

Chessa Pandu Aditirta
155610037
STMIK AKAKOM YOGYAKARTA
2018

Memberikan perintah seperti yang sudah disiapkan di `examples/src/main/python`

Korelasi

```
Welcome to

      / \_/_/_/_/_/_/_/_/_/_\_/
     / \_/_/_/_/_/_/_/_/_/_\_/
    / \_/_/_/_/_/_/_/_/_/_\_/
   / \_/_/_/_/_/_/_/_/_/_\_/
  / \_/_/_/_/_/_/_/_/_/_\_/
 / \_/_/_/_/_/_/_/_/_/_\_/
/_/_/_/_/_/_/_/_/_/_\_/

version 2.2.0

Using Python version 2.7.13 (default, Jan 19 2017 14:48:08)
SparkSession available as 'spark'.
>>> from pyspark.ml.linalg import Vectors
>>> from pyspark.ml.stat import Correlation
>>> data = [(Vectors.sparse(4, [(0, 1.0), (3, -2.0)])),
... (Vectors.dense([4.0, 5.0, 0.0, 3.0])),
... (Vectors.dense([6.0, 7.0, 0.0, 8.0])),
... (Vectors.sparse(4, [(0, 9.0), (3, 1.0)]))])
>>> df = spark.createDataFrame(data, ["features"])
>>> r1 = Correlation.corr(df, "features").head()
18/01/18 22:58:34 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
18/01/18 22:58:34 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
18/01/18 22:58:34 WARN PearsonCorrelation: Pearson correlation matrix contains NaN values.
>>> print("Pearson correlation matrix:\n" + str(r1[0]))
Pearson correlation matrix:
DenseMatrix([[ 1.,          0.05564149,         nan,  0.40047142],
 [ 0.05564149,  1.,          nan,  0.91359586],
 [          nan,          nan,  1.,          nan],
 [ 0.40047142,  0.91359586,         nan,  1.]])
>>> r2 = Correlation.corr(df, "features", "spearman").head()
18/01/18 22:59:23 WARN PearsonCorrelation: Pearson correlation matrix contains NaN values.
>>> print("Spearman correlation matrix:\n" + str(r2[0]))
Spearman correlation matrix:
DenseMatrix([[ 1.,          0.10540926,         nan,  0.4       ],
 [ 0.10540926,  1.,          nan,  0.9486833 ],
 [          nan,          nan,  1.,          nan],
 [ 0.4        ,  0.9486833 ,         nan,  1.]])
>>>
```

Hypothesis testing

```
>>> from pyspark.ml.linalg import Vectors
>>> from pyspark.ml.stat import ChiSquareTest
>>> data = [(0.0, Vectors.dense(0.5, 10.0)),
... (0.0, Vectors.dense(1.5, 20.0)),
... (1.0, Vectors.dense(1.5, 30.0)),
... (0.0, Vectors.dense(3.5, 30.0)),
... (0.0, Vectors.dense(3.5, 40.0)),
... (1.0, Vectors.dense(3.5, 40.0))]
>>> df = spark.createDataFrame(data, ["label", "features"])
>>> r = ChiSquareTest.test(df, "features", "label").head()
>>> print("pValues: " + str(r.pValues))
pValues: [0.687289278791,0.682270330336]
>>> print("degreesOfFreedom: " + str(r.degreesOfFreedom))
degreesOfFreedom: [2, 3]
>>> print("statistics: " + str(r.statistics))
statistics: [0.75,1.5]
```