Second Homework Assignment (20% of grade)

Submission Guidelines:

Please follow carefully the instructions posted on the course's github page when submitting your solutions (https://github.com/MSIA/bigdatacourse/blob/master/README.md). Failure to follow the instructions will result in lost points.

Deliverables:

1. Your spark source code file: name the file lastname_x.py (scala,java)
2. Output files: name them lastname_x.txt
3. Short write-up of findings as instructed below: name them lastname_findings_x.txt

Here 'x' is the number of the task.

You have to use spark. You can use scala, python, or java and you can use all libraries available in spark. You are not allowed to grab other code form the internet that is based on spark. (It is allowed to use python specific libraries such as nltk, scikit-learn, etc.)

Note:

All input files are available in /home/public/crime. Please copy them directly from /home/public/crime to HDFS and not to your home directory on wolf.

Due on Wednesday, May 20 at 5 pm.

Crime in Chicago

Yes, Chicago has crime, and 6 million events since 2001. If we live in a wonderland, there would be no Spark homework assignment. But we don't.

The Chicago crime data is available in /home/public/crime. The file has the header that explains many fields. Less obvious fields: block = the first 5 characters correspond to the block code and the rest specify the street location; IUCR = Illinois Uniform Crime Reporting code; X/Y coordinates = to visualize the data on a map, not needed in the assignment; District, Beat = police jurisdiction geographical partition; the region is partitioned in several districts; each district is partitioned in several beats; http://gis.chicagopolice.org/pdfs/district_beat.pdf; community areas and wards: https://www.chicago.gov/city/en/depts/dgs/supp_info/citywide_maps.html

Perform the following tasks.

1. By using SparkSQL, generate a histogram of average crime events by month. Find an explanation of results. (10 pts)

2. By using plain Spark (RDDs): (1) find the top 10 blocks in crime events in the last 3 years; (2) find the two beats that are adjacent with the highest correlation in the number of crime events (this will require you looking at the map to determine if the correlated beats are adjacent to each other) over the last 5 years (3) establish if the number of crime events is different between Majors Daly and Emanuel at a granularity of your choice (not only at the city level). Find an explanation of results. (20 pts)
3. Predict the number of crime events in the next week at the beat level. Violent crime events represent a greater threat to the public and thus it is desirable that they are forecasted more accurately (IUCR codes available here: https://data.cityofchicago.org/widgets/c7ck-438e). (45 pts) You are encouraged to bring in additional data sets. (extra 10 pts if you mix the existing data with an exogenous data set) Report the accuracy of your models. You must use Spark dataframes and ML pipelines.
4. Find patterns of crimes with arrest with respect to time of the day, day of the week, and month. Use whatever method in spark you would like. (25 pts)