livelessons

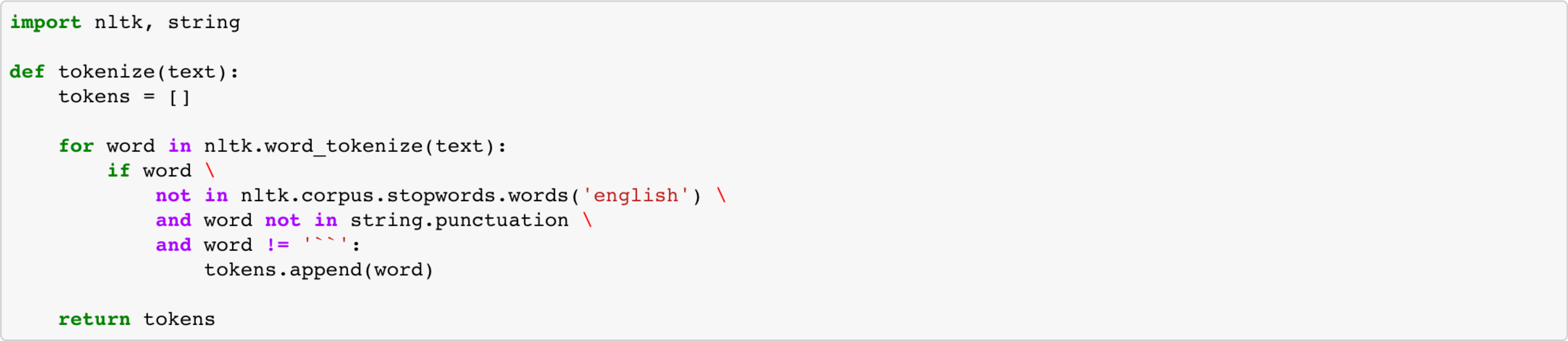
video instruction from technology experts

Lesson 3: Your First Spark Application

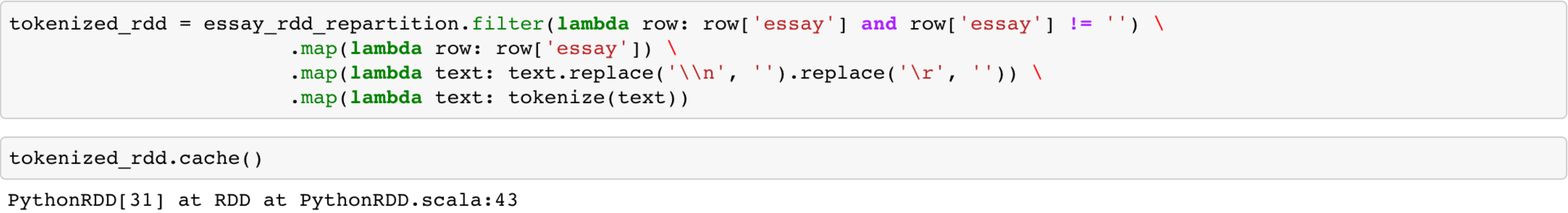
3.7 Tokenization and Vectorization with Spark

livelessons

Tokenization



```
import nltk, string

def tokenize(text):
    tokens = []

    for word in nltk.word_tokenize(text):
        if word \
            not in nltk.corpus.stopwords.words('english') \
            and word not in string.punctuation \
            and word != '``':
                tokens.append(word)

    return tokens
```

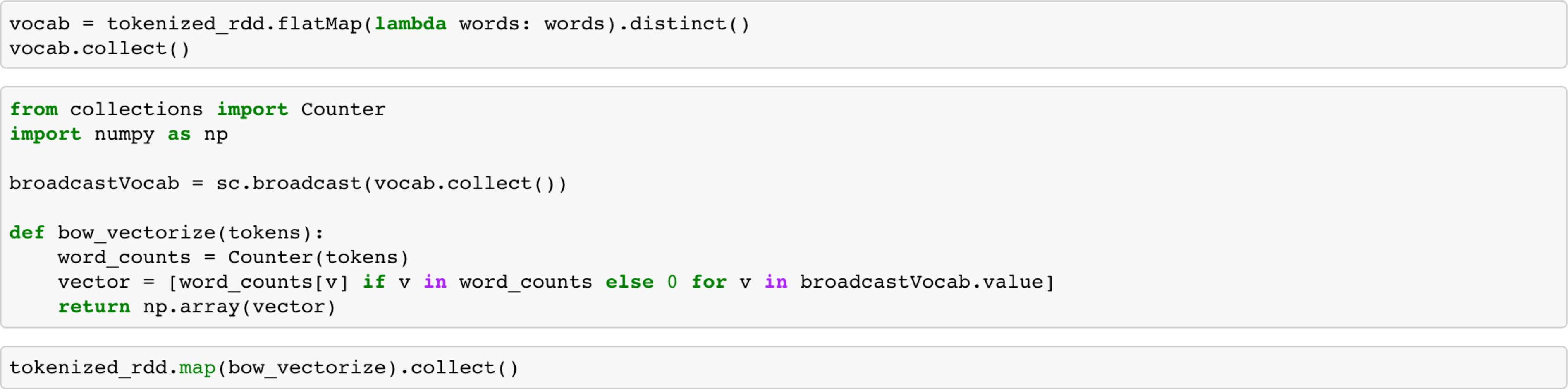


```
tokenized_rdd = essay_rdd_repartition.filter(lambda row: row['essay'] and row['essay'] != '') \
                       .map(lambda row: row['essay']) \
                       .map(lambda text: text.replace('\\n', '').replace('\r', '')) \
                       .map(lambda text: tokenize(text))
```

```
tokenized_rdd.cache()
```

```
PythonRDD[31] at RDD at PythonRDD.scala:43
```

Vectorization



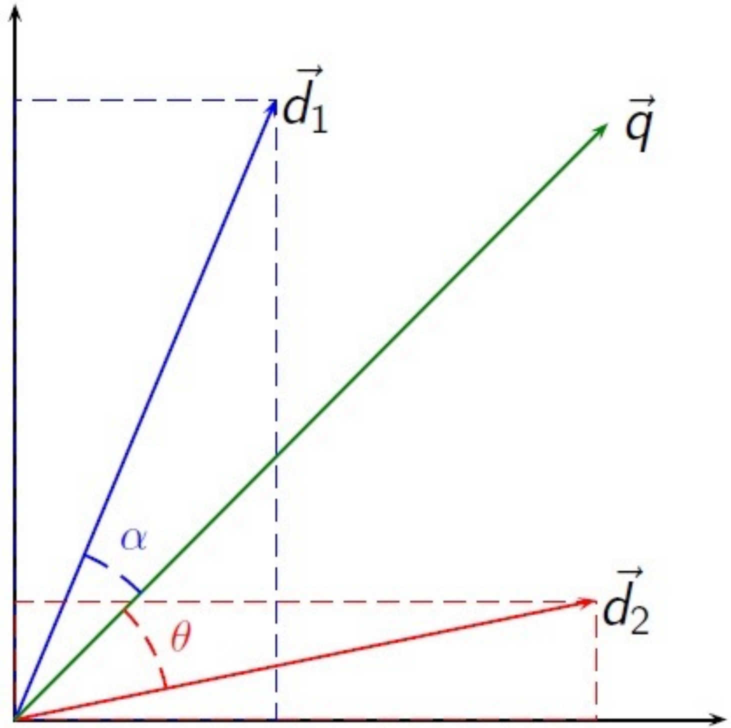
```
vocab = tokenized_rdd.flatMap(lambda words: words).distinct()
vocab.collect()
```

```
from collections import Counter
import numpy as np

broadcastVocab = sc.broadcast(vocab.collect())

def bow_vectorize(tokens):
    word_counts = Counter(tokens)
    vector = [word_counts[v] if v in word_counts else 0 for v in broadcastVocab.value]
    return np.array(vector)
```

```
tokenized_rdd.map(bow_vectorize).collect()
```

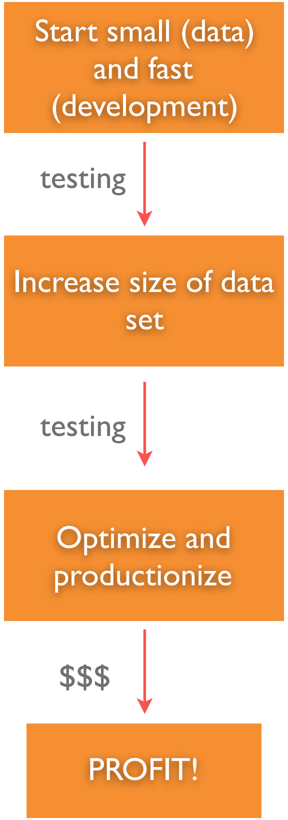
Vector Space Model

Similarity is a measure of “distance”



By Riclas (Own work) CC BY 3.0 , via Wikimedia Commons

Interlude: How to Scale



Start small (data) and fast (development)

testing

Increase size of data set

testing

Optimize and productionize

$$$

PROFIT!