CS236 Class Project (Due Thursday June11th)
Spring 2020
TA: Pritom Ahmed (pahme002@ucr.edu)

Project Summary

For this project, you will be provided with two real datasets. The first dataset contains station information from weather stations across the world. The second provides individual recordings for the stations over a 4-year period. The goal of the project is to find out which states in the US have the most stable rainfall. That is, the result should provide the US states ordered (in ascending order) by D, where D is the difference between the two months with the highest and lowest rainfall. You will do the project individually. The project should be completed on Spark. You will be using a local instance of Spark. The TA will upload instructions on how to quickly setup local Spark in a Unix environment.

Suggested Approach

Task 1 (25 points) : As your datasets contain data from stations around the world while what you are asked is about US states, you would first need to identify which stations are within the United States. Then you would need to **group stations by state**.
Task 2 (25 points) : For each **state** with readings, you will first need to find the **average precipitation recorded for each month** (ignoring year).
Task 3 (25 points) : Then find the **months** with the **highest and lowest averages** for each **state**.
Task 4 (25 points) : You will need to **order** the states by the difference between the **highest and lowest month** average, in ascending order.
In your result, for each state, you should return:
The state abbreviation, e.g. “CA”
The average precipitation and name of the highest month, e.g. “90, July”
The average precipitation and name of the lowest month, e.g. “50, January”
The difference between the two, e.g. “40”

Dataset Information

Please find the dataset here.

The *LOCATIONS* dataset is a single .csv file (*WeatherStationLocations.csv*), containing the metadata for every station across the world. To identify that a station is in the US, you need to look for stations where the “CTRY” field is “US” and the “ST” field is non-empty. Keep in mind that the first row of this file is Header. Here are the fields for this dataset:
USAF = Air Force station ID. May contain a letter in the first position.
WBAN = NCDC WBAN number
CTRY = FIPS country ID
ST = State for US stations

LAT = Latitude in thousandths of decimal degrees
LON = Longitude in thousandths of decimal degrees
ELEV = Elevation in meters
BEGIN = Beginning Period Of Record (YYYYMMDD).
END = Ending Period Of Record (YYYYMMDD).
Sample Row:
"724920","23237","STOCKTON METROPOLITAN AIRPORT","US","CA","+37.889","-121.226","+0007.9","20050101","20140403"

The *RECORDINGS* dataset is contained in four files, one for each year (2006-2009). The "**STN---**" value will match with the "USAF" field in the locations dataset. These files are concatenated from many small files, so keep in mind that there will be Header lines through the files. Here are the fields for this dataset:
STN--- = The station ID (USAF)
WBAN = NCDC WBAN number
YEARMODA = The datestamp
TEMP = Ignore for this project
DEWP = Ignore for this project
SLP = Ignore for this project
STP = Ignore for this project
VISIB = Ignore for this project (Visibility)
WDSP = Ignore for this project
MXSPD = Ignore for this project
GUST = Ignore for this project
MAX = Ignore for this project (Max Temperature for the day)
MIN = Ignore for this project (Min Temperature for the day)
PRCP = Precipitation
NDP = Ignore for this project
FRSHTT = Ignore for this project

Sample Row:
997781 99999 20061121 42.4 13 9999.9 0 9999.9 0 9999.9 0 999.9 0 17.5 13 22.0 999.9 46.2* 39.0* 0.00I 999.9 000000

Calculation of Precipitation

A value of 99.99 means that there is no data (You can either treat these as 0 or exclude them from your calculations, though the latter is the more correct option)

There is a letter at the end of the recordings. Here is a table of what the letters mean (Basically the letter tells you how long the precipitation was accumulated before recording).

A - 6 hours worth of precipitation

B - 12 hours...

C - 18 hours...

D - 24 hours...

E - 12 hours... (slightly different from B but the same for this project).

F - 24 hours ... (slightly different from D but the same for this project).

G - 24 hours ... (slightly different from D but the same for this project).

H - station recorded a 0 for the day (although there was some recorded instance of precipitation).

I - station recorded a 0 for the day (and there was NO recorded instance of precipitation).

A simple solution would be to multiply. For instance, if they recorded 12 hours worth of precipitation, multiply it by 2 to extrapolate 24 hours worth. How you treat these is up to you (just let me know in your README)

Due Date

This project will be due Thursday, June11th before Midnight. Email the TA your submissions.

Deliverables

A zipped file containing:

1. The script to run your job
2. The full source code
3. A *README* describing your project, including:

1. An overall description of how you chose to divide the problem into different spark operations, with reasoning.

2. A description of each step job

3. Total runtime for the whole project

4. A description of anything you did extra for the project, such as adding a combiner. If there is anything that you feel deserves extra credit, put it here.

Do not include the two original datasets.

The script should take as input three things:

1. A folder containing the Locations file
2. A folder containing the Recordings files
3. An output folder

Hints

1. Make sure you parse correctly. One file is a csv, the other is not. One file has a single row of Headers, the other has many.
2. Think about how many steps you should use. How much work should each step do?
3. It is a good idea to read about different types of RDDs and various operations they allow users to do.
4. You may want to create a subset of the reading dataset and use that while you are developing, instead of using the whole dataset. This can save some time.
5. Make sure you start early.

Potential Extra Credit (Optional)

Please feel free to try to beef up your project for extra credit. There are many ways that you can do this. Here are a few examples:

- A clever way to achieve faster execution time (fewer steps); if you do that show how much improvement you get.
- **Bigger Bonus:** Include the stations with “CTRY” as “US” that do not have a state tag, finding a way to estimate the state using a spatial distance with the known stations. There are some stations that are Ocean Buoys so you may want to have a maximum distance to be required in order to be included in a state, or you could create a separate “state” representing the “pacific” and “atlantic” ocean (Checked by using coordinates). There is a lot of potential work here so the extra credit could be large.

Whatever you try to do let the TA know in your *README*. Make sure that you complete the project before going for extra credit. If you aren’t sure whether your idea is worth extra credit or not, just email the TA.