CMPT 898 HOME EXERCISE 7
INITIAL EXPOSURES TO SPARK CASSANDRA DATA SOURCE AND BREEZE-VIZ

This exercise will provide an experience of manipulating Cassandra data in Spark, and further experience in creating figures using Spark data in Breeze-Viz.

Because of the computational loads involved, the configuration challenges and the newness of some of the concepts, students are requested where possible to pursue this exercise in pairs. Such pair programming is also likely to improve the quality of the results, and the timeliness with which they can be produced.

Please note that this exercise will require that you start the system with the Cassandra connector enabled. You will also need to undergo the steps covered in class.

1. As in class, please create a dataframe (*dfGPS*) from the GPS table for Cassandra study 73, but filter it to only include rows with at least modest accuracy (value of the "accuracy" field less than [i.e., finer grained than] 50), only include data from user_id 231.
2. Please call *cache()* on the above, in order to retain the dataframe in memory if possible (i.e., to avoid recomputing the dataset, including loading it again from cassandra). Please note that given Spark's laziness, this caching will not occur until the dataframe is in fact computed!
3. Please count the number of rows in that dataframe
4. Please now compute and show a dataframe giving the count of records associated with each value of the "provider" column. Please note that to count the number of records aggregated in each group, one can perform an agg(count("*"))
5. Please now compute and show a dataframe giving the count of records associated with bins (of size 2) of the accuracy reported, ordering by that accuracy. Please note that in order to do this, you may wish to use the Spark SQL "round" function to operate on a column. To do so, please *import org.apache.spark.sql.functions.round*
6. Please now filter the data further to only include values with at "accuracy" field less than 30, and including only the "speed" column. Please cache and "show" the result.
7. Based on the experience above, did you notice any difference in speed between the first time getting values with *dfGPS* and subsequent uses? Please comment as to a plausible explanation as to this difference.
8. Please count the number of records associated with this further-filtered dataframe.
9. Using the collect() function and map, create an Double array of accuracy values. Please note that in performing this, you will want to perform getFloat(0) on each Row within the array resulting from collect(), and then call .toDouble to convert this to a double.
10. Using breeze-viz's "hist" function (as used in the previous take-home exercise), please create a histogram of the array of double precision accuracy values from the above. Please note that similar to our work in class, you will want to pass this as a DenseVector to "hist". You are recommended to use the xlim property (which is assigned a pair of doubles giving the minimum and maximum points on the x range).
11. Similarly, please create a histogram for the accuracy field (for those times when it is less than 30).

Due in class November 3, 2016