stu_nlp_stopwords by ars0107

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

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

# Show DataFrame
df.show()
```

```
+--------------------+
|             Reviews|
+--------------------+
|The pasta was a d...|
|We ate the fish i...|
|My family did not...|
|The girl even tri...|
|this is his job a...|
|I'm always greete...|
+--------------------+
```

```
%pyspark (/U4G66226D/spaces)
# Tokenize DataFrame
review_data = Tokenizer(inputCol="Reviews", outputCol="Words")
```

Interpreter: spark.pyspark. **FINISHED** Took 231 millisec. Updated by ars0107 on February 04 2019, 8:51:53 AM (CST)

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

```
+--------------------+--------------------+
|             Reviews|               Words|
+--------------------+--------------------+
|The pasta was a d...|[the, pasta, was,...|
|We ate the fish i...|[we, ate, the, fi...|
|My family did not...|[my, family, did,...|
|The girl even tri...|[the, girl, even,...|
|this is his job a...|[this, is, his, j...|
|I'm always greete...|[i'm, always, gre...|
+--------------------+--------------------+
```

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

```
%pyspark (/U4G66226D/spaces)
# Transform new DataFrame
newFrame = remover.transform(reviewed)
newFrame.show()
```

```
+--------------------+--------------------+--------------------+
|             Reviews|               Words|            filtered|
+--------------------+--------------------+--------------------+
|The pasta was a d...|[the, pasta, was,...|       [pasta, dish]|
|We ate the fish i...|[we, ate, the, fi...|  [ate, fish, tasty]|
|My family did not...|[my, family, did,...|[family, like, food]|
|The girl even tri...|[the, girl, even,...|[girl, even, trie...|
|this is his job a...|[this, is, his, j...|[job, since, prob...|
|I'm always greete...|[i'm, always, gre...|[i'm, always, gre...|
+--------------------+--------------------+--------------------+
```

```
%pyspark (/U4G66226D/spaces)
# Show simplified review
newFrame.select("filtered").show(truncate=False)
```

```
+-----------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------+
|filtered
                                                             |
+-----------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------+
|[pasta, dish]
                                                             |
|[ate, fish, tasty]
                                                             |
|[family, like, food]
                                                             |
|[girl, even, tried, spread, half, cover, roll.]
                                                             |
|[job, since, probably, slowest, time, day, would, least, expect, take, order, put, sandwich, through., course, apol
ogy., wait, woman, finish, couple, came, finally, got, sandwich.]|
|[i'm, always, greeted, , employees, always, seem, eager, help.]
                                                             |
+-----------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------+
```