(/U4G66226D/spaces)

udf by ars0107

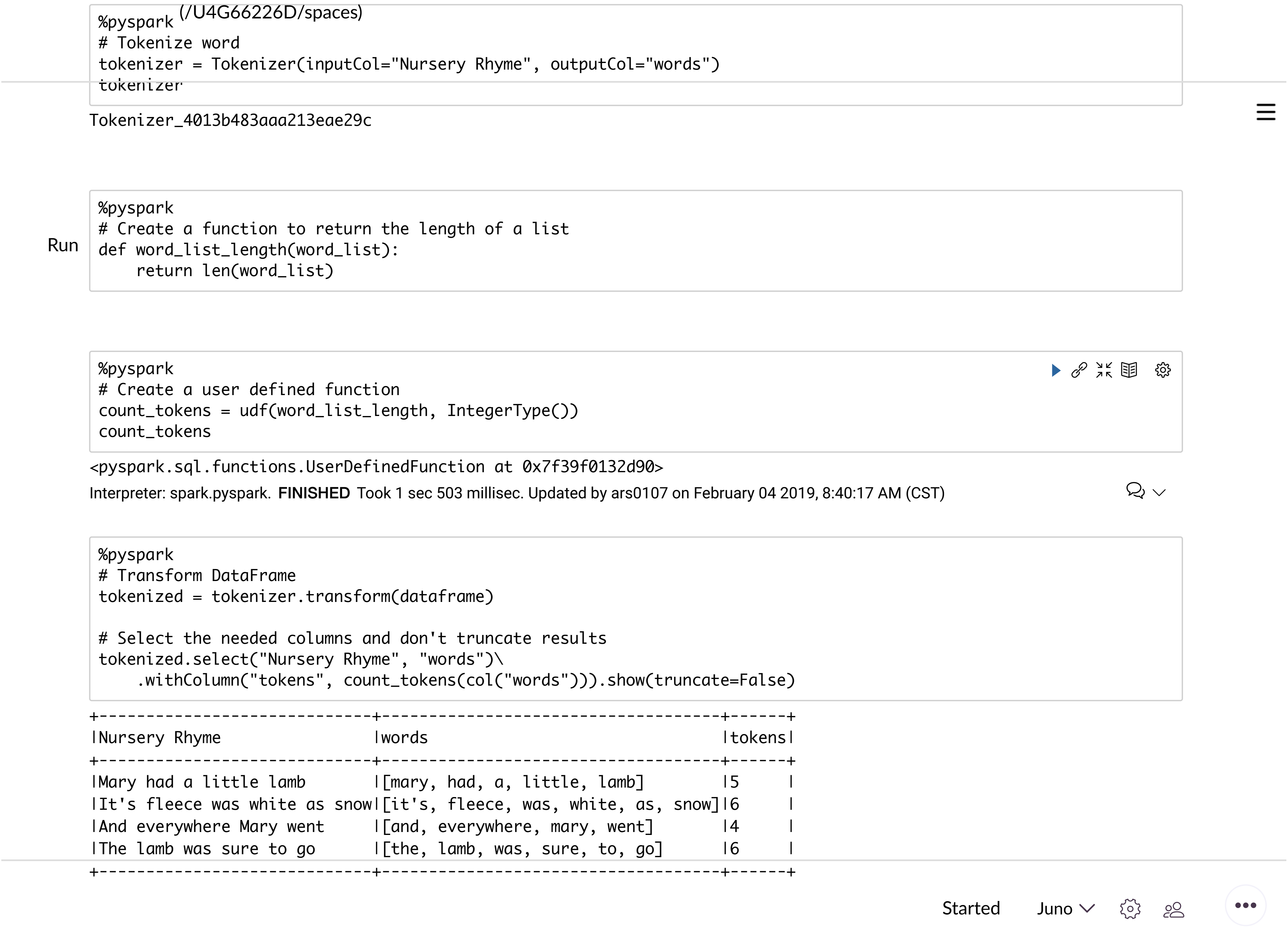
Run

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

```
%pyspark
dataframe = spark.createDataFrame([
    (0, "Mary had a little lamb"),
    (1, "It's fleece was white as snow"),
    (2, "And everywhere Mary went"),
    (3, "The lamb was sure to go")
], ["id", "Nursery Rhyme"])
dataframe.show()
```

```
+---+--------------------+
| id|       Nursery Rhyme|
+---+--------------------+
|  0|Mary had a little...|
|  1|It's fleece was w...|
|  2|And everywhere Ma...|
|  3|The lamb was sure...|
+---+--------------------+
```

Started Juno



```
%pyspark
# Tokenize word
tokenizer = Tokenizer(inputCol="Nursery Rhyme", outputCol="words")
tokenizer
```

```
Tokenizer_4013b483aaa213eae29c
```

```
%pyspark
# Create a function to return the length of a list
def word_list_length(word_list):
    return len(word_list)
```

```
%pyspark
# Create a user defined function
count_tokens = udf(word_list_length, IntegerType())
count_tokens
```

```
<pyspark.sql.functions.UserDefinedFunction at 0x7f39f0132d90>
```

Interpreter: spark.pyspark. **FINISHED** Took 1 sec 503 millisec. Updated by ars0107 on February 04 2019, 8:40:17 AM (CST)

```
%pyspark
# Transform DataFrame
tokenized = tokenizer.transform(dataframe)

# Select the needed columns and don't truncate results
tokenized.select("Nursery Rhyme", "words")\
    .withColumn("tokens", count_tokens(col("words"))).show(truncate=False)
```

```
+-----------------------------+-------------------------------------+------+
|Nursery Rhyme                |words                                |tokens|
+-----------------------------+-------------------------------------+------+
|Mary had a little lamb       |[mary, had, a, little, lamb]         |5     |
|It's fleece was white as snow|[it's, fleece, was, white, as, snow]|6     |
|And everywhere Mary went     |[and, everywhere, mary, went]        |4     |
|The lamb was sure to go      |[the, lamb, was, sure, to, go]       |6     |
+-----------------------------+-------------------------------------+------+
```

(/U4G66226D/spaces)

Run

Started Juno