Raphaël Gavache

TF-IDF

Les deux textes sont dans ~/input
Il y a 4 jobs MapReduce:

1. **WordCount**

On transforme les majuscules en minuscules puis en on retire les caractères spéciaux ()/:;$^… en filtrant avec la regex "[^\\w\\s]"

input -> output1
Calcule le total de chaque mot par texte et écris :
old@callwild.txt -> "*3*"

2. **WordCountDoc**

output1 -> output2
Calcule le nombre total de mot dans chaque texte :
and@callwild.txt -> "1464;31778"

3. **DocCountWord**

output2 -> output3
Calcule le tf-idfs d'un mot
the@callwild.txt -> 0.07168

4. **SortKeyByValue**

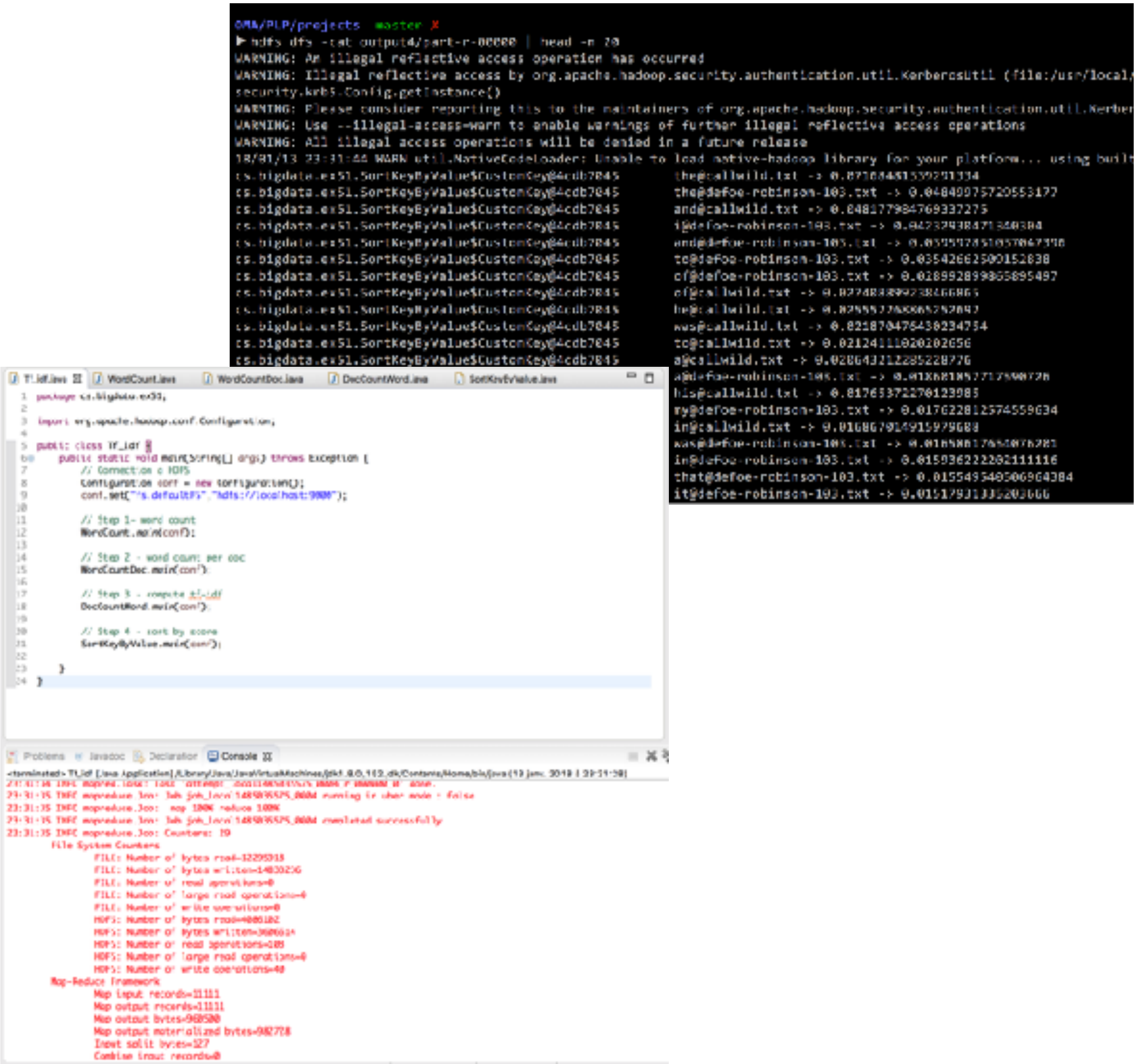
output3 -> output4
Ordonne par ordre décroissant les tf-idfs avec une clef custom dans un job Map. Renvoie
customKey ->the@callwild.txt -> 0.07168481339291334

Résultat - Top 20

Mot	Texte	Tf-Idf
the	callwild	0.07168
the	robinson	0.04850
and	callwild	0.04178
i	robinson	0.04233
and	robinson	0.03960
to	robinson	0.03543
of	robinson	0.02899
of	callwild	0.02740
he	callwild	0.02555
was	callwild	0.02187
to	callwild	0.02124
a	callwild	0.02064
a	robinson	0.01860

his	callwild	0.01765
my	robinson	0.01762
in	callwild	0.01687
was	robinson	0.01658
in	robinson	0.01594
that	robinson	0.01554
it	robinson	0.01517

Screens



```
OMA/PLP/projects  master ✗
▶ hdfs dfs -cat output4/part-r-00000 | head -n 20
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/
security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.Kerber
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
18/01/13 23:31:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       the@callwild.txt -> 0.07108461539251334
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       the@defoe-robinson-103.txt -> 0.04849975720553177
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       and@callwild.txt -> 0.048177984769337275
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       i@defoe-robinson-103.txt -> 0.04232438471340304
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       and@defoe-robinson-103.txt -> 0.035977851057067398
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       to@defoe-robinson-103.txt -> 0.03542662500152838
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       of@defoe-robinson-103.txt -> 0.028992899865895497
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       of@callwild.txt -> 0.027488809923846665
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       he@callwild.txt -> 0.0255527268665252047
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       was@callwild.txt -> 0.021870476430234754
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       to@callwild.txt -> 0.02124111020202656
cs.bigdata.ex51.SortKeyByValue$CustomKey@4cdb7845       a@callwild.txt -> 0.020643212285228776
                                                        a@defoe-robinson-103.txt -> 0.018681857717590728
                                                        his@callwild.txt -> 0.01765372270123985
                                                        my@defoe-robinson-103.txt -> 0.017622812574559634
                                                        in@callwild.txt -> 0.016867014915979688
                                                        was@defoe-robinson-103.txt -> 0.016580817654076281
                                                        in@defoe-robinson-103.txt -> 0.01593622202111116
                                                        that@defoe-robinson-103.txt -> 0.015549540506964384
                                                        it@defoe-robinson-103.txt -> 0.01517931335203666
```

```java
package cs.bigdata.ex53;

import org.apache.hadoop.conf.Configuration;

public class Tf_idf {
    public static void main(String[] args) throws Exception {
        // Connection a HDFS
        Configuration conf = new Configuration();
        conf.set("fs.defaultFS","hdfs://localhost:9000");

        // Step 1- word count
        WordCount.main(conf);

        // Step 2 - word count per doc
        WordCountDoc.main(conf);

        // Step 3 - compute tf-idf
        DocCountWord.main(conf);

        // Step 4 - sort by score
        SortKeyByValue.main(conf);

    }
}
```

```
<terminated> TF_Idf [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_152.jdk/Contents/Home/bin/java (13 janv. 2018 à 23:31:30)
23:31:34 INFO mapred.Task: Task 'attempt_local1485095575_0004_r_000000_0' done.
23:31:35 INFO mapreduce.Job: Job job_local1485095575_0004 running in uber mode : false
23:31:35 INFO mapreduce.Job:  map 100% reduce 100%
23:31:35 INFO mapreduce.Job: Job job_local1485095575_0004 completed successfully
23:31:35 INFO mapreduce.Job: Counters: 39
	File System Counters
		FILE: Number of bytes read=12295318
		FILE: Number of bytes written=14830256
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=4006102
		HDFS: Number of bytes written=3600614
		HDFS: Number of read operations=108
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=40
	Map-Reduce Framework
		Map input records=11111
		Map output records=11111
		Map output bytes=960500
		Map output materialized bytes=982728
		Input split bytes=127
		Combine input records=0
```