# Spark hands on

Lei Pi

## Q1: Install Spark

I downloaded a virtualbox image and imported it to create a virtual machine.

Screenshot:



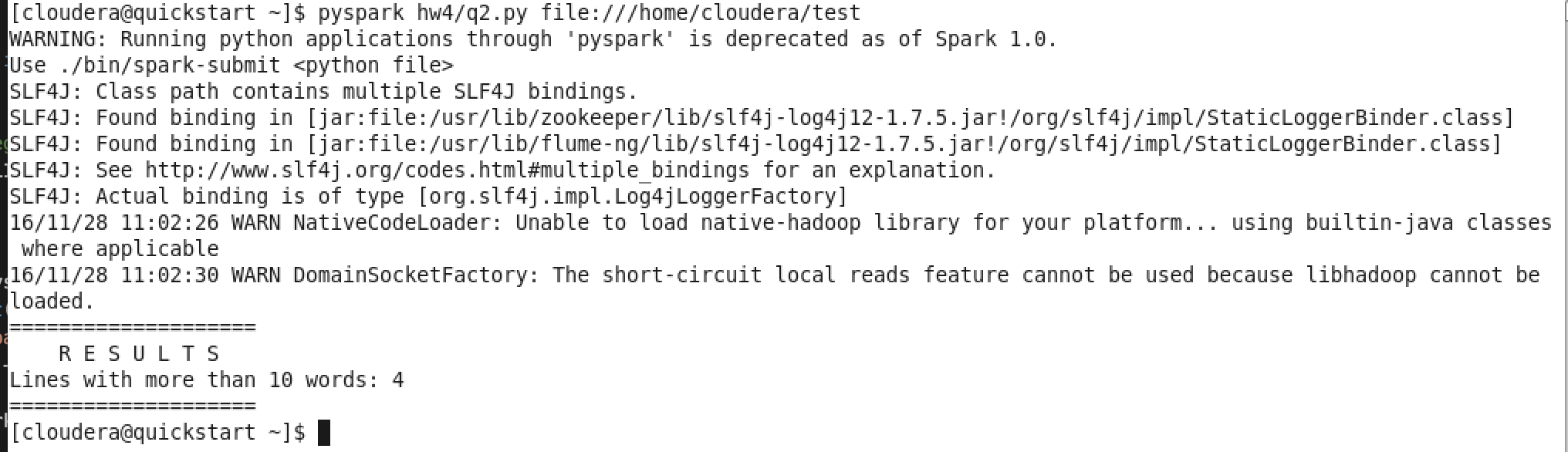
## Q2: Line Count

Please see my source code for this question in q2.py.

In brief,

1. Use “sc.textFile(…)” to read the file
2. Use a regular expression to split the words
3. Use “filter(expression)” method to get lines with more than 10 words
4. Use “count()” method to count the result

Screenshot of the result:



## Q3: Common Words

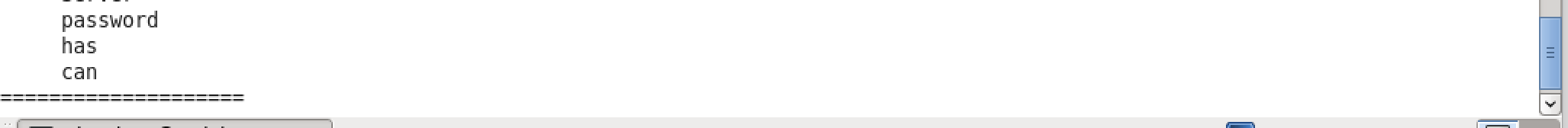
Please see my source code for question 3 in q3.py.

In brief,

1. Create a flatmap of the first text file
   1. Split lines with a regular expression to extract words
   2. Filter out articles such as “a, an and the”
2. For the second file
   1. Extract each word in the same way
   2. Find if there is a same word in the flatmap of the first file.

Screenshot of the result:





## Q4. Inverted Index

Please see my source code for this problem in q4.py.

In brief:

1. Use zipWithIndex method to add line number to map function calls.
2. Extract each word using flatmap and emit result for each word in form of   
   *[word, [line\_number]]*
3. Reduce by key, for each two items with the same key, merge their values.  
   (Explain: key is word in the flat map, value is the line number list for each word)

Screenshot of results:

