HottestTemperature

November 28, 2017

```
In [116]: from pyspark.sql.types import StructType, StructField, FloatType, LongType, StringTyp
          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 [117]: df = spark.read.format('csv').option('sep', '\t').schema(schema).load('inputs/nam_20
```

```
In [118]: import pyspark.sql.functions as sf
          from pyspark.sql import Column as col
          max_temp = df.select(sf.max(df.temperature_surface).alias("max_temperature_surface"))
          max_temp_list = [float(x.max_temperature_surface) for x in max_temp.collect()]
          print(max_temp_list)
          [row.Geohash for row in df[df.temperature_surface.isin(max_temp_list)].collect()]
```

```
[330.67431640625]
```

```
Out[118]: ['d5f0jqerq27b']
```

```
In [74]: # Creating an SQL 'table'
         df.createOrReplaceTempView("FEATURE_DF")

         # What's the maximum value?
         MaxTempValues = spark.sql("SELECT Geohash,temperature_surface FROM FEATURE_DF WHERE t

         MaxTempValues
```

```
Out[74]: [Row(Geohash='d75zuxsuqtpb', temperature_surface=320.95361328125),
          Row(Geohash='d59d5yttuc5b', temperature_surface=320.95361328125)]
```