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

Interpreter: spark.pyspark. FINISHED Took 111 millisec. Updated by bcheek on February 01 2020, 9:39:30 AM (CST)

```
%spark.pyspark

from pyspark import SparkFiles

sc.addFile('http://zdata/CleanTrumpTweets.csv')
df = spark.read.format('csv').options(delimiter=',', header='true', inferSchema='true').load(SparkFiles.get('CleanTrumpTweets.csv'))
df.show(5)
```

```
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+
|            source|                text|     created_at|retweet_count|favorite_count|is_retweet|     id_str|         tidy_tweets|absolute_tidy_tweets|          tweetdate|
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+
|Twitter for iPhone|The United States...|1/1/2018 12:12|     101056.0|      304676.0|      null|9.47803E+17|The United States...|The United States...|2018-01-01 00:00:00|
|Twitter for iPhone|Iran is failing a...|1/1/2018 12:44|      29046.0|      111467.0|      null|9.47811E+17|Iran is failing a...|Iran is failing a...|2018-01-01 00:00:00|
|Twitter for iPhone|Will be leaving F...|1/1/2018 13:37|      16884.0|      114754.0|      null|9.47824E+17|Will be leaving F...|Will be leaving F...|2018-01-01 00:00:00|
|Twitter for iPhone|The people of Ira...|1/2/2018 12:09|      28227.0|      105965.0|      null|9.48164E+17|The people of Ira...|The people of Ira...|2018-01-02 00:00:00|
|Twitter for iPhone|Crooked Hillary C...|1/2/2018 12:48|      37561.0|      130933.0|      null|9.48174E+17|Crooked Hillary C...|Crooked Hillary C...|2018-01-02 00:00:00|
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. FINISHED Took 20 sec 674 millisec. Updated by bcheek on February 01 2020, 10:37:20 AM (CST)

```
%pyspark
# Tokenize DataFrame
review_data = Tokenizer(inputCol="absolute_tidy_tweets", outputCol="Words")
```

Interpreter: spark.pyspark. FINISHED Took 114 millisec. Updated by bcheek on January 31 2020, 11:46:54 PM (CST)

```
%pyspark
# Transform DataFrame
reviewed = review_data.transform(df)
reviewed.show(5)
```

```
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+--------------------+
|            source|                text|     created_at|retweet_count|favorite_count|is_retweet|     id_str|         tidy_tweets|absolute_tidy_tweets|          tweetdate|               Words|
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+--------------------+
|Twitter for iPhone|The United States...|1/1/2018 12:12|     101056.0|      304676.0|      null|9.47803E+17|The United States...|The United States...|2018-01-01 00:00:00|[the, united, sta...|
|Twitter for iPhone|Iran is failing a...|1/1/2018 12:44|      29046.0|      111467.0|      null|9.47811E+17|Iran is failing a...|Iran is failing a...|2018-01-01 00:00:00|[iran, is, failin...|
|Twitter for iPhone|Will be leaving F...|1/1/2018 13:37|      16884.0|      114754.0|      null|9.47824E+17|Will be leaving F...|Will be leaving F...|2018-01-01 00:00:00|[will, be, leavin...|
|Twitter for iPhone|The people of Ira...|1/2/2018 12:09|      28227.0|      105965.0|      null|9.48164E+17|The people of Ira...|The people of Ira...|2018-01-02 00:00:00|[the, people, of,...|
|Twitter for iPhone|Crooked Hillary C...|1/2/2018 12:48|      37561.0|      130933.0|      null|9.48174E+17|Crooked Hillary C...|Crooked Hillary C...|2018-01-02 00:00:00|[crooked, hillary...|
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+--------------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. FINISHED Took 419 millisec. Updated by bcheek on January 31 2020, 11:33:30 PM (CST)

```
%pyspark
# Remove stop words
remover = StopWordsRemover(inputCol="Words", outputCol="filtered")
```

Interpreter: spark.pyspark. FINISHED Took 215 millisec. Updated by bcheek on January 31 2020, 11:33:34 PM (CST)

```
%pyspark
# Transform new DataFrame
newFrame = remover.transform(reviewed)
newFrame.show(5)
```

```
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+--------------------
+--------------------+
|            source|                text|     created_at|retweet_count|favorite_count|is_retweet|     id_str|         tidy_tweets|absolute_tidy_tweets|          tweetdate|               Words|
filtered|
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+--------------------
+--------------------+
|Twitter for iPhone|The United States...|1/1/2018 12:12|     101056.0|      304676.0|      null|9.47803E+17|The United States...|The United States...|2018-01-01 00:00:00|[the, united, sta...|[uni
ted, states, ...|
|Twitter for iPhone|Iran is failing a...|1/1/2018 12:44|      29046.0|      111467.0|      null|9.47811E+17|Iran is failing a...|Iran is failing a...|2018-01-01 00:00:00|[iran, is, failin...|[ira
n, failing, e...|
|Twitter for iPhone|Will be leaving F...|1/1/2018 13:37|      16884.0|      114754.0|      null|9.47824E+17|Will be leaving F...|Will be leaving F...|2018-01-01 00:00:00|[will, be, leavin...|[lea
ving, florida...|
|Twitter for iPhone|The people of Ira...|1/2/2018 12:09|      28227.0|      105965.0|      null|9.48164E+17|The people of Ira...|The people of Ira...|2018-01-02 00:00:00|[the, people, of,...|[peo
ple, iran, fi...|
|Twitter for iPhone|Crooked Hillary C...|1/2/2018 12:48|      37561.0|      130933.0|      null|9.48174E+17|Crooked Hillary C...|Crooked Hillary C...|2018-01-02 00:00:00|[crooked, hillary...|[cro
oked, hillary...|
+------------------+--------------------+---------------+-------------+--------------+----------+-----------+--------------------+--------------------+-------------------+--------------------
+--------------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. FINISHED Took 418 millisec. Updated by bcheek on January 31 2020, 11:33:40 PM (CST)

```
%pyspark
# Show simplified review
newFrame.select("filtered").show(truncate=False)
```

```
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|filtered
|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|[united, states, foolishly, given, pakistan, , billion, dollars, aid, last, , years, given, us, nothing, lies, amp, deceit, thinking, leaders, fools, give, safe, haven, terrorists, hunt, afgh
anistan, little, help]                                                               |
|[iran, failing, every, level, despite, terrible, deal, made, obama, administration, great, iranian, people, repressed, many, years, hungry, food, amp, freedom, along, human, rights, wealth, i
ran, looted, time, fo, change]                                                       |
|[leaving, florida, washington, dc, today, , pm, much, work, done, great, new, year]
|
|[people, iran, finally, acting, brutal, corrupt, iranian, regime, money, president, obama, foolishly, gave, went, terrorism, pockets, people, little, food, big, inflation, human, rights, us,
watching]                                                                            |
|[crooked, hillary, clintons, top, aid, huma, abedin, accused, disregarding, basic, security, protocols, put, classified, passwords, hands, foreign, agentsemember, sailors, pictures, submarin
e, jail, deep, state, justice, dept, must, finally, act, also, comey, amp, others]|
|[thank, brandon, judd, national, border, patrol, council, kind, words, well, border, bringing, amp, great, folks, build, desperately, needed, wall, foxandfriends]
|
|[companies, giving, big, bonuses, workers, tax, cut, billeally, great]
|
|[sanctions, pressures, beginning, big, impact, north, korea, soldiers, dangerously, fleeing, south, koreaocket, man, wants, talk, south, korea, first, time, perhaps, good, news, perhaps, , se
e]                                                                                   |
|[since, taking, office, strict, commercial, aviation, good, news, , reported, zero, deaths, , best, safest, year, record]
|
|[failing, new, york, times, new, publisher, ag, sulzberger, congratulations, last, chance, times, fulfill, vision, founder, adolph, ochs, give, news, impartially, without, fear, favo, regardl
```

```
|[failing, new, york, times, new, publisher, ag, sulzberger, congratulations, last, chance, times, fulfill, vision, founder, adolph, ochs, give, news, impartially, without, fear, favo, regardl
ess, party, sect, interests, involved, get]                                                |
|[impartial, journalists, much, higher, standard, lose, phony, nonexistent, sources, treat, president, united, states, faily, next, time, people, win, wont, write, apology, readers, job, poorl
y, done, gl]                                                                               |
|[democrats, nothing, daca, , interested, politics, daca, activists, hispanics, go, hard, dems, start, falling, love, withepublicans, president, aboutesults]
|
```

Interpreter: spark.pyspark. FINISHED Took 319 millisec. Updated by bcheek on January 31 2020, 11:33:48 PM (CST)

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

# Transform into a DF
hashed_df = hashing.transform(newFrame)
hashed_df.show(5)
```

```
+------------------+--------------------+-------------+-------------+--------------+-----------+-----------+--------------------+--------------------+-------------------+--------------------
+---------------+--------------------+
|            source|                text|   created_at|retweet_count|favorite_count|is_retweet|     id_str|         tidy_tweets|absolute_tidy_tweets|          tweetdate|               Words|
filtered|        hashedValues|
+------------------+--------------------+-------------+-------------+--------------+-----------+-----------+--------------------+--------------------+-------------------+--------------------
+---------------+--------------------+
|Twitter for iPhone|The United States...|1/1/2018 12:12|     101056.0|      304676.0|      null|9.47803E+17|The United States...|The United States...|2018-01-01 00:00:00|[the, united, sta...|[uni
ted, states, ...|(16,[0,2,3,4,5,6,...|
|Twitter for iPhone|Iran is failing a...|1/1/2018 12:44|      29046.0|      111467.0|      null|9.47811E+17|Iran is failing a...|Iran is failing a...|2018-01-01 00:00:00|[iran, is, failin...|[ira
n, failing, e...|(16,[0,2,3,4,5,6,...|
|Twitter for iPhone|Will be leaving F...|1/1/2018 13:37|      16884.0|      114754.0|      null|9.47824E+17|Will be leaving F...|Will be leaving F...|2018-01-01 00:00:00|[will, be, leavin...|[lea
ving, florida...|(16,[0,2,3,4,5,6,...|
|Twitter for iPhone|The people of Ira...|1/2/2018 12:09|      28227.0|      105965.0|      null|9.48164E+17|The people of Ira...|The people of Ira...|2018-01-02 00:00:00|[the, people, of,...|[peo
ple, iran, fi...|(16,[0,2,3,4,6,7,...|
|Twitter for iPhone|Crooked Hillary C...|1/2/2018 12:48|      37561.0|      130933.0|      null|9.48174E+17|Crooked Hillary C...|Crooked Hillary C...|2018-01-02 00:00:00|[crooked, hillary...|[cro
oked, hillary...|(16,[0,1,2,3,4,5,...|
+------------------+--------------------+-------------+-------------+--------------+-----------+-----------+--------------------+--------------------+-------------------+--------------------
+---------------+--------------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. FINISHED Took 517 millisec. Updated by bcheek on January 31 2020, 11:34:07 PM (CST)

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

```
Py4JJavaErrorTraceback (most recent call last)
<ipython-input-43-bb59b99c93f3> in <module>()
      1 # Fit the IDF on the data set
      2 idf = IDF(inputCol="hashedValues", outputCol="features")
----> 3 idfModel = idf.fit(hashed_df)
      4 rescaledData = idfModel.transform(hashed_df)

/usr/zepl/spark-2.2/python/lib/pyspark.zip/pyspark/ml/base.py in fit(self, dataset, params)
     62                 return self.copy(params)._fit(dataset)
     63             else:
---> 64                 return self._fit(dataset)
     65         else:
     66             raise ValueError("Params must be either a param map or a list/tuple of param maps, "

/usr/zepl/spark-2.2/python/lib/pyspark.zip/pyspark/ml/wrapper.py in _fit(self, dataset)
    263
    264     def _fit(self, dataset):
--> 265         java_model = self._fit_java(dataset)
    266         return self._create_model(java_model)
    267

/usr/zepl/spark-2.2/python/lib/pyspark.zip/pyspark/ml/wrapper.py in _fit_java(self, dataset)
    260         """
    261         self._transfer_params_to_java()
--> 262         return self._java_obj.fit(dataset._jdf)
    263
    264     def _fit(self, dataset):

/usr/zepl/spark-2.2/python/lib/py4j-0.10.4-src.zip/py4j/java_gateway.py in __call__(self, *args)
```

Interpreter: spark.pyspark. ERROR Took 417 millisec. Updated by bcheek on January 31 2020, 11:47:32 PM (CST)

```
%pyspark
```

Interpreter: spark.pyspark. FINISHED Took 116 millisec. Updated by bcheek on January 31 2020, 11:47:08 PM (CST)

```
%pyspark
# Show simplified review
hashed_df.select("hashedValues").show(truncate=False)
```

```
+-------------------------------------------------------------------------------------+
|hashedValues                                                                         |
+-------------------------------------------------------------------------------------+
|(16,[0,2,3,4,5,6,7,9,10,11,12,13,14,15],[1.0,1.0,1.0,2.0,1.0,3.0,3.0,3.0,3.0,2.0,4.0,2.0,2.0,1.0])|
|(16,[0,2,3,4,5,6,8,9,10,11,12,13,14,15],[2.0,1.0,1.0,2.0,2.0,3.0,3.0,4.0,2.0,1.0,2.0,2.0,3.0,1.0])|
|(16,[0,2,3,4,5,6,7,8,9,12],[1.0,1.0,2.0,2.0,1.0,1.0,1.0,1.0,1.0,2.0])                |
|(16,[0,2,3,4,6,7,8,9,10,11,13,14,15],[2.0,2.0,2.0,1.0,1.0,1.0,4.0,2.0,2.0,2.0,1.0,4.0,1.0])    |
|(16,[0,1,2,3,4,5,6,7,9,10,13,14,15],[3.0,2.0,3.0,2.0,3.0,3.0,4.0,3.0,1.0,4.0,2.0,2.0,1.0])    |
|(16,[0,3,4,5,6,7,8,9,10,11,12,13,15],[2.0,1.0,2.0,1.0,1.0,1.0,2.0,1.0,2.0,1.0,2.0,1.0,3.0])   |
|(16,[0,4,5,6,9,12],[2.0,1.0,3.0,1.0,1.0,1.0])                                        |
|(16,[1,2,5,6,7,8,11,12,13,14,15],[2.0,2.0,1.0,6.0,2.0,4.0,1.0,1.0,2.0,3.0,1.0])      |
|(16,[0,1,3,5,6,7,8,9,11,12,14],[2.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,2.0,2.0,1.0])        |
|(16,[1,2,3,4,5,7,8,9,11,12,13,14],[2.0,3.0,1.0,3.0,4.0,3.0,1.0,3.0,3.0,2.0,3.0,1.0]) |
|(16,[1,2,3,5,6,7,8,9,11,12,13,14],[2.0,1.0,2.0,4.0,2.0,2.0,1.0,3.0,1.0,3.0,3.0,2.0]) |
|(16,[0,4,5,7,8,9,12,13,15],[1.0,1.0,3.0,1.0,3.0,1.0,4.0,2.0,2.0])                    |
|(16,[0,2,4,7,8,9,10,11,13,14],[3.0,1.0,4.0,1.0,1.0,1.0,2.0,1.0,3.0,2.0])             |
|(16,[0,1,2,4,6,7,9,12,13,14,15],[2.0,1.0,1.0,1.0,2.0,3.0,3.0,2.0,2.0,2.0,2.0])       |
|(16,[0,3,4,6,7,8,9,11,12,13,14,15],[2.0,1.0,2.0,3.0,2.0,1.0,2.0,7.0,1.0,1.0,2.0,1.0])|
|(16,[1,3,4,5,6,7,10,11,12,13,14],[2.0,1.0,3.0,2.0,1.0,2.0,1.0,1.0,2.0,2.0,3.0])      |
|(16,[0,1,2,4,5,7,8,9,10,11,12,13,14],[2.0,1.0,3.0,3.0,1.0,1.0,3.0,1.0,1.0,3.0,3.0,1.0,4.0])   |
|(16,[1,2,4,5,6,7,8,10,11,12,14,15],[3.0,1.0,4.0,1.0,2.0,2.0,1.0,2.0,2.0,2.0,2.0,1.0])|
|(16,[0,1,3,4,6,8,9,10,12,13,14,15],[2.0,3.0,7.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])|
|(16,[2,4,6,7,8,9,10,11,13,14,15],[1.0,1.0,2.0,2.0,1.0,1.0,1.0,1.0,1.0,3.0,1.0])      |
+-------------------------------------------------------------------------------------+
only showing top 20 rows
```

Interpreter: spark.pyspark. FINISHED Took 165 millisec. Updated by bcheek on January 31 2020, 10:56:12 PM (CST)

Interpreter: spark.