Forum - Top10 tags

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

你已经在这道题上花费大量时间，是否题目有让你困惑的地方？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!

You shall run the code mostly on your VMs. The dataset is in the file forum_data.tar.gz. To unarchive it, download it to your VM, put in the data directory and run:

tar zxvf forum_data.tar.gz

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.

Write a MapReduce program that would output Top 10 tags, ordered by the number of questions they appear in.

Please note that you should only look at tags appearing in questions themselves (i.e. nodes with node_type "question"), not on answers or comments.

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 when using the test data set provided. The output of your code should include all of the rows below, aside from the columns headers. The order of the rows does matter for this question. Tags that tie for the same number of counts can be switched with one another, but the list as a whole needs to show tags starting with the highest count and descending.

Tag	Counts
cs101	8

cs253	5
discussion	5
issues	3
welcome	3
homework	2
jobs	2
lessons	2
meta	2
nationalities	2

Hint:

Hint is not available for this exercise.

Time limit:

Hard time limitation: 60 seconds

Standard answer spent time: 50 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\"add

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

def      (    ):
    stuInfo = line.replace("\n"," ").split("\t")
    if    (       ) == 19:
        tagnames = getValue(stuInfo[2])
                  =         (       [5])
        if node_type == "question":
                         =         .     (' ')
            for item in tagnamesList:
                print "{0}\t{1}".      (    , 1)

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

          =     .     ().     ("\t")
    if len(items) > 4 and getValue(items[0]).isdigit() and getValue(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: 30 (submitted on 2015-05-31 14:34)

```
=================================[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]===========================
Hadoop failure

Timeout (exceeds hard time limit 60 seconds)
Current step: 2
Hadoop command: hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -input smar
Detailed hadoop log:
Step 1: hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -input smart_progra
15/05/31 14:36:16 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files i
15/05/31 14:36:17 INFO client.RMProxy: Connecting to ResourceManager at eden/172.16.21.236:8032
15/05/31 14:36:18 INFO client.RMProxy: Connecting to ResourceManager at eden/172.16.21.236:8032
15/05/31 14:36:18 INFO mapred.FileInputFormat: Total input paths to process : 1
15/05/31 14:36:18 INFO mapreduce.JobSubmitter: number of splits:2
15/05/31 14:36:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1433003308145_0151
15/05/31 14:36:19 INFO impl.YarnClientImpl: Submitted application application_1433003308145_0151
15/05/31 14:36:19 INFO mapreduce.Job: The url to track the job: http://eden:8088/proxy/application_143300
15/05/31 14:36:19 INFO mapreduce.Job: Running job: job_1433003308145_0151
15/05/31 14:36:25 INFO mapreduce.Job: Job job_1433003308145_0151 running in uber mode : false
15/05/31 14:36:25 INFO mapreduce.Job:  map 0% reduce 0%
15/05/31 14:36:32 INFO mapreduce.Job:  map 100% reduce 0%
15/05/31 14:36:38 INFO mapreduce.Job:  map 100% reduce 17%
15/05/31 14:36:40 INFO mapreduce.Job:  map 100% reduce 67%
15/05/31 14:36:41 INFO mapreduce.Job:  map 100% reduce 83%
15/05/31 14:36:43 INFO mapreduce.Job:  map 100% reduce 100%
```

```
15/05/31 14:36:43 INFO mapreduce.Job: Job job_1433003308145_0151 completed successfully
15/05/31 14:36:43 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=135787
                FILE: Number of bytes written=1232930
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=120319220
                HDFS: Number of bytes written=706
                HDFS: Number of read operations=24
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=12
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=6
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=11532
                Total time spent by all reduces in occupied slots (ms)=29138
                Total time spent by all map tasks (ms)=11532
                Total time spent by all reduce tasks (ms)=29138
                Total vcore-seconds taken by all map tasks=11532
                Total vcore-seconds taken by all reduce tasks=29138
                Total megabyte-seconds taken by all map tasks=11808768
                Total megabyte-seconds taken by all reduce tasks=29837312
        Map-Reduce Framework
                Map input records=922570
                Map output records=100698
                Map output bytes=977197
                Map output materialized bytes=162505
                Input split bytes=266
                Combine input records=0
                Combine output records=0
                Reduce input groups=8131
                Reduce shuffle bytes=162505
                Reduce input records=100698
                Reduce output records=60
                Spilled Records=201396
                Shuffled Maps =12
                Failed Shuffles=0
                Merged Map outputs=12
                GC time elapsed (ms)=571
                CPU time spent (ms)=16180
                Physical memory (bytes) snapshot=2645811200
                Virtual memory (bytes) snapshot=22003052544
                Total committed heap usage (bytes)=3206021120
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=120318954
        File Output Format Counters
                Bytes Written=706
15/05/31 14:36:43 INFO streaming.StreamJob: Output directory: smart_programmer/tmp/19424/tmp_result/
Step 2: hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -input smart_progra
15/05/31 14:36:51 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files i
15/05/31 14:36:52 INFO client.RMProxy: Connecting to ResourceManager at eden/172.16.21.236:8032
15/05/31 14:36:52 INFO client.RMProxy: Connecting to ResourceManager at eden/172.16.21.236:8032
15/05/31 14:36:53 INFO mapred.FileInputFormat: Total input paths to process : 6
15/05/31 14:36:54 INFO mapreduce.JobSubmitter: number of splits:6
15/05/31 14:36:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1433003308145_0152
15/05/31 14:36:54 INFO impl.YarnClientImpl: Submitted application application_1433003308145_0152
15/05/31 14:36:54 INFO mapreduce.Job: The url to track the job: http://eden:8088/proxy/application_1433003
15/05/31 14:36:54 INFO mapreduce.Job: Running job: job_1433003308145_0152
15/05/31 14:37:00 INFO mapreduce.Job: Job job_1433003308145_0152 running in uber mode : false
15/05/31 14:37:00 INFO mapreduce.Job:  map 0% reduce 0%
15/05/31 14:37:05 INFO mapreduce.Job:  map 17% reduce 0%
15/05/31 14:37:06 INFO mapreduce.Job:  map 83% reduce 0%
15/05/31 14:37:09 INFO mapreduce.Job:  map 100% reduce 0%
15/05/31 14:37:13 INFO mapreduce.Job:  map 100% reduce 67%
15/05/31 14:37:14 INFO mapreduce.Job:  map 100% reduce 83%
```

Your submission record:

Submission Date	Grade	Diff
2015-05-31 14:34	30	Diff
2015-05-31 14:06	100	Diff
2015-05-31 13:33	30	Diff
2015-05-31 12:24	30	Diff
2015-05-31 12:21	30	Diff
2015-05-31 12:18	30	Diff
2015-05-31 12:10	30	Diff
2015-05-31 11:51	30	Diff
2015-05-31 11:49	30	Diff
2015-05-31 11:44	30	Diff
2015-05-31 10:17	30	Diff
2015-05-30 17:40	30	Diff
2015-05-30 17:31	23	Diff
2015-05-30 17:15	10	Diff
2015-05-30 11:17	16	Diff
2015-05-29 11:12	16	Diff

Top performers in the class:

Name	Points	Total Trials	Last Submission
DIS	100	16	2015-05-31 14:34
彭翌	100	10	2015-06-01 01:48
test	100	15	2015-06-01 02:36
温伟力	100	6	2015-06-01 11:23
我是大傻逼	100	1	2015-06-01 13:13
把我飘准的普通发凡我	100	13	2015-06-01 13:34
持盾卫士	100	8	2015-06-01 13:58
魏杰伟	100	24	2015-06-01 14:44
聂照昌	30	4	2015-05-28 22:38
萨瓦迪卡	30	3	2015-05-31 11:33

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:

<table>
<tr><th>Class</th><th>Comments</th></tr>
<tr><td>BDSE_2015</td><td>这是要逼疯人的节奏啊</td></tr>
<tr><td>BDSE_2015</td><td></td></tr>
<tr><td>BDSE_2015</td><td><pre>
==
==============================
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

==
==============================
</pre></td></tr>
<tr><td>BDSE_2015</td><td>题目和内容不一致，把Student Group 的题贴到这里来了，还以为看错了。</td></tr>
</table>

Please contact **Dr. Wang (roadlit at gmail.com)** to report 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