Tutorial Apache Spark Mencoba Machine Learning Statistics
PYSPARK

Chessa Pandu Aditirta
155610037
STMIK AKAKOM YOGYAKARTA
2018

Saya Asumsikan Anda telah menginstall python 2 dan apache spark

Memberikan perintah pyspark untuk masuk pada shell spark

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

Korelasi

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]
```