Fifth Homework Assignment (25% of grade)

Due on Friday, June 12 at 5 pm.

Deliver the following in folder /home/public/assignment (you should have write permissions; after you copy every file, execute: chmod 400 file_name) for each exercise:

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 grabbed other code form the internet. The installed version of spark is 1.3.

Crime in Chicago

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

The Chicago crime data is available in /home/public/course/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: http://www.cityofchicago.org/city/en/depts/doit/dataset/boundaries_-_communityareas.html; wards: http://www.cityofchicago.org/city/en/depts/doit/dataset/boundaries_-_wards.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 (no Spark SQL): find the top 10 blocks in crime events in the last 3 years, find the two beats that are adjacent with the highest correlation in the number of crime events, 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. The higher the IUCR is, the more severe the crime is. Violent crime events are more important and thus it is desirable that they are forecasted more accurately. (45 pts) You are encouraged to bring in additional data sets. (extra 50 pts if you mix the existing data with an exogenous data set) Report the accuracy of your models.
4. Find patterns of crimes with arrest with respect to time of the day, day of the week, and month. (25 pts)