Quiz #4

Instructions

PySpark using DataFrame

Solution must be provided in PySpark using DataFrames and SQL. You must NOT use RDDs at all. I will discuss this assignment in details on Tuesday, May 19, 2020. This assignment involves some basic Spark processing using the White House Visitor Log.

Input:

You can find the entire input dataset: here

https://obamawhitehouse.archives.gov/sites/default/files/disclosures/whitehouse_waves-2016_12.csv_.zip

You MUST not edit the input provided.

The attributes in this dataset are defined
in the first record of downloaded file.

The attributes are:

NAMELAST,
NAMEFIRST,
NAMEMID,
...
visitee_namelast,
visitee_namefirst,
...

Data Clean up:

1. All rows must be dropped if NAMELAST is null.
2. All rows must be dropped if visitee_namelast is null.
3. All data must be converted to lowercase
4. If a record is empty, then drop it

PySpark Algorithm:

You are required to write **efficient** PySpark program
using Spark's DataFrames and actions to find the
following information:

(a) The 10 most frequent visitors to the White House.
    (NAMELAST, NAMEFIRST)

(b) The 10 most frequently visited people in
    the White House.
    (visitee_namelast, visitee_namefirst)

(c) The 10 most frequent visitor-visitee combinations.

(d) The number of records dropped (due to filtering)

(e) The number of records processed

Requirements:

a) NAMELAST can not be null/empty.
b) visitee_namelast can not be null/empty.

Input Data Format:

Your data is comprised of a set of records, where each
record contains a single visitor log, where fields
are separated by ",".

Sample Input Example (10 records):

NAMELAST,NAMEFIRST,NAMEMID,UIN,BDGNBR,ACCESS_TYPE,TOA,POA,TOD,POD,APPT_MADE_DATE,APPT_START_DATE,APPT_END_DATE,APPT_CANCEL_DATE,Total_People,LAST_UPDATEDBY,POST,LASTENTRYDATE,TERMINAL_SUFFIX,visitee_namelast,visitee_namefirst,MEETING_LOC,MEETING_ROOM,CALLER_NAME_LAST,CALLER_NAME_FIRST,CALLER_ROOM,DESCRIPTION,Release_Date
TAJOURIBESSASSI,HANENE,,U22101,,VA,,,,,9/2/2015 0:00,10/1/2015 3:00,10/1/2015 23:59,,1,AR,WIN,9/2/20

```
15 11:38,AR,Pelofsky,Eric,OEOB,226,ROWBERRY,ARIANA,,,1/29/2016
bageant,laura,j,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/2015 13:
42,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
Broemson,Earl,H,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/2015 14:
41,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
Jackling Jr,William,C,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/20
15 13:42,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
McCrary,Richard,L,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/2015 1
3:42,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
Mulcahy,Joshua,E,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/2015 1
3:42,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
Ryan,Oliver,J,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/2015 14:4
1,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
Smith Jr,William,T,U30528,,VA,,,,,9/29/2015 0:00,10/1/2015 5:00,9/30/2016 23:59,,7,WW,WIN,9/29/2015
 13:42,WW,Baskerville,Steven,WH,WH Grounds,WARDEN,WILLIAM,,,1/29/2016
Keeler,Douglas,E,U21657,,VA,,,,,9/1/2015 0:00,10/1/2015 6:30,10/1/2015 23:59,,1,LD,WIN,9/1/2015 11:0
4,LD,Goldstein,Jeff,NEOB,7013,DUKE,LAURA,,,1/29/2016
```

The total number of records is 970,505
(which includes the header line.)

Generic Solution

Your solution will be a PySpark solution,
which can be run by ./bin/submit-spark command
as (assume my name is Alex Smith):

```
./bin/submit-spark quiz4_alex_smith.py  N  input-path
```

where N is an integer: N=5 means Top-5, N=10 means Top-10

NOTE-1: If a given record is missing visitor
or visitee then that record is dropped
from all calculations

NOTE-2: Your solution has to be generic and
should be able to handle billions of records

NOTE-3: You have to pass 2 input parameters
to your PySpark program

Expected output: CLEARLY IDENTIFY OUTPUT Sections

(a) The 10 most frequent visitors to the White House.
visitor is (NAMELAST, NAMEFIRST)

<visitor> <frequency>

(b) The 10 most frequently visited people
    in the White House.
visitee is (visitee_namelast, visitee_namefirst)

<visitee> <frequency>

(c) The 10 most frequent visitor-visitee

combinations.

<visitor-visitee> <frequency>

(d) The number of records dropped

(e) The number of records processed

PySpark Solution:

1. For this assignment, your PySpark program is comprised of Spark DataFrames transformations and actions

2. Your PySpark Solution must be a generic solution and work for any size data. But you will show your PySpark solution step-by-step (as I presented in class) with the input provided.

3. Apply proper Spark DataFrames transformations for the given Input and show your work in detail (step-by-step transformations and actions)

4. You will read 2 input parameters:
   Parameter 1: N (denotes Top-N)
   Parameter 2: white house visitor log file