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 [ ]:
```
#2.5 min on mini sample data
```