Correlation

November 30, 2017

```
In [22]: from pyspark.sql.types import StructType, StructField, FloatType, LongType, StringType
         from pyspark.shell import spark

         feats = []
         f = open('features.txt')
         for line_num, line in enumerate(f):
             if line_num == 0:
                 # Timestamp
                 feats.append(StructField(line.strip(), LongType(), True))
             elif line_num == 1:
                 # Geohash
                 feats.append(StructField(line.strip(), StringType(), True))
             else:
                 # Other features
                 feats.append(StructField(line.strip(), FloatType(), True))

         schema = StructType(feats)

In [23]: df = spark.read.format('csv').option('sep', '\t').schema(schema).load('inputs/mini-sa

In [24]: col_names = []
         for i in range(2,len(df.columns)):
             col_names.append(df.columns[i])
         df_features = df.select(*col_names)
         rdd_df = df_features.rdd

In [25]: from pyspark.mllib.stat import Statistics
         coeff = Statistics.corr(rdd_df.map(list),method="pearson")

In [26]: import numpy as np
         np.savetxt('./heatmap-generation/correlation_matrix.txt', coeff)

In [27]: #2.5 min on mini sample data

In [61]: list_corr_pairs_coeffs = []

         for i in range(0,56):
             for j in range(0,56):
                 if (i != j):
                     corr_pair_coeff = []
                     corr_pair_coeff.append(col_names[i]+" , "+col_names[j])
                     corr_pair_coeff.append(float(coeff[i][j]))
                     list_corr_pairs_coeffs.append(tuple(corr_pair_coeff))

In [62]: df_corr_coeff_col_names = []
         df_corr_coeff_col_names.append(StructField("Feature_Pair", StringType(), True))
         df_corr_coeff_col_names.append(StructField("Pearson_Coeff", FloatType(), True))
         df_corr_coeff = spark.createDataFrame(list_corr_pairs_coeffs,StructType(df_corr_coeff_
         sort_coeff_df = df_corr_coeff.sort(df_corr_coeff.Pearson_Coeff.desc())

In [63]: f = open("Feature_pair_sorted_coeff.txt","w")
         sort_coeff_df_list = sort_coeff_df.collect()
         for i in range(0,len(sort_coeff_df_list)):
             f.write(sort_coeff_df_list[i].Feature_Pair + " " + (str)(sort_coeff_df_list[i].Pea
             f.write("\n")
         f.close()
```