NYU Big Data **Spring 2023**

Assignment 3 – Spark & Dataframes
(100 points + 30 points extra credit)

Due Date: Sunday, March 12, 11:55PM Eastern

Submission Rules:

- Give attribution to any code you use that is not your original code
- Submit the homework as a Jupyter Notebook. Use your netid: e.g. jcr365-hw2.ipynb
- If we cannot run your notebook, you will not get full credit.

Datasets are in Brightspace.

1. 15 points

Datafile: BreadBasket_DMS.csv

Solve: What is the most popular (most sold) between the 8:00AM and 8:59AM for each day?

Example output (not actual solution)

2016-10-30, Pastry
2016-10-31, Coffee
:
:

2. 15 points

Datafile: BreadBasket_DMS.csv

Solve: What is the most common item bought along with “Brownie”? (items bought in the same transaction)

3. 10 Points

Dataset: Restaurants_in_Durham_County_NC.csv
NOTE* This file is colon delimited (not comma) ***

Solve: How many years are represented in this dataset?

4. 20 Points

Dataset: Restaurants_in_Durham_County_NC.csv

Solve: Show the type and count of restaurant **opened** during the 90's (1990-1999 inclusive). Note: type="Rpt_Area_Desc"

Example (not the actual result):
"Swimming Pools", 13
"Tatoo Establishment", 2
:

5. 25 Points

Dataset: populationbycountry19802010millions.csv

Solve: For region, compute the ***percentage change*** in population, year over year. Note the year 1980 will not have a preceding year. For each year, display the region with the top population ***decrease***.

Example (not actual results):

1981, North America, -2%
1982, Aruba, -7%…

6. 15 Points

Dataset: romeo-juliet-pg1777.txt

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

7. Extra credit – 30 points

Datasets:

Restaurants_in_Durham_County_NC.csv
durham-nc-foreclosure-2006-2016.json

Solve: For each restaurant ('Restaurants_in_Durham_County_NC.csv) with "status"="ACTIVE" **and** ""rpt_area_desc"="Food Service", show the number of foreclosures ('durham-nc-foreclosure-2006-2016') within a radius **of 1 mile** of the restaurant's coordinates.

Note: Use any assumption for the shape of Earth…

Or you can use the Haversine distance. https://pypi.org/project/haversine/

Note: UDF, or user defined functions, is part of next week's lecture.