Forum - Students and Posting Time

Course: 大数据软件工程

Exercise type: Hadoop Python Streaming

Soft Due: 2015-06-01 18:00

Hard Due: 2015-06-01 18:00

Note: hard due has passed. You can no longer submit answer.

Author: 王欣明(TEACHER)

Email: wangxm35@mail.sysu.edu.cn

你已经接近或者达到满分。在完成这道题后，如果愿意的话，请你在下面**Feedback**那里评价一下这道题。你的反馈将同时用邮件发送给作者。

Description:

0 Questions and 0 Answers | Ask New Question

In this project you will work with discussion forum (also sometimes called discussion board) data. It is one type of user generated content that you can find all around the web. Most popular websites have some kind of a forum, and the things you will do in this project can transfer to other similar projects.

This particular dataset is taken from an online forum similar to the popular StackOverflow forum. The basic structure is - the forum has nodes. All nodes have a body and author_id. Top level nodes are called questions, and will also have a title and tags. Questions can have answers. Both questions and answers can have comments. If you are not sure how that all looks, please go to StackOverflow and look around!

There are 2 files in the dataset. The first is "forum_nodes.tsv", and that contains all forum questions and answers in one table. It was exported from the RDBMS by using tab as a separator, and enclosing all fields in double quotes. You can find the field names in the first line of the file "forum_node.tsv". The ones that are the most relevant to the task are:

- "id": id of the node
- "title": title of the node. in case "node_type" is "answer" or "comment", this field will be empty
- "tagnames": space separated list of tags
- "author_id": id of the author
- "body": content of the post
- "node_type": type of the node, either "question", "answer" or "comment"
- "parent_id": node under which the post is located, will be empty for "questions"
- "abs_parent_id": top node where the post is located
- "added_at": date added

The second table is "forum_users.tsv". It contains fields for "user_ptr_id" - the id of the user. "Reputation" - the reputation, or karma of the user, earned when other users upvote their posts, and the number of "gold", "silver" and "bronze" badges earned. The actual database has more fields in this table, like user name nickname, bio (if set) etc, but we have removed this information here.

Task1: Students and Posting Time

We have a lot of passionate students that bring a lot of value to forums. Forums also sometimes need a watchful eye on them, to make sure that posts are tagged in a way that helps to find them, that the tone on forums stays positive, and in general - they need people who can perform some management tasks - forum moderators. These are usually chosen from students who already have shown that they are active and helpful forum participants.

Our students come from all around the world, so we need to know both at what times of day the activity is the highest, and to know which of the students are active at that time.

In this exercise your task is to find for each student: what is the hour during which the student has posted the most posts. Output from reducers should be:

```
author_id    hour
```

For example:

```
13431511\t13
```

```
54525254141\t21
```

If there is a tie: there are multiple hours during which a student has posted a maximum number of posts, please print the student-hour pairs on separate lines. The order in which these lines appear in your output does not matter.

You can ignore the time-zone offset for all times - for example in the following line: "2012-02-25 08:11:01.623548+00" - you can ignore the +00 offset.

Note: In order to find the hour posted, please use the added_at field and NOT the last_activity_at field.

To make sure your code is running properly, we have put together a smaller data set and set of expected outputs for you to use to check your work. The name of the test data set is student_test_posts.csv.

Run the following command to display your code’s output: $ cat student_test_posts.csv | python mapper.py | sort | python reducer.py

Below you will find the output expected for this exercise when using the test data set provided. The output of your code should include all of the rows below, aside from the columns headers, but the order of the rows may be switched around.

```
Student ID |  Hour

100000005   1

100000066   1

100000066   5

100002460   12

100003192   8

100003268   15

100004467   12

100004467   20

100004819   4

100005156   17

100007808   12

100008254   22

100010128   14

100012200   5

100019875   5

100020526   14

100071170   12
```

Hint:

Hint is not available for this exercise.

Time limit:

Hard time limitation: 60 seconds

Standard answer spent time: 28 seconds

Your answer:

Already pass (hard) due date: June 1, 2015, 6 p.m. You can no longer submit your code to smart_programmer by yourself.

We don't accept late submission in principle. But if you really have a convincing reason, send the answer code to your TA by email and ask her/him to submit it for you. Don't copy the standard answer. It will be considered as plagirism.

If you want to study the standard answer, you can add it as an optional assignment.

Add this exercise as an optional assignment

Hard due has passed. The standard answer is unlocked:

[*]mapper1.py | combiner1.py | [*]reducer1.py | mapper2.py | combiner2.py | reducer2.py | mapper3.py | combiner3.py | reducer3.py

```
#!/usr/bin/python
import sys
```

```

title1 = 'id\ttitle\ttagnames\tauthor_id\tbody\tnode_type\tparent_id\tabs_parent_id\tadded_at\tscore\tstate_string\tlast_
       = '\"id\"\t\"title\"\t\"tagnames\"\t\"author_id\"\t\"body\"\t\"node_type\"\t\"parent_id\"\t\"abs_parent_id\"\t\"ad

def         (    ):
    if str_.startswith("\"") and str_.endswith("\""):
        return     [1:-1]
    else:
        return

def      (    ):
    stuInfo = line.replace("\n"," ").split("\t")
    if    (       ) == 19:
        autoId = getVelue(stuInfo[3])
             =         (       [8])
        hourInfo = time.split(" ")
        if    (        ) == 2:
            hour = hourInfo[1].split(":")[0]
            print "{0}\t{1}".      (      ,    (    ))

     = True
currLine = None
for      in     .     :
    if line == title1 or line == title2:
        continue

          =     .     ().     ("\t")
    if len(items) > 4 and getVelue(items[0]).isdigit() and getVelue(items[3]).isdigit():
        if          != None:
            mapTo(currLine)
                 =
    else:
        if          == None:
            currLine = line
        else:
            currLine += line

if currLine != None:
         (        )
```

Random test input generator ():

```

```

Discussion:

Discussion is not available for this exercise.

Latest Submission Grade: 100 (submitted on 2015-05-31 09:42)

```
==================================[Phase1: check_plagirism_cloudera]========================================
Pass.

==================================[Phase2: cloudera_prepare]================================================
Pass syntax checking. You got 10 points.

==================================[Phase3: cloudera_hadoop_py_streaming_local]==============================
Pass all local test cases. Good job! You get 20 out of totally 20 points.

==================================[Phase4: cloudera_hadoop_py_streaming_cluster]============================
Spent time: 29.3648030758 seconds. The standard execution time is 28 seconds.
Your hadoop program runs roughly the same fast as the standard answer. You get 100% of the correctness poi

Correctness points: 70
```

Your submission record:

Submission Date	Grade	Diff
2015-05-31 09:42	100	Diff

2015-05-30 17:25	30	Diff
2015-05-30 17:24	16	Diff
2015-05-29 10:36	10	Diff
2015-05-29 10:24	10	Diff
2015-05-28 20:48	16	Diff
2015-05-28 20:45	10	Diff
2015-05-28 20:44	0	Diff
2015-05-28 20:40	10	Diff
2015-05-28 20:34	10	Diff
2015-05-28 20:29	10	Diff
2015-05-28 20:19	16	Diff
2015-05-28 20:10	16	Diff
2015-05-28 19:55	16	Diff
2015-05-28 19:43	10	Diff
2015-05-28 17:08	10	Diff
2015-05-28 16:49	10	Diff

Top performers in the class:

Name	Points	Total Trials	Last Submission
贾同学	100	24	2015-05-26 17:01
DIS	100	17	2015-05-31 09:42
魏杰伟	100	11	2015-05-31 10:30
把我飘准的普通发凡我	100	8	2015-05-31 14:43
彭翌	100	23	2015-06-01 01:40
test	100	13	2015-06-01 02:25
温伟力	100	6	2015-06-01 09:15
嘟嘟噜	100	8	2015-06-01 11:20
萨瓦迪卡	100	13	2015-06-01 11:43
我是大傻逼	100	1	2015-06-01 13:19

Feedback: (这里不是填答案的地方！)

Warning: Don't put your answer into the feedback, as they will be seen by all other students. If you have doubts on your answer, send an email to the author of this assignment instead.

为减轻服务器负担，在评论中插入大图像时建议用链接，小图像的时候建议用base64编码直接嵌入，即打开html编辑页面插入像<img src="data:image/png;base64,iVBORw0KGgoAAAANSclgAAAAEAAA">一样的标签 (图片转base64编码的工具)。注意，评论的大小不能超过100k字节，如果超出就会显示Invalid form data。

难度(选填): ------------------------------

题意表达是否清楚(选填): ------------------------------

对学习的帮助(选填): ------------------------------

你对本题的评价(选填):

Font Family Font Size Paragraph

Path: p

Upload

Feedbacks from students:

Class	Comments
BDSE_2015	== ============================== Unexpected error occurs when the system is grading your su bmission. As a result, your grade might be lower than what you shall have gotten. Please report the follow ing problem description to wangxm35@mail.sysu.edu.cn befor e re-submitting your answer. == ============================== local variable 'line_count' referenced before assignment Traceback (most recent call last): File "/projects/smart_programmer/labsite/tasks/task_clou dera_hadoop_py_streaming_local.py", line 166, in hadoop_py _streaming_local result, error_log, _ = hadoop_py_streaming_command(Non e, submission, "local", file_names, "output"+str(test_inpu t.input_index), compile_dir, None, None) File "/projects/smart_programmer/labsite/tasks/hadoop_py _streaming_common.py", line 256, in hadoop_py_streaming_co mmand line_count+=1 UnboundLocalError: local variable 'line_count' referenced before assignment == ==============================
BDSE_2015	============[Phase4: cloudera_hadoop_py_streaming_cluster] == Hadoop failure Timeout (exceeds hard time limit 60 seconds) Current step: 1 Hadoop command: hadoop jar /opt/cloudera/parcels/CDH/lib/h adoop-mapreduce/hadoop-streaming.jar -input smart_programm er/udacity_forum/input1 -mapper mapper1.py -file mapper1.p y -reducer reducer1.py -file reducer1.py -output smart_p rogrammer/tmp/24860/tmp_result/ Detailed hadoop log: Step 1: hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-ma preduce/hadoop-streaming.jar -input smart_programmer/udaci ty_forum/input1 -mapper mapper1.py -file mapper1.py -redu cer reducer1.py -file reducer1.py -output smart_programme r/tmp/24860/tmp_result/ 15/06/01 15:53:34 WARN streaming.StreamJob: -file option i s deprecated, please use generic option -files instead. 15/06/01 15:53:35 INFO client.RMProxy: Connecting to Resou rceManager at eden/172.16.21.236:8032 15/06/01 15:53:35 INFO client.RMProxy: Connecting to Resou rceManager at eden/172.16.21.236:8032 15/06/01 15:53:36 INFO mapred.FileInputFormat: Total input paths to process : 1 15/06/01 15:53:36 INFO mapreduce.JobSubmitter: number of s

	plits:2 15/06/01 15:53:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1433116351930_0435 15/06/01 15:53:36 INFO impl.YarnClientImpl: Submitted appl ication application_1433116351930_0435 15/06/01 15:53:36 INFO mapreduce.Job: The url to track the job: http://eden:8088/proxy/application_1433116351930_043 5/ 15/06/01 15:53:36 INFO mapreduce.Job: Running job: job_143 3116351930_0435 这个集群方式要怎么写
BDSE_2015	
BDSE_2015	=== ============================== Unexpected error occurs when the system is grading your su bmission. As a result, your grade might be lower than what you shall have gotten. Please report the follow ing problem description to wangxm35@mail.sysu.edu.cn befor e re-submitting your answer. === ============================== local variable 'line_count' referenced before assignment Traceback (most recent call last): File "/projects/smart_programmer/labsite/tasks/task_clou dera_hadoop_py_streaming_local.py", line 166, in hadoop_py _streaming_local result, error_log, _ = hadoop_py_streaming_command(Non e, submission, "local", file_names, "output"+str(test_inpu t.input_index), compile_dir, None, None) File "/projects/smart_programmer/labsite/tasks/hadoop_py _streaming_common.py", line 256, in hadoop_py_streaming_co mmand line_count+=1 UnboundLocalError: local variable 'line_count' referenced before assignment === ============================== 这是什么问题?

BDSE_2015	Failed Test Case #1
	
	Fail to execute test oracle script:
	# Python script snippet to implement test oracles on outpu
	t files
	# Two input variables: standard_output, submission_output,
	 task_points
	# Two output variables: compare_info, grade
	
	compare_info=""
	set1=set()
	for line in standard_output.splitlines():
	    set1.add(" ".join(line.split()))
	
	set2=set()
	for line in submission_output.splitlines():
	    set2.add(" ".join(line.split()))
	
	if set1==set2:
	    grade = task_points
	    compare_info=""
	else:
	    grade = 0
	    compare_info="Your result is different from the standa
	rd answer:\n\n"

```
    compare_info+="\n============lines that the standard a
nswer contains, but your submission does not============\n
"
    if len(set1.difference(set2))!=0:
        for line in set1.difference(set2):
            compare_info+=str(line)+"\n"
    if len(set2.difference(set1))!=0:
        compare_info+="\n============lines that your submi
ssion contains, but standard answer does not============\n
"
        for line in set2.difference(set1):
            compare_info+=str(line)+"\n"



Exception:
can only concatenate list (not "str") to list
```

BDSE_2015

问题要求如果同一人在多个时间段达到最多，是都要输出的，可是判定标准是只给出一个。上次老师给的答案有同样的问题，下面有人反馈这个问题，老师好像没注意啊。

BDSE_2015

```
Fail to execute test oracle script:
# Python script snippet to implement test oracles on outpu
t files
# Two input variables: standard_output, submission_output,
 task_points
# Two output variables: compare_info, grade

compare_info=""
set1=set()
for line in standard_output.splitlines():
    set1.add(" ".join(line.split()))

set2=set()
for line in submission_output.splitlines():
    set2.add(" ".join(line.split()))

if set1==set2:
    grade = task_points
    compare_info=""
else:
    grade = 0
    compare_info="Your result is different from the standa
rd answer:\n\n"
    compare_info+="\n============lines that the standard a
nswer contains, but your submission does not============\n
"
    if len(set1.difference(set2))!=0:
        for line in set1.difference(set2):
            compare_info+=str(line)+"\n"
    if len(set2.difference(set1))!=0:
        compare_info+="\n============lines that your submi
ssion contains, but standard answer does not============\n
"
        for line in set2.difference(set1):
            compare_info+=str(line)+"\n"



Exception:
can only concatenate list (not "str") to list
```

bugs.

Thanks to:

- 吴浩坚同学 and 梁展瑞同学
- Django, Gunicorn, TinyMCE, Sandbox, Nginx
- Valgrind, Google Gode Style SOClone

Recent messages:

- [06/01 14:50] From 王欣明: 最近一段时间经常发生有些同学的作业卡队列的现象。如果发现请及时通知我重启服务队列。
- [06/01 14:49] From 王欣明: 本周是大数据软件工程的期末课程项目，题目会在周二晚上6点开放，同学们可以自己在宿舍做，一周时间内完成。
- [05/31 00:38] From 谢议尊: test

Send message to an user with real name or nick name:

Name:

Message:

Send Message