CS-GY-6513 Big Data FALL 2024

Assignment 2 - Spark (130 points + 35 extra credit)

Due Date: Friday, October 25, 11:59PM

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.

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

Instructions

Refer to the notebook **HW2.ipynb** and the ***data*** folder in the course website.

*** ALL DATASETS ARE AVAILABLE IN THE JUPYTERHUB SHARED FOLDER

1. 25 points **Data**: shared/data/Bakery.csv

Show the highest selling **item for Mondays, per hour**, for the 7AM to 11AM hours. Note that "weekday", "period" have to be computed.

For example (these are made up numbers….)
Item qty, weekday, Date , Hour-period, qty
Bread, 102, Monday, 2016-10-31, 7AM
Coffee, 132, Monday, 2016-10-31, 8AM
:

2. 25 points **Data**: shared/data/Bakery.csv

Show the top 2 (by qty) items bought **by Daypart, by DayType.**
Note:

Daypart = Breakfast if 6AM – 10:59AM, Lunch if 11:01AM – 3:59PM, Dinner otherwise
DayType = Weekend if Sat, Sun, Weekday otherwise

For example (not necessarily the right numbers….)
Weekend, Breakfast, (coffee, Muffin)
Weekend, Lunch, (cookies, pastry)
:
** The Answer **MUST** include the 2 items in a single column

3. 20 Points **Data:** shared/data/Restaurants_in_Durham_County_NC.json

Show the number of entities by “fields.rpt_area_desc”

Example (not true numbers):
“Food Service”, 13
“Tatoo Establishment”, 2
:

4. 20 Points. **Data:** shared/data/populationbycountry19802010millions.csv

Show the country or region with ***the biggest percentage increase*** in population AND the country with **biggest percentage decrease** in population, between the years 1990 and 2000. Use only the countries, not ‘World’.

Example (Not the real answer):

North America, 2.30% <- assuming North America was max
Aruba, -22.2%… <- assuming Aruba was min

5. 20 Points **Data**: hw1text (from HW1).

Solve: do WordCount

Do ***word count*** exercise using pyspark.
Ignore punctuation and normalize to ***lower case***.
i.e. replace characters in NOT in this set: **[0-9a-z]** with **space.**

HINT: You can use the sparkml package.

6. 20 Points **Data**: hw1text (from HW1)

Find the 10 most common bigrams

HINT: You can use the sparkml package.

7. Extra credit – 40 points

Data:

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

a) Find food service and active restaurants (“status” = “ACTIVE” **and** “"rpt_area_desc" = "Food Service”) ***closest*** to the following coordinate: **of 35.994914, -78.897133,** and show it.
b) With that restaurant in (a) as your center point, find the number of foreclosures within a 1 mile radius

You can use an external library for calculating coordinate distances.
The ***haversine*** library is available in Jupyterhub’s bigdata environment.