(/U4G66226D/spaces)

airline_hashing by ars0107

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

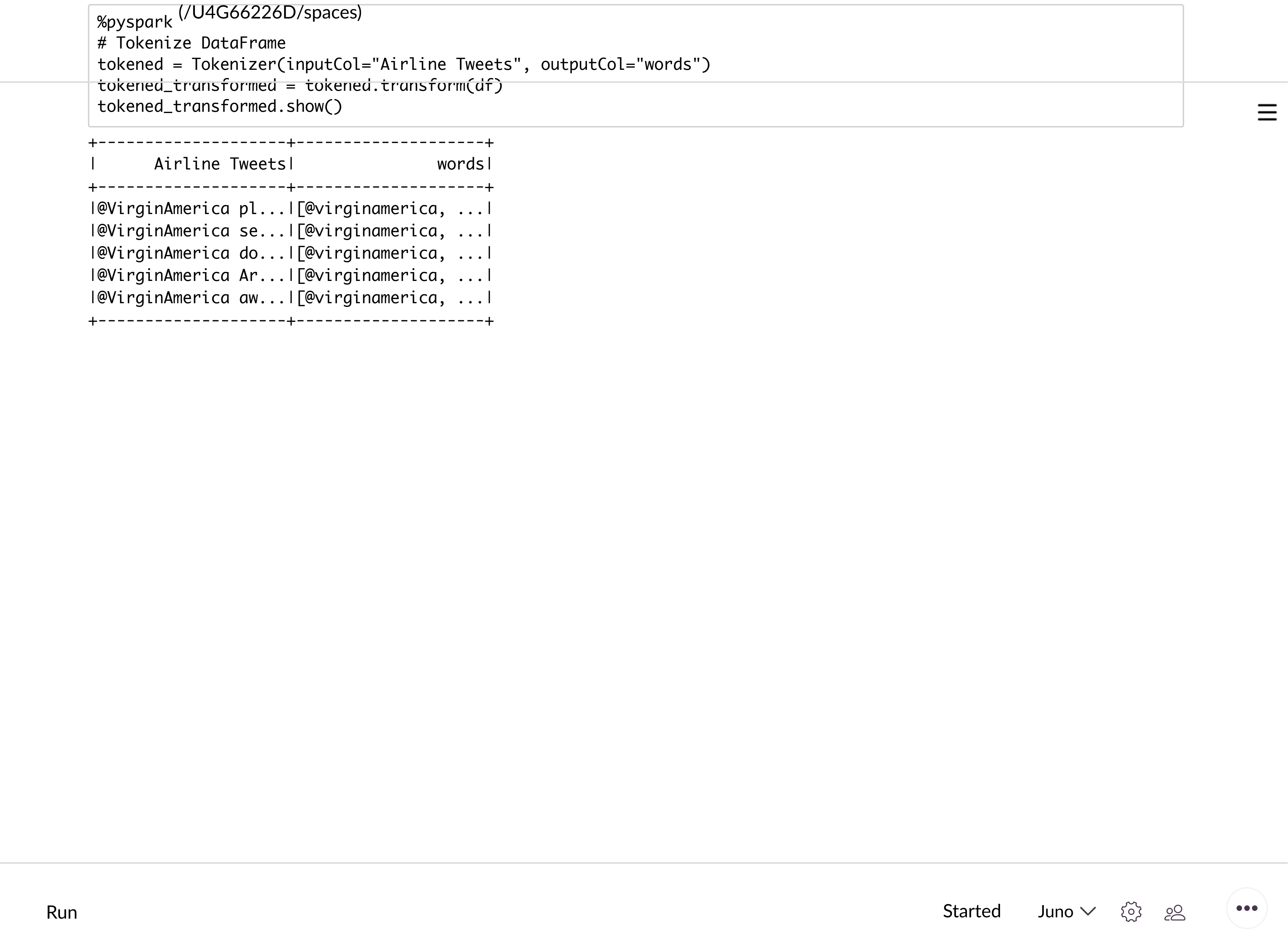
```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://s3.amazonaws.com/dataviz-curriculum/day_2/airlines.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("airlines.csv"), sep=",", header=True)

# Show DataFrame
df.show()
```

```
+--------------------+
|      Airline Tweets|
+--------------------+
|@VirginAmerica pl...|
|@VirginAmerica se...|
|@VirginAmerica do...|
|@VirginAmerica Ar...|
|@VirginAmerica aw...|
+--------------------+
```

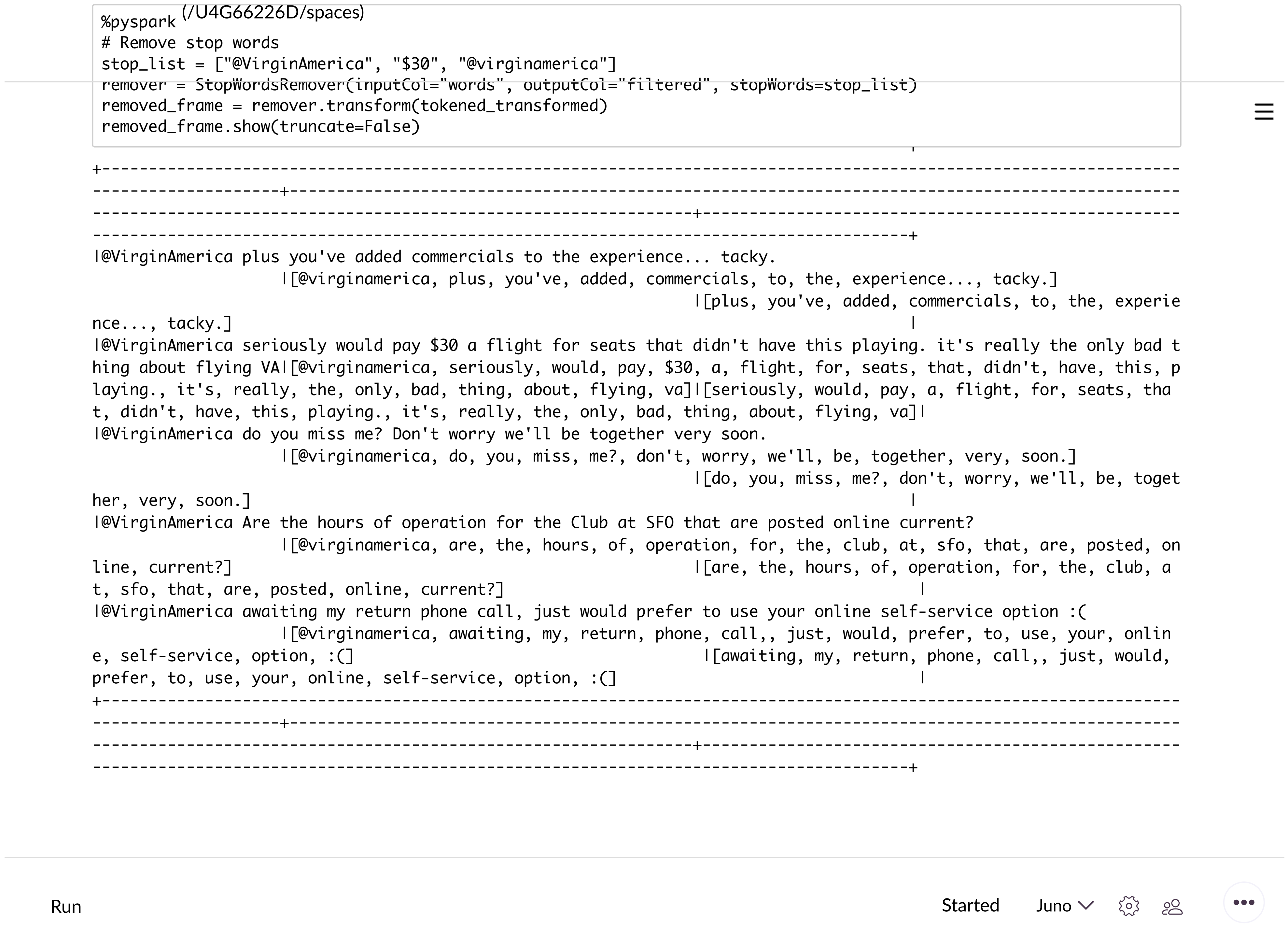
Interpreter: spark.pyspark. **FINISHED** Took 7 sec 551 millisec. Updated by ars0107 on February 04 2019, 9:01:00 AM (CST)

Run Started Juno



```
%pyspark
# Tokenize DataFrame
tokened = Tokenizer(inputCol="Airline Tweets", outputCol="words")
tokened_transformed = tokened.transform(df)
tokened_transformed.show()
```

```
+--------------------+--------------------+
|      Airline Tweets|               words|
+--------------------+--------------------+
|@VirginAmerica pl...|[@virginamerica, ...|
|@VirginAmerica se...|[@virginamerica, ...|
|@VirginAmerica do...|[@virginamerica, ...|
|@VirginAmerica Ar...|[@virginamerica, ...|
|@VirginAmerica aw...|[@virginamerica, ...|
+--------------------+--------------------+
```



```
%pyspark
# Remove stop words
stop_list = ["@VirginAmerica", "$30", "@virginamerica"]
remover = StopWordsRemover(inputCol="words", outputCol="filtered", stopWords=stop_list)
removed_frame = remover.transform(tokened_transformed)
removed_frame.show(truncate=False)
```

```
                                                                                      |
+------------------------------------------------------------------------------------------------------------------
--------------------+------------------------------------------------------------------------------------------------
----------------------------------------------------------------+--------------------------------------------------
--------------------------------------------------------------------------------------+
|@VirginAmerica plus you've added commercials to the experience... tacky.
                    |[@virginamerica, plus, you've, added, commercials, to, the, experience..., tacky.]
                                                                |[plus, you've, added, commercials, to, the, experie
nce..., tacky.]                                                                       |
|@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad t
hing about flying VA|[@virginamerica, seriously, would, pay, $30, a, flight, for, seats, that, didn't, have, this, p
laying., it's, really, the, only, bad, thing, about, flying, va]|[seriously, would, pay, a, flight, for, seats, tha
t, didn't, have, this, playing., it's, really, the, only, bad, thing, about, flying, va]|
|@VirginAmerica do you miss me? Don't worry we'll be together very soon.
                    |[@virginamerica, do, you, miss, me?, don't, worry, we'll, be, together, very, soon.]
                                                                |[do, you, miss, me?, don't, worry, we'll, be, toget
her, very, soon.]                                                                     |
|@VirginAmerica Are the hours of operation for the Club at SFO that are posted online current?
                    |[@virginamerica, are, the, hours, of, operation, for, the, club, at, sfo, that, are, posted, on
line, current?]                                                 |[are, the, hours, of, operation, for, the, club, a
t, sfo, that, are, posted, online, current?]                                           |
|@VirginAmerica awaiting my return phone call, just would prefer to use your online self-service option :(
                    |[@virginamerica, awaiting, my, return, phone, call,, just, would, prefer, to, use, your, onlin
e, self-service, option, :(]                                     |[awaiting, my, return, phone, call,, just, would,
prefer, to, use, your, online, self-service, option, :(]                               |
+------------------------------------------------------------------------------------------------------------------
--------------------+------------------------------------------------------------------------------------------------
----------------------------------------------------------------+--------------------------------------------------
--------------------------------------------------------------------------------------+
```

(/U4G66226D/spaces)

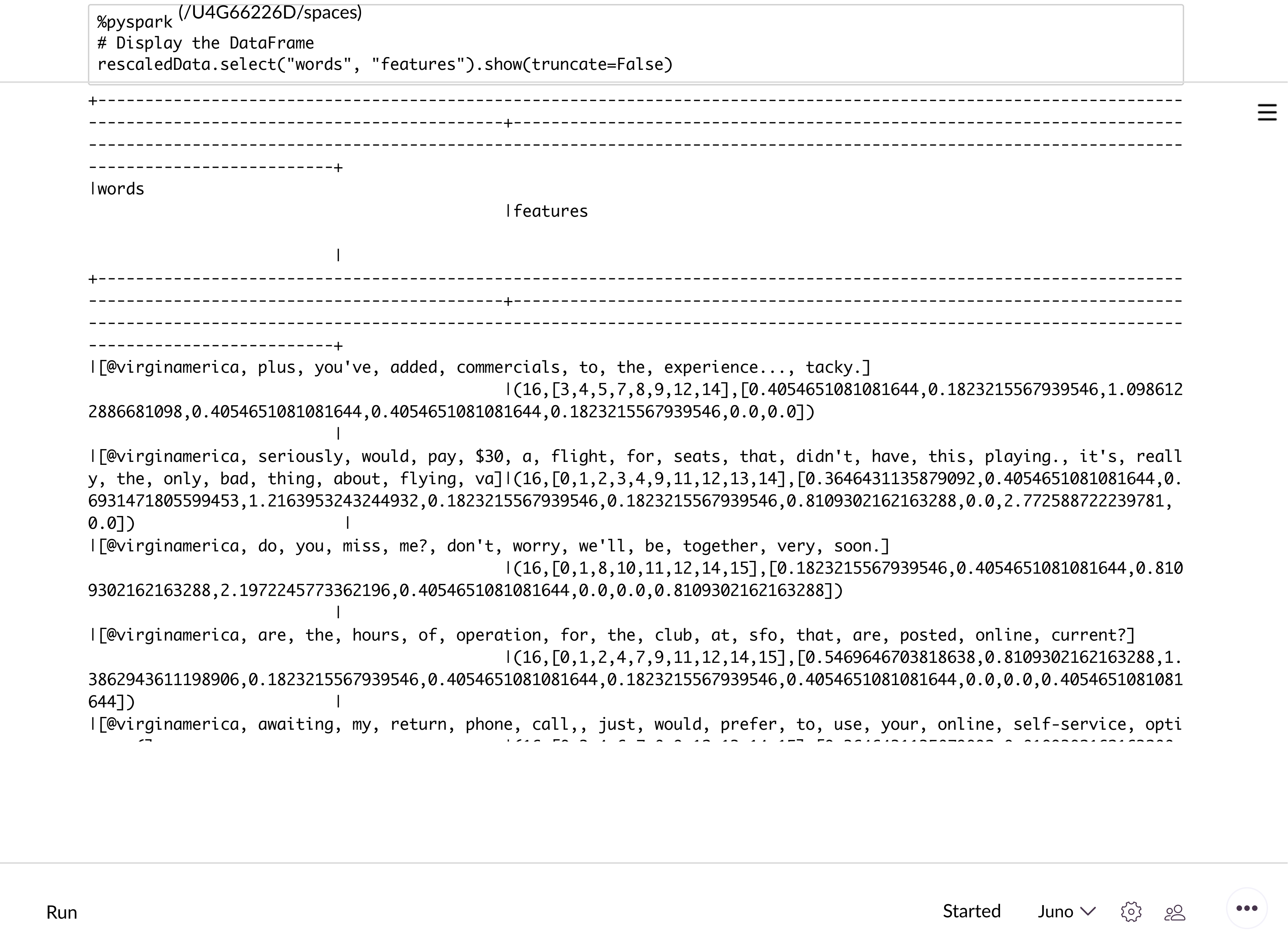
```
%pyspark
# Run the hashing term frequency
hashing = HashingTF(inputCol="filtered", outputCol="hashedValues", numFeatures=pow(2,4))

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

```
+--------------------+--------------------+--------------------+--------------------+
|      Airline Tweets|               words|            filtered|        hashedValues|
+--------------------+--------------------+--------------------+--------------------+
|@VirginAmerica pl...|[@virginamerica, ...|[plus, you've, ad...|(16,[3,4,5,7,8,9,...|
|@VirginAmerica se...|[@virginamerica, ...|[seriously, would...|(16,[0,1,2,3,4,9,...|
|@VirginAmerica do...|[@virginamerica, ...|[do, you, miss, m...|(16,[0,1,8,10,11,...|
|@VirginAmerica Ar...|[@virginamerica, ...|[are, the, hours,...|(16,[0,1,2,4,7,9,...|
|@VirginAmerica aw...|[@virginamerica, ...|[awaiting, my, re...|(16,[0,3,4,6,7,8,...|
+--------------------+--------------------+--------------------+--------------------+
```

```
%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

                          |
+------------------------------------------------------------------------------------------------------------------
--------------------------------------------+---------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------
--------------------------+
|[@virginamerica, plus, you've, added, commercials, to, the, experience..., tacky.]
                                            |(16,[3,4,5,7,8,9,12,14],[0.4054651081081644,0.1823215567939546,1.098612
2886681098,0.4054651081081644,0.4054651081081644,0.1823215567939546,0.0,0.0])
                          |
|[@virginamerica, seriously, would, pay, $30, a, flight, for, seats, that, didn't, have, this, playing., it's, reall
y, the, only, bad, thing, about, flying, va]|(16,[0,1,2,3,4,9,11,12,13,14],[0.3646431135879092,0.4054651081081644,0.
6931471805599453,1.2163953243244932,0.1823215567939546,0.1823215567939546,0.8109302162163288,0.0,2.772588722239781,
0.0])                      |
|[@virginamerica, do, you, miss, me?, don't, worry, we'll, be, together, very, soon.]
                                            |(16,[0,1,8,10,11,12,14,15],[0.1823215567939546,0.4054651081081644,0.810
9302162163288,2.1972245773362196,0.4054651081081644,0.0,0.0,0.8109302162163288])
                          |
|[@virginamerica, are, the, hours, of, operation, for, the, club, at, sfo, that, are, posted, online, current?]
                                            |(16,[0,1,2,4,7,9,11,12,14,15],[0.5469646703818638,0.8109302162163288,1.
3862943611198906,0.1823215567939546,0.4054651081081644,0.1823215567939546,0.4054651081081644,0.0,0.0,0.4054651081081
644])                     |
|[@virginamerica, awaiting, my, return, phone, call,, just, would, prefer, to, use, your, online, self-service, opti
```