ITMO

spam comment detection with data collection from flickr

Ma Cheng Yuan
372818

Statement

This project is a showcase to the study in this semester , it is incomplete due to some technical problems , so dataset does not meet the criteria of “big data”

Reason :
i follow instruction in pyspark website to set the environment in jupyter notebook , but udf can not find downloaded library
So i use pytorch to complete this practice . It does not show problem when i use small datasets

Problem 1 : In jupyter , i am unable to load flickrapi into worker executor environment
Problem 2 : using environment in solution 1 , and tried to set worker executor environment and main environment to same environment , but still had problem when i used udf

project intro

ITMO

Inappropriate text has been a problem for several years since the popularity of social media, and the end result is detecting inappropriate comments with emotion.

website intro

photo-sharing Web site owned by SmugMug and headquartered in San Francisco, California. Flickr is an **ad-supported service, free to the general public**, that allows users to **upload digital photographs** from their own computers and **share them online** with either private groups or the world at large

tool intro

flickrapi - retrival data

pyspark - store , process

transformers - emotion detection

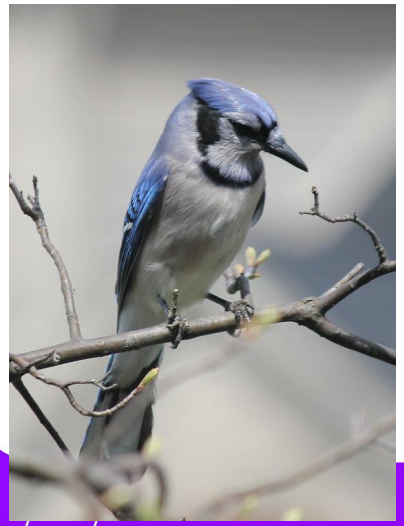
pytorch - model implementation

datasets intro

category : birds

features : color

blue jay (blue) , northern cardinal(red) , american goldfinch(yellow)



data retrival

library : flickrapi

```
API_KEY = '513e24d8fc90dd4f767670d583ba1e41'
API_SECRET = '975b054fcd3c3371'
api = flickrapi.FlickrAPI(api_key=API_KEY, secret=API_SECRET, format='parsed-json')
```

```
photos = api.photos.search(text=label, sort='relevance', media='photos',
                           extras='url_c', per_page=size, page=page)['photos']['photo']
```

	id	owner	secret	server	farm	title	ispublic	isfriend	isfamily	
0	14143181441	75378295@N07	ffdc52e7c5	5546	6	Blue Jay	1	0	0	https://live.staticflickr.com/5546/1414
1	22506516545	51102294@N05	fe40197b38	715	1	Blue jay	1	0	0	https://live.staticflickr.com/715/225065

requirement can be customized : image size , media type , location

transformation

library : transformation

a pre-trained model to detect six different emotion from text :
joy , love , anger , surprise , sadness , fear

data process

library : pyspark.pandas

Here are result with 30000 rows of data (authors with most photos)

attention ranking by comments

owner	count
38802090@N06	930
75709980@N08	922
88978913@N08	487
50177794@N07	469
78998859@N04	460
25643797@N03	458
49114357@N08	421
122545282@N04	420
10226995@N02	248
94330142@N08	238
129376752@N03	224
65166049@N06	217
38635198@N04	213
79813986@N02	206
129524245@N08	187
126304632@N05	173
93549366@N06	169
58737825@N07	154
38487871@N05	154
51433704@N06	150

only showing top 20 rows

comparison by owner_id - blue jay :

owner	count
50177794@N07	468
25643797@N03	451
61768576@N04	80
9773589@N08	73
36678894@N06	67
100278916@N04	60
94330142@N08	58
75558468@N07	57
46610212@N07	55
39871363@N04	55

only showing top 10 rows

comparison by owner_id - northern cardinal :

owner	count
38802090@N06	913
75709980@N08	727
78998859@N04	431
49114357@N08	251
65166049@N06	206
38635198@N04	157
129376752@N03	148
129524245@N08	146
38487871@N05	143
58737825@N07	142

only showing top 10 rows

comparison by owner_id - american goldfinch :

owner	count
88978913@N08	485
122545282@N04	413
75709980@N08	195
49114357@N08	156
65148649@N04	132
52919773@N08	124
143249210@N06	113
92361032@N05	110
51433704@N06	106
126304632@N05	102

only showing top 10 rows

Warning

from here

i am only able to show example with few datasets

Result

here is the dataframe after query for preference topic and top3_fan of each author , authors duplicate because the main column here is photo id

label	comments	authors	sentiment	comment_size	top	id	prefer_topic3	id	top3_fan
2	[The perfect pose...	[40724294@N04, 60...	joy	4	3	110466196@N07	[House finch, Ame...	110466196@N07	[40724294@N04, 15...
2	[The perfect pose...	[40724294@N04, 60...	joy	4	3	110466196@N07	[House finch, Ame...	110466196@N07	[16159474@N00, 62...
2	[Cracking shot Jo...	[157646645@N07, 4...	joy	2	2	110466196@N07	[House finch, Ame...	110466196@N07	[40724294@N04, 15...
2	[Cracking shot Jo...	[157646645@N07, 4...	joy	2	2	110466196@N07	[House finch, Ame...	110466196@N07	[16159474@N00, 62...
2	[The perfect pose...	[40724294@N04, 60...	joy	4	3	110466196@N07	[House finch, Ame...	110466196@N07	[40724294@N04, 15...
2	[The perfect pose...	[40724294@N04, 60...	joy	4	3	110466196@N07	[House finch, Ame...	110466196@N07	[16159474@N00, 62...
2	[Cracking shot Jo...	[157646645@N07, 4...	joy	2	2	110466196@N07	[House finch, Ame...	110466196@N07	[40724294@N04, 15...
2	[Cracking shot Jo...	[157646645@N07, 4...	joy	2	2	110466196@N07	[House finch, Ame...	110466196@N07	[16159474@N00, 62...
0	null	null	null	0	3	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[21838055@N08, 85...
0	null	null	null	0	3	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[8152356@N08, 503...
0	null	null	null	0	2	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[21838055@N08, 85...
0	null	null	null	0	2	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[8152356@N08, 503...
0	null	null	null	0	3	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[21838055@N08, 85...
0	null	null	null	0	3	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[8152356@N08, 503...
0	null	null	null	0	2	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[21838055@N08, 85...
0	null	null	null	0	2	24218402@N06	[IMG_3082, ST. Th...	24218402@N06	[8152356@N08, 503...
2	[Beautiful shot, ...	[81938471@N00, 35...	joy	71	4	33900425@N03	[Menges Fameflowe...	33900425@N03	[45445559@N04, 95...
1	[Wow, Nicole, sup...	[129060298@N06, 1...	joy	21	5	38802090@N06	[Étourneau sanson...	38802090@N06	[38487871@N05, 12...
1	[Wow, Nicole, sup...	[129060298@N06, 1...	joy	21	5	38802090@N06	[Étourneau sanson...	38802090@N06	[38487871@N05, 12...
1	[Wow, Nicole, sup...	[129060298@N06, 1...	joy	21	5	38802090@N06	[Étourneau sanson...	38802090@N06	[38487871@N05, 12...

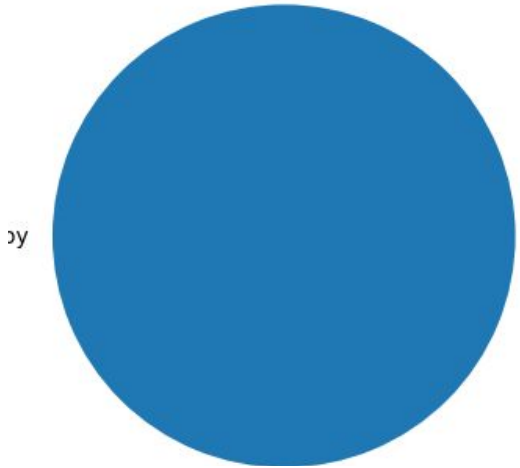
emotion proportion

İTMO

sentiment	size
joy	2
null	0

sentiment	size
joy	5

sentiment	size
joy	4
null	0



spam comment

criteria is when there is one emotion reach 80 % , others are spam

shit_comment	shit_author	sentiment	id
Lovely colours in...	118948217@N06	love	22847153945
Lovely bluejay ca...	46353871@N00	love	22847153945
Cracker!	9750464@N02	anger	22847153945
I love these bird...	55032983@N07	love	22847153945
Lovely shot of a ...	124011531@N04	love	22847153945
...beautiful imag...	38741307@N07	love	22847153945
Lovely ....nice BG	7522188@N02	love	22847153945
I wish they would...	127727047@N05	sadness	22847153945
Nice	33856622@N07	anger	22847153945
Magnifique capture	125881398@N07	anger	22847153945
Muito lindo..	65548569@N07	anger	22847153945
Mooi hoor!!	99745284@N00	anger	22847153945
Amazing capture S...	75715068@N02	surprise	22847153945
<img src="https:/...	45762667@N08	anger	22847153945
Lovely capture !	52614128@N07	love	22847153945
Lovely ... Wonder...	128236239@N05	love	22847153945
Muito lindo	136735947@N04	anger	22847153945
Lovely blue jay!	125716890@N02	love	22847153945
Mordecai :)	124415191@N06	anger	22847153945
Lovely Blue Jay, ...	105765827@N03	love	22847153945

conclusion

after this practice , i realized the powerful part of parallel computing from pyspark compared to pandas library . In return , there is complex setting of environment , moreover any transformation can influence the performance of computation .

Further improvement , i would like to use the proper environment using multiple workers with massive datasets to complete this project .

THANK YOU FOR YOUR TIME!

ITMO more than a UNIVERSITY