NYU Big Data Fall 2022

Assignment 2 - Spark Dataframes (100 points + 25 points extra credit)

Due Date: Thursday, October 13

2PM Eastern

**** Give attribution to any code you use that is not your original code ****

SUBMIT YOUR SOLUTION AS A JUPYTER NOTEBOOK.
Use your netid: e.g. jcr365-hw2.ipynb

If I cannot run your notebook, you will not get full credit.

Datasets are in jupyterhub's shared folder.

1. 15 points

Datafile: BreadBasket_DMS.csv

Solve: Show the top 5 items bought (count) for the time period between 9:00AM inclusive and 11:00PM exclusive.

2. 15 Points

Dataset: Restaurants_in_Durham_County_NC.csv

NOTE* This file is colon delimited (not comma). Do not preprocess it; read it with spark.read…**

Solve: Summarize the number of entities by "rpt_area_desc"

Example:
"Swimming Pools", 13
"Tatoo Establishment", 2
:

3. 50 Points

Dataset: populationbycountry19802010millions.csv

Solve: For each year and region, compute percentage increase in population, ***year over year***. Note the year 1980 will not have a preceding year.

For each year, display the top and bottom country in terms of global growth

Example:

Year, Region, yearly increase, percent of global year increase (these results are made up)

1981, North America, 1.30%
1981, Bermuda, 0.1%
1982, Aruba,

4. 20 Points

Dataset: romeo-juliet-pg1777.txt

Solve: WordCount

Do a word count exercise using pyspark. Ignore punctuation, and normalize to ***lower case***. Accept only the characters in this set: **[0-9a-zA-Z]**

5. Extra credit – 25 points

Datasets:
durham-nc-foreclosure-2006-2016.json
Restaurants_in_Durham_County_NC.json

Solve: For each restaurant ('Restaurants_in_Durham_County_NC.json') classified as "status":"ACTIVE" **and** ""rpt_area_desc": "Food Service":

For each restaurant, show the number of foreclosures ('durham-nc-foreclosure-2006-2016') within a radius **of 10 miles** of the restaurant's coordinates.

Note: Assume the shape of Earth is a sphere. You can use the Haversine distance.
https://pypi.org/project/haversine/