## CS-GY-6513 Big Data Spring 2022

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

**Due Date: Sunday, March 13th, 11:55PM 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.

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

### Instructions

Refer to the jupyter notebook **HW2.ipynb**, and the ***data*** folder in the resources section of the course website.

### *** ALL DATASETS ARE IN THE JUPYTER HUB SHARED FOLDER ***

### 1. 15 points

**Datafile**: Bakery.csv

**Solve**: Show the total number bought **by item, per day, per hour range**

For example (not necessarily the right numbers….)

Bread, 2016-10-30, 09, 1
Bread, 2016-10-30, 10, 11
:

### 2. 15 points

**Datafile**: Bakery.csv

**Solve**: Show the highest 3 items bought **by Daypart, by DayType**

For example (not necessarily the right numbers….)

Morning, (bread, pastry, Muffin), Weekend

:

## 3. 15 Points

**Dataset:** Restaurants_in_Durham_County_NC.json

**Solve:** Summarize the number of entities by "rpt_area_desc"

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

## 4. 25 Points

**Dataset:** populationbycountry19802010millions.csv

**Solve:** For each year, show the region with the biggest percentage increase in population, year over year. Ignore year 1980.

For example, the year over year for North America in 1981 is:
(324.44694 - 320.27638) / 320.2763 = 1.30%

Example:

1981, North America, 1.30% <- assuming North America was the max increase
1982, Aruba, … <- assuming Aruba was max that year

## 5. 15 Points

**Dataset**: internet_archive.scifi_v3.txt

**Solve:** WordCount

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

## 6. 15 Points

**Dataset**: internet_archive_scifi_v3.txt

Find the 10 most common bigrams

## 7. 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":

Show the number of foreclosures ('durham-nc-foreclosure-2006-2016'**) within a radius of 1 minute** of the restaurant's coordinates.