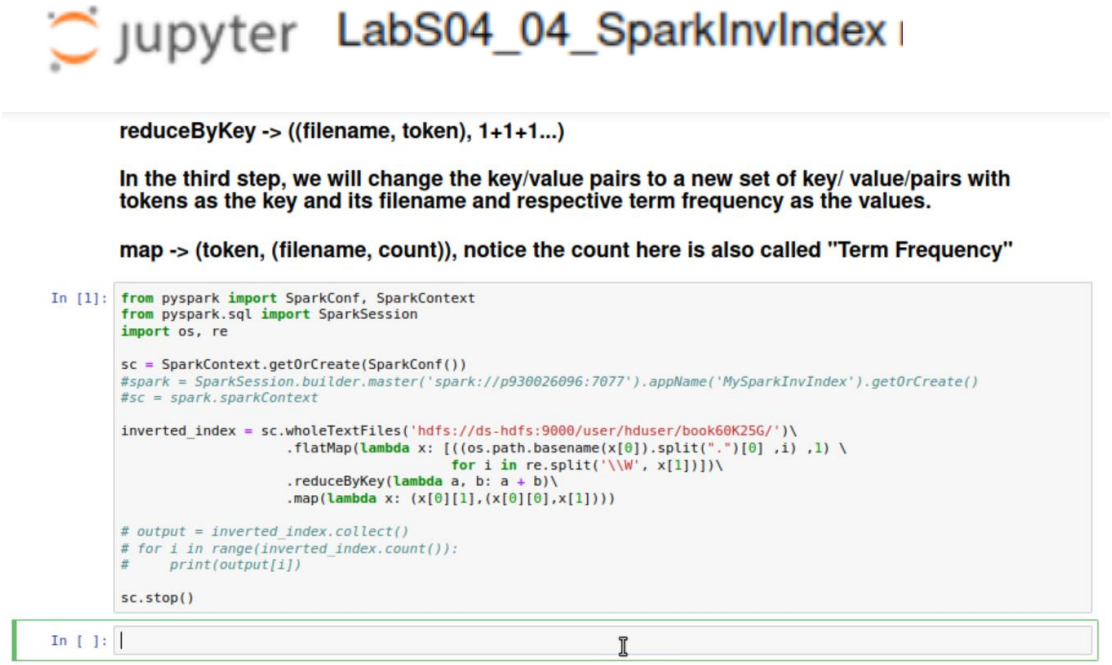
Bonus

Bonus 1.

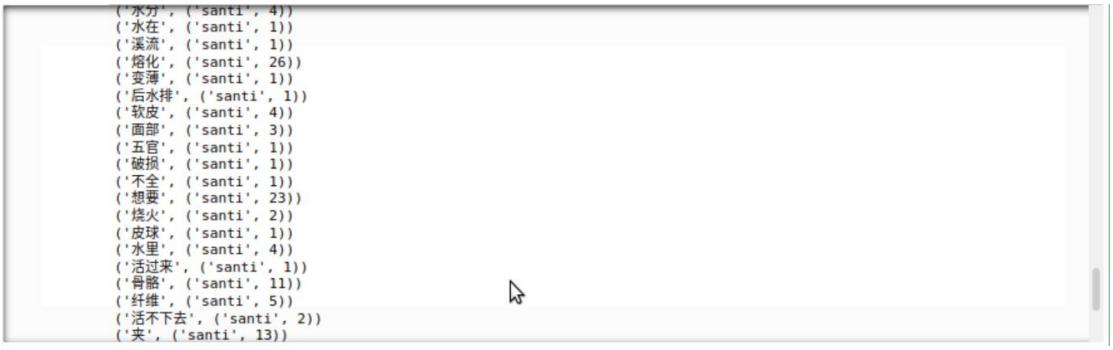


Run a SparkInvIndex with more than 1G files. Here we use HDFS directory to store our files. It contains 60 thousand books. Since if we print out the result on the website, it will crush since they are huge. So, I comment the output code lines. The result turns out to be good. The code block run successfully.

Bonus 2

```
In [2]: from pyspark import SparkConf, SparkContext
from pyspark.sql import SparkSession
import os, re
import jieba

sc = SparkContext.getOrCreate(SparkConf())

inverted_index = sc.wholeTextFiles('./iitfidf_ch/*.txt')\
                   .flatMap(lambda x: [((os.path.basename(x[0]).split(".")[0] ,i) ,1) \
                                       for i in jieba.cut(x[1])])\
                   .reduceByKey(lambda a, b: a + b)\
                   .map(lambda x: (x[0][1],(x[0][0],x[1])))

output = inverted_index.collect()
for i in range(inverted_index.count()):
    print(output[i])

sc.stop()
```

Get the inverse index for a Chinese text file. We create a local directory to store three Chinese text files (santi.txt, sanguo.txt, douluodalu.txt). Since we should use Jieba to split the words.



The result contains a Chinese word, the file that contains the word and the term frequency.