113172480	7584	31127	This is Koda. He is a South Australian decksha...
11	889880896479866880	5116	28208	This is Bruno. He is a service shark. Only get...
12	889665388333682688	8502	38745	Here's a puppo that seems to be on the fence a...
13	889638837579907072	4705	27633	This is Ted. He does his best. Sometimes that'...

	id	retweet_count	favorite_count	full_text
14	889531135344209920	2309	15329	This is Stuart. He's sporting his favorite fan...
15	889278841981685760	5635	25712	This is Oliver. You're witnessing one of his m...
16	888917238123831296	4681	29555	This is Jim. He found a fren. Taught him how t...
17	888804989199671296	4535	26021	This is Zeke. He has a new stick. Very proud o...
18	888554962724278272	3722	20267	This is Ralphus. He's powering up. Attempting ...
19	888078434458587136	3637	22144	This is Gerald. He was just told he didn't get...
20	887705289381826560	5584	30690	This is Jeffrey. He has a monopoly on the pool...
21	887517139158093824	12053	46940	I've yet to rate a Venezuelan Hover Wiener. Th...
22	887473957103951872	18813	70007	This is Canela. She attempted some fancy porch...
23	887343217045368832	10713	34223	You may not have known you needed to see this ...
24	887101392804085760	6147	31045	This... is a Jubilant Antarctic House Bear. We...
25	886983233522544640	8045	35786	This is Maya. She's very shy. Rarely leaves he...
26	886736880519319552	3420	12286	This is Mingus. He's a wonderful father to his...
27	886680336477933568	4597	22802	This is Derek. He's late for a dog meeting. 13...

	id	retweet_count	favorite_count	full_text
28	886366144734445568	3297	21488	This is Roscoe. Another pupper fallen victim t...
29	886267009285017600	4	117	@NonWhiteHat @MayhewMayhem omg hello tanner yo...
...	...	...	...	...
2322	666411507551481856	337	457	This is quite the dog. Gets really excited whe...
2323	666407126856765440	43	113	This is a southern Vesuvius bumblegruff. Can d...
2324	666396247373291520	91	171	Oh goodness. A super rare northeast Qdoba kang...
2325	666373753744588800	99	194	Those are sunglasses and a jean jacket. 11/10 ...
2326	666362758909284352	590	801	Unique dog here. Very small. Lives in containe...
2327	666353288456101888	76	228	Here we have a mixed Asiago from the Galápagos...
2328	666345417576210432	146	308	Look at this jokester thinking seat belt laws ...
2329	666337882303524864	96	203	This is an extremely rare horned Parthenon. No...
2330	666293911632134144	365	519	This is a funny dog. Weird toes. Won't come do...
2331	666287406224695296	71	152	This is an Albanian 3 1/2 legged Episcopalian...
2332	666273097616637952	81	183	Can take selfies 11/10 https://t.co/ws2AMaNwPW
2333	666268910803644416	37	108	Very concerned about fellow dog trapped in com...

	id	retweet_count	favorite_count	full_text
2334	666104133288665088	6835	14703	Not familiar with this breed. No tail (weird)....
2335	666102155909144576	15	81	Oh my. Here you are seeing an Adobe Setter giv...
2336	666099513787052032	73	160	Can stand on stump for what seems like a while...
2337	666094000022159360	78	168	This appears to be a Mongolian Presbyterian mi...
2338	666082916733198336	47	121	Here we have a well-established sunblockerspan...
2339	666073100786774016	173	334	Let's hope this flight isn't Malaysian (lol). ...
2340	666071193221509120	67	154	Here we have a northern speckled Rhododendron....
2341	666063827256086528	230	494	This is the happiest dog you will ever see. Ve...
2342	666058600524156928	61	117	Here is the Rand Paul of retrievers folks! He'...
2343	666057090499244032	146	304	My oh my. This is a rare blond Canadian terrie...
2344	666055525042405376	261	449	Here is a Siberian heavily armored polar bear ...
2345	666051853826850816	877	1250	This is an odd dog. Hard on the outside but lo...
2346	666050758794694656	60	136	This is a truly beautiful English Wilson Staff...
2347	666049248165822464	41	111	Here we have a 1949 1st generation vulpix. Enj...

	id	retweet_count	favorite_count	full_text
2348	666044226329800704	147	309	This is a purebred Piers Morgan. Loves to Netf...
2349	666033412701032448	47	128	Here is a very happy pup. Big fan of well-main...
2350	666029285002620928	48	132	This is a western brown Mitsubishi terrier. Up...
2351	666020888022790144	530	2528	Here we have a Japanese Irish Setter. Lost eye...

2352 rows × 5 columns

In [90]:

```
# rating_denominator不是10的需要清理掉
witter_archive_enhanced_clean = witter_archive_enhanced[witter_archive_enhanced["rating_denomina
tor"]==10]
witter_archive_enhanced_clean.info()
witter_archive_enhanced_clean
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2333 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2333 non-null int64
in_reply_to_status_id         73 non-null float64
in_reply_to_user_id           73 non-null float64
timestamp                     2333 non-null object
source                        2333 non-null object
text                          2333 non-null object
retweeted_status_id           180 non-null float64
retweeted_status_user_id      180 non-null float64
retweeted_status_timestamp    180 non-null object
expanded_urls                 2278 non-null object
rating_numerator              2333 non-null int64
rating_denominator            2333 non-null int64
name                          2333 non-null object
doggo                         2333 non-null object
floofer                       2333 non-null object
pupper                        2333 non-null object
puppo                         2333 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 328.1+ KB
```

Out[90]:

	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 hr r..
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a hr r..
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a hr r..
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a hr r..
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a hr r..
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a hr r..
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a hr r..
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40 +0000	<a hr r..
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51 +0000	<a hr r..
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51 +0000	<a hr r..
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53 +0000	<a hr r..
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32 +0000	<a hr r..
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02 +0000	<a hr r..
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04 +0000	<a hr r..
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32 +0000	<a hr r..
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39 +0000	<a hr r..
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37 +0000	<a hr r..
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06 +0000	<a hr r..
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36 +0000	<a hr r..
20	888078434458587136	NaN	NaN	2017-07-20 16:49:33 +0000	<a hr r..
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09 +0000	<a hr r..
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34 +0000	<a hr r..
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03 +0000	<a hr r..
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08 +0000	<a hr r..
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36 +0000	<a hr r..
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41 +0000	<a hr r..
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00 +0000	<a hr r..
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31 +0000	<a hr r..
...	...	...	...	...	...
2325	666418789513326592	NaN	NaN	2015-11-17 00:53:15 +0000	<a hr r..
2326	666411507551481857	NaN	NaN	2015-11-17 00:24:19 +0000	<a hr r..
2327	666407126856765440	NaN	NaN	2015-11-17 00:06:54 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2328	666396247373291520	NaN	NaN	2015-11-16 23:23:41 +0000	<a hr r..
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18 +0000	<a hr r..
2330	666362758909284353	NaN	NaN	2015-11-16 21:10:36 +0000	<a hr r..
2331	666353288456101888	NaN	NaN	2015-11-16 20:32:58 +0000	<a hr r..
2332	666345417576210432	NaN	NaN	2015-11-16 20:01:42 +0000	<a hr r..
2333	666337882303524864	NaN	NaN	2015-11-16 19:31:45 +0000	<a hr r..
2334	666293911632134144	NaN	NaN	2015-11-16 16:37:02 +0000	<a hr r..
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19 +0000	<a hr r..
2337	666268910803644416	NaN	NaN	2015-11-16 14:57:41 +0000	<a hr r..
2338	666104133288665088	NaN	NaN	2015-11-16 04:02:55 +0000	<a hr r..
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04 +0000	<a hr r..
2340	666099513787052032	NaN	NaN	2015-11-16 03:44:34 +0000	<a hr r..
2341	666094000022159362	NaN	NaN	2015-11-16 03:22:39 +0000	<a hr r..
2342	666082916733198337	NaN	NaN	2015-11-16 02:38:37 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2343	666073100786774016	NaN	NaN	2015-11-16 01:59:36 +0000	<a hr r..
2344	666071193221509120	NaN	NaN	2015-11-16 01:52:02 +0000	<a hr r..
2345	666063827256086533	NaN	NaN	2015-11-16 01:22:45 +0000	<a hr r..
2346	666058600524156928	NaN	NaN	2015-11-16 01:01:59 +0000	<a hr r..
2347	666057090499244032	NaN	NaN	2015-11-16 00:55:59 +0000	<a hr r..
2348	666055525042405380	NaN	NaN	2015-11-16 00:49:46 +0000	<a hr r..
2349	666051853826850816	NaN	NaN	2015-11-16 00:35:11 +0000	<a hr r..
2350	666050758794694657	NaN	NaN	2015-11-16 00:30:50 +0000	<a hr r..
2351	666049248165822465	NaN	NaN	2015-11-16 00:24:50 +0000	<a hr r..
2352	666044226329800704	NaN	NaN	2015-11-16 00:04:52 +0000	<a hr r..
2353	666033412701032449	NaN	NaN	2015-11-15 23:21:54 +0000	<a hr r..
2354	666029285002620928	NaN	NaN	2015-11-15 23:05:30 +0000	<a hr r..
2355	666020888022790149	NaN	NaN	2015-11-15 22:32:08 +0000	<a hr r..

2333 rows × 17 columns

In [91]:

```
# rating_numerator分子小于rating_denominator的需要清理掉
witter_archive_enhanced_clean = witter_archive_enhanced_clean[witter_archive_enhanced_clean["rat
ing_numerator"] > witter_archive_enhanced_clean["rating_denominator"]]
witter_archive_enhanced_clean.info()
witter_archive_enhanced_clean
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1438 entries, 0 to 2339
Data columns (total 17 columns):
tweet_id                      1438 non-null int64
in_reply_to_status_id         56 non-null float64
in_reply_to_user_id           56 non-null float64
timestamp                     1438 non-null object
source                        1438 non-null object
text                          1438 non-null object
retweeted_status_id           153 non-null float64
retweeted_status_user_id      153 non-null float64
retweeted_status_timestamp    153 non-null object
expanded_urls                 1398 non-null object
rating_numerator              1438 non-null int64
rating_denominator            1438 non-null int64
name                          1438 non-null object
doggo                         1438 non-null object
floofer                       1438 non-null object
pupper                        1438 non-null object
puppo                         1438 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 202.2+ KB
```

Out[91]:

	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 hr r..
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a hr r..
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a hr r..
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a hr r..
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a hr r..
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a hr r..
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a hr r..
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40 +0000	<a hr r..
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51 +0000	<a hr r..
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51 +0000	<a hr r..
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53 +0000	<a hr r..
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32 +0000	<a hr r..
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02 +0000	<a hr r..
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04 +0000	<a hr r..
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32 +0000	<a hr r..
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39 +0000	<a hr r..
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37 +0000	<a hr r..
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06 +0000	<a hr r..
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36 +0000	<a hr r..
20	888078434458587136	NaN	NaN	2017-07-20 16:49:33 +0000	<a hr r..
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09 +0000	<a hr r..
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34 +0000	<a hr r..
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03 +0000	<a hr r..
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08 +0000	<a hr r..
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36 +0000	<a hr r..
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41 +0000	<a hr r..
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00 +0000	<a hr r..
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31 +0000	<a hr r..
...	...	...	...	...	...
2201	668645506898350081	NaN	NaN	2015-11-23 04:21:26 +0000	<a hr r..
2212	668587383441514497	NaN	NaN	2015-11-23 00:30:28 +0000	<a re
2213	668567822092664832	NaN	NaN	2015-11-22 23:12:44 +0000	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2217	668528771708952576	NaN	NaN	2015-11-22 20:37:34 +0000	<a hr r..
2221	668480044826800133	NaN	NaN	2015-11-22 17:23:57 +0000	<a hr r..
2225	668286279830867968	NaN	NaN	2015-11-22 04:33:59 +0000	<a hr r..
2228	668256321989451776	NaN	NaN	2015-11-22 02:34:57 +0000	<a hr r..
2234	668190681446379520	NaN	NaN	2015-11-21 22:14:07 +0000	<a hr r..
2244	667886921285246976	NaN	NaN	2015-11-21 02:07:05 +0000	<a hr r..
2250	667832474953625600	NaN	NaN	2015-11-20 22:30:44 +0000	<a hr r..
2252	667801013445750784	NaN	NaN	2015-11-20 20:25:43 +0000	<a hr r..
2257	667728196545200128	NaN	NaN	2015-11-20 15:36:22 +0000	<a re
2259	667550904950915073	NaN	NaN	2015-11-20 03:51:52 +0000	<a re
2267	667524857454854144	NaN	NaN	2015-11-20 02:08:22 +0000	<a re
2269	667509364010450944	NaN	NaN	2015-11-20 01:06:48 +0000	<a re
2270	667502640335572993	NaN	NaN	2015-11-20 00:40:05 +0000	<a re
2273	667470559035432960	NaN	NaN	2015-11-19 22:32:36 +0000	<a re

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2275	667453023279554560	NaN	NaN	2015-11-19 21:22:56 +0000	<a re
2278	667435689202614272	NaN	NaN	2015-11-19 20:14:03 +0000	<a hr r..
2283	667200525029539841	NaN	NaN	2015-11-19 04:39:35 +0000	<a hr r..
2284	667192066997374976	NaN	NaN	2015-11-19 04:05:59 +0000	<a hr r..
2292	667160273090932737	NaN	NaN	2015-11-19 01:59:39 +0000	<a hr r..
2293	667152164079423490	NaN	NaN	2015-11-19 01:27:25 +0000	<a hr r..
2301	667044094246576128	NaN	NaN	2015-11-18 18:17:59 +0000	<a hr r..
2304	666983947667116034	NaN	NaN	2015-11-18 14:18:59 +0000	<a hr r..
2307	666826780179869698	NaN	NaN	2015-11-18 03:54:28 +0000	<a hr r..
2324	666421158376562688	NaN	NaN	2015-11-17 01:02:40 +0000	<a hr r..
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18 +0000	<a hr r..
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19 +0000	<a hr r..
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04 +0000	<a hr r..

1438 rows × 17 columns

In [92]:

```
witter_archive_enhanced_clean.name.value_counts()
```

Out[92]:

```
None        437
a            17
Charlie      10
Oliver        9
Tucker        9
Cooper        9
Bo            9
Lucy          7
Bailey        6
Daisy         6
Koda          6
Jack          6
Lola          6
Penny         6
Rusty         5
Winston       5
Buddy         5
Leo           5
Scout         5
Loki          4
Carl          4
Maddie        4
Luna          4
Dave          4
Sunny         4
Oscar         4
Louis         4
Alfie         4
Phil          4
Maximus       4
             ...
Brutus        1
Newt          1
Pilot         1
Carper        1
Blue          1
Miley         1
Tuck          1
Bayley        1
Opal          1
Alice         1
Thor          1
Molly         1
Skittle       1
Puff          1
Gabe          1
Robin         1
Tripp         1
Calbert       1
Arlen         1
Elliot        1
Griswold      1
Tebow         1
Mookie        1
Spanky        1
Goose         1
Sailor        1
Mya           1
Liam          1
DonDon        1
```

```
Bonbon          1

Lacy            1
Name: name, Length: 639, dtype: int64
```

In [93]:

```
# timestamp、retweeted_status_timestamp现在是object类型，要转换为时间类型
witter_archive_enhanced_clean['timestamp'] = pd.to_datetime(witter_archive_enhanced_clean['times
tamp'],format='%Y-%m-%d %H:%M:%S')
```

```
C:\Anaconda2\lib\site-packages\ipykernel\__main__.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/
indexing.html#indexing-view-versus-copy
  from ipykernel import kernelapp as app
```

In [94]:

```
witter_archive_enhanced_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1438 entries, 0 to 2339
Data columns (total 17 columns):
tweet_id                    1438 non-null int64
in_reply_to_status_id       56 non-null float64
in_reply_to_user_id         56 non-null float64
timestamp                   1438 non-null datetime64[ns]
source                      1438 non-null object
text                        1438 non-null object
retweeted_status_id         153 non-null float64
retweeted_status_user_id    153 non-null float64
retweeted_status_timestamp  153 non-null object
expanded_urls               1398 non-null object
rating_numerator            1438 non-null int64
rating_denominator          1438 non-null int64
name                        1438 non-null object
doggo                       1438 non-null object
floofer                     1438 non-null object
pupper                      1438 non-null object
puppo                       1438 non-null object
dtypes: datetime64[ns](1), float64(4), int64(3), object(9)
memory usage: 202.2+ KB
```

In [95]:

```
# expanded_urls的单条记录中有重复的url地址，清理清理只保留一个
witter_archive_enhanced_clean['expanded_urls_one'], witter_archive_enhanced_clean['expanded_urls
_others'] = witter_archive_enhanced_clean['expanded_urls'].str.split(',', 1).str
witter_archive_enhanced_clean
```

```
C:\Anaconda2\lib\site-packages\ipykernel\__main__.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/
indexing.html#indexing-view-versus-copy
  from ipykernel import kernelapp as app
```

Out[95]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56	<a hr r..
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27	<a hr r..
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03	<a hr r..
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51	<a hr r..
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24	<a hr r..
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17	<a hr r..
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12	<a hr r..
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40	<a hr r..
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51	<a hr r..
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51	<a hr r..
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25	<a hr r..
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53	<a hr r..
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32	<a hr r..
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04	<a hr r..
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32	<a hr r..
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39	<a hr r..
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37	<a hr r..
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06	<a hr r..
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36	<a hr r..
20	887078434458587136	NaN	NaN	2017-07-20 16:49:33	<a hr r..
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48	<a hr r..
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09	<a hr r..
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34	<a hr r..
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03	<a hr r..
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08	<a hr r..
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36	<a hr r..
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00	<a hr r..
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31	<a hr r..
...	...	...	...	...	...
2201	668645506898350081	NaN	NaN	2015-11-23 04:21:26	<a hr r..
2212	668587383441514497	NaN	NaN	2015-11-23 00:30:28	<a re
2213	668567822092664832	NaN	NaN	2015-11-22 23:12:44	<a hr r..
2217	668528771708952576	NaN	NaN	2015-11-22 20:37:34	<a hr r..
2221	668480044826800133	NaN	NaN	2015-11-22 17:23:57	<a hr r..
2225	668286279830867968	NaN	NaN	2015-11-22 04:33:59	<a hr r..
2228	668256321989451776	NaN	NaN	2015-11-22 02:34:57	<a hr r..
2234	668190681446379520	NaN	NaN	2015-11-21 22:14:07	<a hr r..
2244	667886921285246976	NaN	NaN	2015-11-21 02:07:05	<a hr r..
2250	667832474953625600	NaN	NaN	2015-11-20 22:30:44	<a hr r..
2252	667801013445750784	NaN	NaN	2015-11-20 20:25:43	<a hr r..
2257	667728196545200128	NaN	NaN	2015-11-20 15:36:22	<a re
2259	667550904950915073	NaN	NaN	2015-11-20	<a

				03:51:52	re

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2267	667524857454854144	NaN	NaN	2015-11-20 02:08:22	<a re
2269	667509364010450944	NaN	NaN	2015-11-20 01:06:48	<a re
2270	667502640335572993	NaN	NaN	2015-11-20 00:40:05	<a re
2273	667470559035432960	NaN	NaN	2015-11-19 22:32:36	<a re
2275	667453023279554560	NaN	NaN	2015-11-19 21:22:56	<a re
2278	667435689202614272	NaN	NaN	2015-11-19 20:14:03	<a hr r..
2283	667200525029539841	NaN	NaN	2015-11-19 04:39:35	<a hr r..
2284	667192066997374976	NaN	NaN	2015-11-19 04:05:59	<a hr r..
2292	667160273090932737	NaN	NaN	2015-11-19 01:59:39	<a hr r..
2293	667152164079423490	NaN	NaN	2015-11-19 01:27:25	<a hr r..
2301	667044094246576128	NaN	NaN	2015-11-18 18:17:59	<a hr r..
2304	666983947667116034	NaN	NaN	2015-11-18 14:18:59	<a hr r..
2307	666826780179869698	NaN	NaN	2015-11-18 03:54:28	<a hr r..
2324	666421158376562688	NaN	NaN	2015-11-17 01:02:40	<a hr r..
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19	<a hr r..
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04	<a hr r..

1438 rows × 19 columns

In [96]:

```
# source字段内容是html文本，需要更加简洁，只需要获取该html的text部分即可
witter_archive_enhanced_clean['source1'], witter_archive_enhanced_clean['source2']= witter_archi
ve_enhanced_clean.source.str.split('>', 1).str
witter_archive_enhanced_clean['source_text'], witter_archive_enhanced_clean['source4']= witter_a
rchive_enhanced_clean.source2.str.split('<', 1).str
witter_archive_enhanced_clean = witter_archive_enhanced_clean.drop('source1', axis=1)
witter_archive_enhanced_clean = witter_archive_enhanced_clean.drop('source2', axis=1)
witter_archive_enhanced_clean = witter_archive_enhanced_clean.drop('source4', axis=1)
witter_archive_enhanced_clean
```

```
C:\Anaconda2\lib\site-packages\ipykernel\__main__.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/
indexing.html#indexing-view-versus-copy
  from ipykernel import kernelapp as app
C:\Anaconda2\lib\site-packages\ipykernel\__main__.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/
indexing.html#indexing-view-versus-copy
  app.launch_new_instance()
```

Out[96]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56	<a hr r..
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27	<a hr r..
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03	<a hr r..
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51	<a hr r..
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24	<a hr r..
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17	<a hr r..
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12	<a hr r..
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40	<a hr r..
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51	<a hr r..
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51	<a hr r..
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25	<a hr r..
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53	<a hr r..
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32	<a hr r..
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04	<a hr r..
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32	<a hr r..
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39	<a hr r..
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37	<a hr r..
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06	<a hr r..
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36	<a hr r..
20	887078434458587136	NaN	NaN	2017-07-20 16:49:33	<a hr r..
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48	<a hr r..
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09	<a hr r..
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34	<a hr r..
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03	<a hr r..
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08	<a hr r..
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36	<a hr r..
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00	<a hr r..
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31	<a hr r..
...	...	...	...	...	...
2201	668645506898350081	NaN	NaN	2015-11-23 04:21:26	<a hr r..
2212	668587383441514497	NaN	NaN	2015-11-23 00:30:28	<a re
2213	668567822092664832	NaN	NaN	2015-11-22 23:12:44	<a hr r..
2217	668528771708952576	NaN	NaN	2015-11-22 20:37:34	<a hr r..
2221	668480044826800133	NaN	NaN	2015-11-22 17:23:57	<a hr r..
2225	668286279830867968	NaN	NaN	2015-11-22 04:33:59	<a hr r..
2228	668256321989451776	NaN	NaN	2015-11-22 02:34:57	<a hr r..
2234	668190681446379520	NaN	NaN	2015-11-21 22:14:07	<a hr r..
2244	667886921285246976	NaN	NaN	2015-11-21 02:07:05	<a hr r..
2250	667832474953625600	NaN	NaN	2015-11-20 22:30:44	<a hr r..
2252	667801013445750784	NaN	NaN	2015-11-20 20:25:43	<a hr r..
2257	667728196545200128	NaN	NaN	2015-11-20 15:36:22	<a re
2259	667550904950915073	NaN	NaN	2015-11-20	<a

				03:51:52	re

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2267	667524857454854144	NaN	NaN	2015-11-20 02:08:22	<a re
2269	667509364010450944	NaN	NaN	2015-11-20 01:06:48	<a re
2270	667502640335572993	NaN	NaN	2015-11-20 00:40:05	<a re
2273	667470559035432960	NaN	NaN	2015-11-19 22:32:36	<a re
2275	667453023279554560	NaN	NaN	2015-11-19 21:22:56	<a re
2278	667435689202614272	NaN	NaN	2015-11-19 20:14:03	<a hr r..
2283	667200525029539841	NaN	NaN	2015-11-19 04:39:35	<a hr r..
2284	667192066997374976	NaN	NaN	2015-11-19 04:05:59	<a hr r..
2292	667160273090932737	NaN	NaN	2015-11-19 01:59:39	<a hr r..
2293	667152164079423490	NaN	NaN	2015-11-19 01:27:25	<a hr r..
2301	667044094246576128	NaN	NaN	2015-11-18 18:17:59	<a hr r..
2304	666983947667116034	NaN	NaN	2015-11-18 14:18:59	<a hr r..
2307	666826780179869698	NaN	NaN	2015-11-18 03:54:28	<a hr r..
2324	666421158376562688	NaN	NaN	2015-11-17 01:02:40	<a hr r..
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18	<a hr r..

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19	<a hr r..
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04	<a hr r..

1438 rows × 20 columns

In [97]:

```
# 查看source的分布
witter_archive_enhanced_clean.source_text.value_counts()
```

Out[97]:

```
Twitter for iPhone     1342
Vine - Make a Scene      71
Twitter Web Client       17
TweetDeck                 8
Name: source_text, dtype: int64
```

In [98]:

```
# 修改df_tweet_clean的 id 为 tweet_id
df_tweet_clean['tweet_id'] = df_tweet_clean['id']
# 只保留部分列
index = ['tweet_id', 'retweet_count', 'favorite_count']
df_tweet_clean = df_tweet_clean.reindex_axis(index, axis=1)
index = ['tweet_id', 'name','rating_numerator','rating_denominator','timestamp', 'text', 'source
_text', 'expanded_urls_one','doggo','floofer','pupper','puppo']
witter_archive_enhanced_clean = witter_archive_enhanced_clean.reindex_axis(index, axis=1)
```

In [99]:

```
# 通过 witter_archive_enhanced.tweet_id 和 df_tweet.id关联创建新的表格twitter_archive_master.csv
twitter_archive_master = pd.merge(witter_archive_enhanced_clean, df_tweet_clean,
                          on=['tweet_id'], how='left')
twitter_archive_master.info()
twitter_archive_master
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1438 entries, 0 to 1437
Data columns (total 14 columns):
tweet_id              1438 non-null int64
name                  1438 non-null object
rating_numerator      1438 non-null int64
rating_denominator    1438 non-null int64
timestamp             1438 non-null datetime64[ns]
text                  1438 non-null object
source_text           1438 non-null object
expanded_urls_one     1398 non-null object
doggo                 1438 non-null object
floofer               1438 non-null object
pupper                1438 non-null object
puppo                 1438 non-null object
retweet_count         890 non-null float64
favorite_count        890 non-null float64
dtypes: datetime64[ns](1), float64(2), int64(3), object(8)
memory usage: 168.5+ KB
```

Out[99]:

	tweet_id	name	rating_numerator	rating_denominator	timestar
0	892420643555336193	Phineas	13	10	2017-08-01 16:23:56
1	892177421306343426	Tilly	13	10	2017-08-01 00:17:27
2	891815181378084864	Archie	12	10	2017-07-31 00:18:03
3	891689557279858688	Darla	13	10	2017-07-30 15:58:51
4	891327558926688256	Franklin	12	10	2017-07-29 16:00:24
5	891087950875897856	None	13	10	2017-07-29 00:08:17
6	890971913173991426	Jax	13	10	2017-07-28 16:27:12
7	890729181411237888	None	13	10	2017-07-28 00:22:40
8	890609185150312448	Zoey	13	10	2017-07-27 16:25:51
9	890240255349198849	Cassie	14	10	2017-07-26 15:59:51
10	890006608113172480	Koda	13	10	2017-07-26 00:31:25
11	889880896479866881	Bruno	13	10	2017-07-25 16:11:53
12	889665388333682689	None	13	10	2017-07-25 01:55:32
13	889638837579907072	Ted	12	10	2017-07-25 00:10:02

	tweet_id	name	rating_numerator	rating_denominator	timestar
14	889531135344209921	Stuart	13	10	2017-07-24 17:02:04
15	889278841981685760	Oliver	13	10	2017-07-24 00:19:32
16	888917238123831296	Jim	12	10	2017-07-23 00:22:39
17	888804989199671297	Zeke	13	10	2017-07-22 16:56:37
18	888554962724278272	Ralphus	13	10	2017-07-22 00:23:06
19	888202515573088257	Canela	13	10	2017-07-21 01:02:36
20	888078434458587136	Gerald	12	10	2017-07-20 16:49:33
21	887705289381826560	Jeffrey	13	10	2017-07-19 16:06:48
22	887517139158093824	such	14	10	2017-07-19 03:39:09
23	887473957103951883	Canela	13	10	2017-07-19 00:47:34
24	887343217045368832	None	13	10	2017-07-18 16:08:03
25	887101392804085760	None	12	10	2017-07-18 00:07:08
26	886983233522544640	Maya	13	10	2017-07-17 16:17:36
27	886736880519319552	Mingus	13	10	2017-07-16 23:58:41

	tweet_id	name	rating_numerator	rating_denominator	timestar
28	886680336477933568	Derek	13	10	2017-07-16 20:14:00
29	886366144734445568	Roscoe	12	10	2017-07-15 23:25:31
...	...	...	...	...	...
1408	668645506898350081	None	11	10	2015-11-04:21:26
1409	668587383441514497	the	13	10	2015-11-00:30:28
1410	668567822092664832	Marvin	11	10	2015-11-23:12:44
1411	668528771708952576	Gòrdón	12	10	2015-11-20:37:34
1412	668480044826800133	DayZ	11	10	2015-11-17:23:57
1413	668286279830867968	Rusty	11	10	2015-11-04:33:59
1414	668256321989451776	Jareld	13	10	2015-11-02:34:57
1415	668190681446379520	Skittles	12	10	2015-11-22:14:07
1416	667886921285246976	Erik	11	10	2015-11-02:07:05
1417	667832474953625600	None	12	10	2015-11-22:30:44
1418	667801013445750784	None	12	10	2015-11-20:25:43
1419	667728196545200128	Olive	11	10	2015-11-15:36:22
1420	667550904950915073	None	12	10	2015-11-03:51:52
1421	667524857454854144	None	12	10	2015-11-02:08:22
1422	667509364010450944	None	12	10	2015-11-01:06:48
1423	667502640335572993	Hall	11	10	2015-11-

	[illegible]	Hall	11	10	00:40:05

	tweet_id	name	rating_numerator	rating_denominator	timestar
1424	667470559035432960	a	11	10	2015-11-22:32:36
1425	667453023279554560	Cupcake	11	10	2015-11-21:22:56
1426	667435689202614272	None	12	10	2015-11-20:14:03
1427	667200525029539841	Joshwa	11	10	2015-11-04:39:35
1428	667192066997374976	None	12	10	2015-11-04:05:59
1429	667160273090932737	Bradlay	11	10	2015-11-01:59:39
1430	667152164079423490	Pipsy	12	10	2015-11-01:27:25
1431	667044094246576128	None	12	10	2015-11-18:17:59
1432	666983947667116034	a	11	10	2015-11-14:18:59
1433	666826780179869698	None	12	10	2015-11-03:54:28
1434	666421158376562688	None	12	10	2015-11-01:02:40
1435	666373753744588802	None	11	10	2015-11-21:54:18
1436	666273097616637952	None	11	10	2015-11-15:14:19
1437	666102155909144576	None	11	10	2015-11-03:55:04

1438 rows × 14 columns

In [100]:

```
# 把数据保存到 twitter_archive_master.csv
twitter_archive_master.to_csv('twitter_archive_master.csv', index=False)
```

In [ ]:

```

```