EX 3

create a program wc3.py to select the average number of occurrences of words of the dataset

We wrote the following priogram using map-filter-reduce pattern.
we used `count()` and `sum()` to get the nubers we needed to compute the average.

```
# create a program wc3.py to select the average number of occurrences of words of the dataset

import sys
import re

from pyspark import SparkContext, SparkConf
#output folder in hadoop
_DATA_ = "hdfs:/user/user_lsc_3/labPySparkData/big.txt"

if __name__ == "__main__":
    # create Spark context with necessary configuration
    sc = SparkContext("local","PySpark Word Count Exmaple")

    # read data from text file and split each line into words
    rdd = sc.textFile(_DATA_).flatMap(lambda line: re.split(r"[^\w]*", line.strip().lower()))

    # the count reduced by word
    wordCounts = rdd.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a + b)

    # some numbers
    total_number_of_words = wordCounts.count()
    # same as rdd.count()
    total_word_count = wordCounts.map(lambda x: x[1]).sum()

    # average
    avg = total_word_count / total_number_of_words

    #print
    print (avg)
```

Results

The average occurrence per word is 34