Quiz #2

Instructions

First Program in PySpark

Using PySpark shell, you will read a text file with 6 records outlined below:

```
$ cat /tmp/foxy.txt
a Fox jumped high and high and jumped and jumped
fox of red jumped
fox of blue jumped
a Fox is a red fox of hen
a fox is a high fox
orange fox is high and blue and blue
```

The final output:

```
<unique-word> <frequency>
```

The following rules will be applied:

- if a word's length is smaller than 3, that word will not appear in output
- if a frequency of a unique-word is less than 3, that word will not appear in output
- all words must be converted to lowercase (example: 'fox' is the same as 'Fox', etc.)
- all of your solution must be properly documented
- even though your input is very small, but your program must be as generic as possible to handle any size input
- efficiency is very important for writing this program and all big data programs (it means that your solution should be optimized to handle big volume data without performance penalties)
- you can not change the content of input file