4.More on spark.sql

October 11, 2018

1 Important classes of Spark SQL and DataFrames:

```
- :class:`pyspark.sql.SQLContext`
  Main entry point for :class:`DataFrame` and SQL functionality.
- :class:`pyspark.sql.DataFrame`
  A distributed collection of data grouped into named columns.
- :class:`pyspark.sql.Column`
  A column expression in a :class:`DataFrame`.
- :class:`pyspark.sql.Row`
  A row of data in a :class:`DataFrame`.
- :class:`pyspark.sql.HiveContext`
  Main entry point for accessing data stored in Apache Hive.
- :class:`pyspark.sql.GroupedData`
  Aggregation methods, returned by :func:`DataFrame.groupBy`.
- :class:`pyspark.sql.DataFrameNaFunctions`
  Methods for handling missing data (null values).
- :class:`pyspark.sql.DataFrameStatFunctions`
  Methods for statistics functionality.
- :class:`pyspark.sql.functions`
  List of built-in functions available for :class:`DataFrame`.
- :class:`pyspark.sql.types`
  List of data types available.
- :class:`pyspark.sql.Window`
  For working with window functions.
```

```
In [2]: from pyspark import SparkContext
        #sc.stop()
        sc = SparkContext(master="local[3]")

        from pyspark import SparkContext
        from pyspark.sql import *
        sqlContext = SQLContext(sc)
```

1.1 DataframeStatFunctions

Methods for statistics functionality. documented here

- **approxQuantile(col, probabilities, relativeError)** Calculates the approximate quantiles of a numerical column of a DataFrame.
- **corr(col1, col2[, method])** Calculates the correlation of two columns of a DataFrame as a double value.
- **cov(col1, col2)** Calculate the sample covariance for the given columns, specified by their names, as a double value.
- **crosstab(col1, col2)** Computes a pair-wise frequency table of the given columns.
- **freqItems(cols[, support])** Finding frequent items for columns, possibly with false positives.
- **sampleBy(col, fractions[, seed])** Returns a stratified sample without replacement based on the fraction given on each stratum.

```
In [4]: DataFrameStatFunctions.corr?
```