

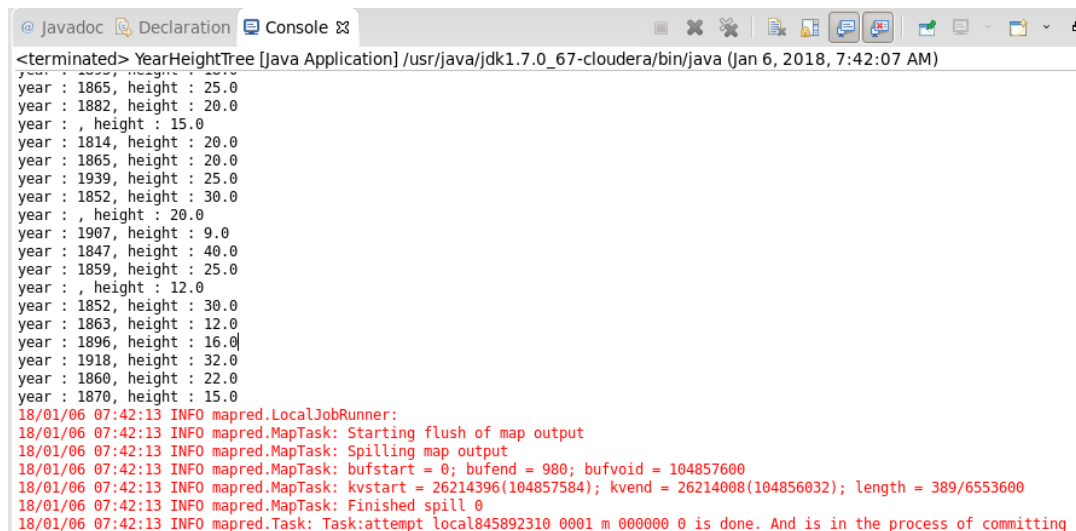
Hamza Haloui

Mohammed El Abridj

Stéphane Multon

## Rapport PLP OMA 2017-2018

### Hadoop MapReduce en Java

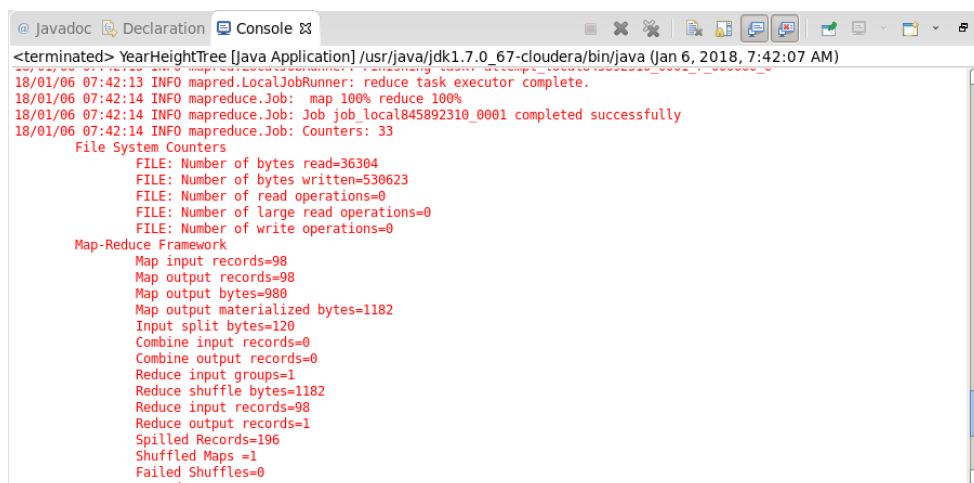


```
@ Javadoc Declaration Console X
<terminated> YearHeightTree [Java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Jan 6, 2018, 7:42:07 AM)
year : 1865, height : 25.0
year : 1882, height : 20.0
year : , height : 15.0
year : 1814, height : 20.0
year : 1865, height : 20.0
year : 1939, height : 25.0
year : 1852, height : 30.0
year : , height : 20.0
year : 1907, height : 9.0
year : 1847, height : 40.0
year : 1859, height : 25.0
year : , height : 12.0
year : 1852, height : 30.0
year : 1863, height : 12.0
year : 1896, height : 16.0
year : 1918, height : 32.0
year : 1860, height : 22.0
year : 1870, height : 15.0
18/01/06 07:42:13 INFO mapred.LocalJobRunner:
18/01/06 07:42:13 INFO mapred.MapTask: Starting flush of map output
18/01/06 07:42:13 INFO mapred.MapTask: Spilling map output
18/01/06 07:42:13 INFO mapred.MapTask: bufstart = 0; bufend = 980; bufvoid = 104857600
18/01/06 07:42:13 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214008(104856032); length = 389/6553600
18/01/06 07:42:13 INFO mapred.MapTask: Finished spill 0
18/01/06 07:42:13 INFO mapred.Task: Task:attempt_local845892310_0001_m_000000_0 is done. And is in the process of committing
```

#### 2.7 Arbres de Paris, affichage de .csv

#### 2.8 Stations NOAA, affichage de .txt

Le contenu du code est là mais nous n'avons pas réussi à faire fonctionner le code en lisant le fichier isd-history.txt depuis HDFS. Nous obtenons une erreur de connexion indiquée par la ligne d'erreur suivante, affichée dans la console :



```
@ Javadoc Declaration Console X
<terminated> YearHeightTree [Java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Jan 6, 2018, 7:42:07 AM)
18/01/06 07:42:13 INFO mapred.LocalJobRunner: reduce task executor complete.
18/01/06 07:42:14 INFO mapreduce.Job: map 100% reduce 100%
18/01/06 07:42:14 INFO mapreduce.Job: Job job_local845892310_0001 completed successfully
18/01/06 07:42:14 INFO mapreduce.Job: Counters: 33
File System Counters
  FILE: Number of bytes read=36304
  FILE: Number of bytes written=530623
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
Map-Reduce Framework
  Map input records=98
  Map output records=98
  Map output bytes=980
  Map output materialized bytes=1182
  Input split bytes=120
  Combine input records=0
  Combine output records=0
  Reduce input groups=1
  Reduce shuffle bytes=1182
  Reduce input records=98
  Reduce output records=1
  Spilled Records=196
  Shuffled Maps =1
  Failed Shuffles=0
```

WARN security.UserGroupInformation: PriviledgedActionException as:cloudera (auth:SIMPLE) cause:java.net.ConnectException: Call From quickstart.cloudera/192.168.1.111 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: <http://wiki.apache.org/hadoop/ConnectionRefused>

## 5.1 TFIDF

Là encore le contenu du code est là, et fonctionne en lisant les fichiers en local, en ligne par ligne plutôt qu'en document par document (ce qui modifie la formule du TF-IDF). Mais nous n'avons donc pas réussi à le faire fonctionner depuis les fichiers en HDFS pour la même raison que ci-dessus au 2.8. Aussi le dernier print censé afficher le top 20 des TF-IDF ne renvoie pas le bon top 20, en tout cas pas au niveau des couples (docid, word) pour une raison que nous n'avons pas réussi à déceler. Mais il semble y avoir les bonnes valeurs de TF-IDF (vis à vis du mode de lecture en ligne par ligne, les valeurs ne sont donc pas celles attendues pour une lecture en document par document). Les top 20 valeurs affichées sont toutes les mêmes mais ça semble être par construction, car la plage des valeurs de wordcount, wordsPerDoc et DocsPerWord n'est pas si large que cela donc il y a plusieurs ex-aequo (dans notre modèle ligne par ligne tout du moins).

Top 20 scores ligne par ligne affiché dans la console:

*Top 20 tfidf are {11.955776270285945, 11.955776270285945,  
11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945, 11.955776270285945}*

## 5.2 PageRank

Pour cette question on a une erreur de compilation sur la variable globale Map, mais on sait pas comment la résoudre. Le reste du code est censé être correct.

### 5.3 Arbres de Paris, calculs

```

<terminated> TreesType [java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Jan 7, 2018, 11:32:13 AM)
18/01/07 11:32:43 INFO mapred.LocalJobRunner: Finishing task: attempt_local821199962_0001_r_000000_0
18/01/07 11:32:43 INFO mapred.LocalJobRunner: reduce task executor complete.
18/01/07 11:32:44 INFO mapreduce.Job: map 100% reduce 100%
18/01/07 11:32:44 INFO mapreduce.Job: Job job_local821199962_0001 completed successfully
18/01/07 11:32:44 INFO mapreduce.Job: Counters: 33
File System Counters
  FILE: Number of bytes read=36780
  FILE: Number of bytes written=532265
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
Map-Reduce Framework
  Map input records=98
  Map output records=98
  Map output bytes=1233
  Map output materialized bytes=1435
  Input split bytes=105
  Combine input records=0
  Combine output records=0
  Reduce input groups=37
  Reduce shuffle bytes=1435
  Reduce input records=98
  Reduce output records=37
  Spilled Records=196

<terminated> TreesType [java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Jan 7, 2018, 11:32:13 AM)
  Reduce shuffle bytes=1435
  Reduce input records=98
  Reduce output records=37
  Spilled Records=196
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=90
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=331227136
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=16778
File Output Format Counters
  Bytes Written=410
  
```

#### TreesType :

```

Acer 3
Aesculus 3
Ailanthus 1
Alnus 1
Araucaria 1
Broussonetia 1
Calocedrus 1
Catalpa 1
Cedrus 4
Celtis 1
Corylus 3
Davidia 1
Diospyros 4
Eucommia 1
Fagus 8
Fraxinus 1
GENRE 1
Ginkgo 5
Gymnocladus 1
Juglans 1
Liriodendron 2
Maclura 1
  
```

Magnolia	1
Paulownia	1
Pinus	5
Platanus	19
Pterocarya	3
Quercus	4
Robinia	1
Sequoia	1
Sequoiadendron	5
Styphnolobium	1
Taxodium	3
Taxus	2
Tilia	1
Ulmus	1
Zelkova	4

#### Et pour la hauteur :

Acer	16	
Aesculus	30	
Ailanthus	35	
Alnus	16	
Araucaria	9	
Broussonetia		12
Calocedrus	20	
Catalpa	15	
Cedrus	30	
Celtis	1	
6Corylus	20	
Davidia	12	
Diospyros	14	
Eucommia	1	
2Fagus	30	
Fraxinus	30	
Genre	0	
Ginkgo	33	
Gymnocladus		101
Juglans	28	
Liriodendron		35
Maclura	1	
3Magnolia	12	
Paulownia	20	
Pinus	30	
Platanus	45	
Pterocarya	30	
Quercus	31	
Robinia	11	
Sequoia	30	
Sequoiadendron		35
Styphnolobium		1
0Taxodium	3	
5Taxus		
13Tilia	20	
Ulmus	1	
5Zelkova	30	

## **Spark**

Les réponses sont apportées directement sur le script suivant :  
<https://github.com/halouih/velib/blob/master/Velib-HH-ME-SM.py>