EX 2

Write a program `wc2.py` *with a map-reduce scheme to select the 3 words starting and ending with a vowel that occur more frequently in the dataset*

We wrote the following priogram using map-filter-reduce pattern.
we used `take()` to get the first three elements after ordering by ke in a descending order.

```
# create a program wc2.py with a map-reduce scheme to select the 3
# words starting and ending with a vowel that occur more frequently in the dataset

import sys

import re

from pyspark import SparkContext, SparkConf

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


def StartsEndsWithVowel(x):
   return re.match(r"\b[aeiou][a-zA-Z]*[aeiou]\b", x)
   # return x[0] in _VOWELS_ and x[-1] in _VOWELS_


if __name__ == "__main__":
    sc = SparkContext("local","PySpark Word Count Exmaple")
    rdd = sc.textFile(_DATA_).flatMap(lambda line: re.split(r"[^\w]*", line.strip().lower()))

    # Contatore parola
    counts = rdd.map(lambda word: (word.lower(), 1)).reduceByKey(lambda a,b:a +b).map(lambda x:

    # Filter - sort
    vowelcounts = sc.parallelize(counts.filter(lambda x: StartsEndsWithVowel(x[1])).sortByKey(as

    # save text into hadoop file system
    vowelcounts.saveAsTextFile(_OUTPUT_)
```

We obtained the following output in a txt file with

```
hdfs dfs -getmerge /user/user_lsc_3/labPySparkData/output/* output.txt
```

```
[user_lsc_3@it EX2]$ hdfs dfs -text /user/user_lsc_3/labPySparkData/output/*
2021-12-11 17:28:09,918 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usi
(1570, u'into')
(1267, u'one')
(743, u'are')
```

which are the most common words starting and ending with a vowel.