(/U4G66226D/spaces)

nlp_hashingTF by ars0107

```
%pyspark
from pyspark.ml.feature import HashingTF, IDF, Tokenizer
```

```
%pyspark
# Sample DataFrame with repeating words
dataframe = spark.createDataFrame([
    (0, "The cow cow jumped and jumped cow"),
    (1, "then the cow said"),
    (2, "I am a cow that jumped")
],["id", "words"])

dataframe.show()
```

```
+---+--------------------+
| id|               words|
+---+--------------------+
|  0|The cow cow jumpe...|
|  1|   then the cow said|
|  2|I am a cow that j...|
+---+--------------------+
```

Run

Started Juno

(/U4G66226D/spaces)

```
%pyspark
# Tokenize the words
tokenizer = Tokenizer(inputCol="words", outputCol="tokens")
wordsData = tokenizer.transform(dataframe)
wordsData.show()
```

```
+---+--------------------+--------------------+
| id|               words|              tokens|
+---+--------------------+--------------------+
|  0|The cow cow jumpe...|[the, cow, cow, j...|
|  1|   then the cow said|[then, the, cow, ...|
|  2|I am a cow that j...|[i, am, a, cow, t...|
+---+--------------------+--------------------+
```

```
%pyspark
# Run the hashing term frequency
hashing = HashingTF(inputCol="tokens", outputCol="hashedValues", numFeatures=pow(2,4))

# Transform into a DF
hashed_df = hashing.transform(wordsData)
```

Run

Started

Juno

(/U4G66226D/spaces)

```
%pyspark
# Display new DataFrame
hashed_df.show(truncate=False)
```

```
+---+--------------------------------+-----------------------------------------+-----------------------------------
-----------+
|id |words                           |tokens                                   |hashedValues
           |
+---+--------------------------------+-----------------------------------------+-----------------------------------
-----------+
|0  |The cow cow jumped and jumped cow|[the, cow, cow, jumped, and, jumped, cow]|(16,[11,13,14,15],[2.0,1.0,1.0,3.
0])          |
|1  |then the cow said               |[then, the, cow, said]                   |(16,[0,13,14,15],[1.0,1.0,1.0,1.0])
           |
|2  |I am a cow that jumped          |[i, am, a, cow, that, jumped]            |(16,[0,1,2,5,11,15],[1.0,1.0,1.0,1.
0,1.0,1.0])|
+---+--------------------------------+-----------------------------------------+-----------------------------------
-----------+
```

```
%pyspark
# Fit the IDF on the data set
idf = IDF(inputCol="hashedValues", outputCol="features")
idfModel = idf.fit(hashed_df)
rescaledData = idfModel.transform(hashed_df)
```

Run

Started

Juno

(/U4G66226D/spaces)

```
%pyspark
# Display the DataFrame
rescaledData.select("words", "features").show(truncate=False)
```

```
+---------------------------------+---------------------------------------------------------------------------------
------------------------------------------+
|words                            |features
                                          |
+---------------------------------+---------------------------------------------------------------------------------
------------------------------------------+
|The cow cow jumped and jumped cow|(16,[11,13,14,15],[0.5753641449035617,0.28768207245178085,0.28768207245178085,0.
0])                                        |
|then the cow said                |(16,[0,13,14,15],[0.28768207245178085,0.28768207245178085,0.28768207245178085,0.
0])                                        |
|I am a cow that jumped           |(16,[0,1,2,5,11,15],[0.28768207245178085,0.6931471805599453,0.6931471805599453,0.
6931471805599453,0.28768207245178085,0.0])|
+---------------------------------+---------------------------------------------------------------------------------
------------------------------------------+
```

Run

Started

Juno

•••