

Bonus

Bonus 1.



`reduceByKey -> ((filename, token), 1+1+1...)`

In the third step, we will change the key/value pairs to a new set of key/ value/pairs with tokens as the key and its filename and respective term frequency as the values.

`map -> (token, (filename, count))`, notice the count here is also called "Term Frequency"

```
In [1]: from pyspark import SparkConf, SparkContext
from pyspark.sql import SparkSession
import os, re

sc = SparkContext.getOrCreate(SparkConf())
#spark = SparkSession.builder.master('spark://p938026096:7077').appName('MySparkInvIndex').getOrCreate()
#sc = spark.sparkContext

inverted_index = sc.wholeTextFiles('hdfs://ds-hdfs:9000/user/hduser/book60K25G/')\
    .flatMap(lambda x: [(os.path.basename(x[0]).split(".")[0], i), 1] \
        for i in re.split('\W', 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()
```

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.

```
( '水分', ('santi', 4))  
( '水在', ('santi', 1))  
( '溪流', ('santi', 1))  
( '熔化', ('santi', 26))  
( '变薄', ('santi', 1))  
( '后水排', ('santi', 1))  
( '软皮', ('santi', 4))  
( '面部', ('santi', 3))  
( '五官', ('santi', 1))  
( '破损', ('santi', 1))  
( '不全', ('santi', 1))  
( '想要', ('santi', 23))  
( '烧火', ('santi', 2))  
( '皮球', ('santi', 1))  
( '水里', ('santi', 4))  
( '活过来', ('santi', 1))  
( '骨骼', ('santi', 11))  
( '纤维', ('santi', 5))  
( '活不下去', ('santi', 2))  
( '夹', ('santi', 13))
```

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