EX 4

create a program wc4.py to select all words starting and ending with a vowel which occur in the dataset for more than the average number computed in (3)

We wrote the following program starting with the one we used for EX3 using map-filter-reduce pattern.

We created a custom filter in a function `myFilter` which calls the self explanatory functions `StartsEndsWithVowel` and `HasHigherThanAverageCount`

```
# create a program wc4.py to select all words starting and ending
# with a vowel which occur in the dataset for more than the average
# number computed in (3)

import re

from pyspark import SparkContext

_DATA_ = "hdfs:/user/user_lsc_3/labPySparkData/big.txt"
_OUTPUT_ = "hdfs:/user/user_lsc_3/labPySparkData/ex4"

def StartsEndsWithVowel(x):
   return re.match(r"\b[aeiou][a-zA-Z]*[aeiou]\b", x)

def HasHigherThanAverageCount(x, avg):
    return x > avg

def myFilter(x, avg):
    return StartsEndsWithVowel(x[1]) and HasHigherThanAverageCount(x[0],avg)

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 counts of ords
    wordCounts = rdd.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a + b)

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

    # average
    avg = total_word_count / total_number_of_words


    counts = wordCounts.map(lambda x: (x[1], x[0]))

    # Filter - sort
    vowelcounts = sc.parallelize(counts.filter(lambda x: myFilter(x,avg)).collect())

    vowelcounts.saveAsTextFile(_OUTPUT_)
```

Results

The result in the `output.txt` file merged with hadoop dfs:

```
(301, u'inside')
(122, u'alone')
(36, u'everywhere')
(1267, u'one')
(59, u'aside')
(50, u'opposite')
(41, u'imagine')
(384, u'once')
(1570, u'into')
(85, u'ago')
(41, u'applause')
(35, u'anywhere')
(104, u'above')
(170, u'office')
(41, u'extra')
(124, u'entrance')
(37, u'envelope')
(36, u'invisible')
(221, u'else')
(49, u'anymore')
(145, u'able')
(193, u'anyone')
(186, u'eye')
(258, u'everyone')
(39, u'awake')
(265, u'onto')
(336, u'uncle')
(37, u'angelina')
(166, u'outside')
(130, u'idea')
(41, u'ernie')
(743, u'are')
(70, u'edge')
(83, u'also')
(144, u'use')
(59, u'alive')
```