Tokenizing

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

```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://s3.amazonaws.com/dataviz-curriculum/day_2/data.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("data.csv"), sep=",", header=True)

# Show DataFrame
df.show()
```

```
+--------------------+
|                Poem|
+--------------------+
|This Autumn midnight|
|Orion’s at my window|
|shouting for his ...|
+--------------------+
```

Interpreter: spark.pyspark. **FINISHED** Took 995 millisec. Updated by ars0107 on February 04 2019, 8:45:08 AM (CST)

```
%pyspark
# Tokenize DataFrame
tokened = Tokenizer(inputCol="Poem", outputCol="words")
```

```
%pyspark
# Transform DataFrame
tokenized = tokened.transform(df)
tokenized.show()
```

```
+--------------------+--------------------+
|                Poem|               words|
+--------------------+--------------------+
|This Autumn midnight|[this, autumn, mi...|
|Orion’s at my window|[orion’s, at, my,...|
|shouting for his ...|[shouting, for, h...|
+--------------------+--------------------+
```

```
%pyspark
# Create a Function to count vowels
def vowel_counter(words):
    vowel_count = 0

    for word in words:
        for vowel in word:
            if vowel in ('a', 'e', 'i', 'o', 'u'):
                vowel_count += 1

    return vowel_count
```

```
%pyspark
# Store a user defined function
count_vowels = udf(vowel_counter, IntegerType())
count_vowels
```

```
<pyspark.sql.functions.UserDefinedFunction at 0x7f1fd48acbd0>
```

```
%pyspark
# Create new DataFrame with the udf
tokenized.select("Poem", "words")\
    .withColumn("vowels", count_vowels(col("words"))).show(truncate=False)
```

```
+--------------------+-------------------------+------+
|Poem                |words                    |vowels|
+--------------------+-------------------------+------+
|This Autumn midnight |[this, autumn, midnight]  |6     |
|Orion’s at my window |[orion’s, at, my, window] |6     |
|shouting for his dog.|[shouting, for, his, dog.]|6     |
+--------------------+-------------------------+------+
```