(/U4G66226D/spaces)

nlp_tokens by ars0107

```
%pyspark
from pyspark.ml.feature import Tokenizer
from pyspark.sql.functions import col, udf
from pyspark.sql.types import IntegerType
```

Run

```
%pyspark
# Create sample DataFrame
dataframe = spark.createDataFrame([
    (0, "Spark is great"),
    (1, "We are learning Spark"),
    (2, "Spark is better than hadoop no doubt")
], ["id", "sentence"])
```

Started Juno

```
%pyspark
# Show DataFrame
dataframe.show()
```

```
+---+--------------------+
| id|            sentence|
+---+--------------------+
|  0|      Spark is great|
|  1|We are learning S...|
|  2|Spark is better t...|
+---+--------------------+
```

Interpreter: spark.pyspark. **FINISHED** Took 3 sec 407 millisec. Updated by ars0107 on February 04 2019, 8:35:21 AM (CST)

Run

```
%pyspark
# Show DataFrame
dataframe.show()
# Tokenize word
tokenizer = Tokenizer(inputCol="sentence", outputCol="words")
tokenizer
```

```
+---+--------------------+
| id|            sentence|
+---+--------------------+
|  0|      Spark is great|
|  1|We are learning S...|
|  2|Spark is better t...|
+---+--------------------+

Tokenizer_42d98c9124d75c3dd868
```

(/U4G66226D/spaces)

```
%pyspark
# Show DataFrame
dataframe.show()
# Transform and show DataFrame
tokenized = tokenizer.transform(dataframe)
tokenized.show(truncate=False)
```

```
+---+--------------------+
| id|            sentence|
+---+--------------------+
|  0|      Spark is great|
|  1|We are learning S...|
|  2|Spark is better t...|
+---+--------------------+

+---+-----------------------------------+-------------------------------------------+
|id |sentence                           |words                                      |
+---+-----------------------------------+-------------------------------------------+
|0  |Spark is great                     |[spark, is, great]                         |
|1  |We are learning Spark              |[we, are, learning, spark]                 |
|2  |Spark is better than hadoop no doubt|[spark, is, better, than, hadoop, no, doubt]|
+---+-----------------------------------+-------------------------------------------+
```

Run