Big Data Algorithms, Techniques and Platforms

Flavie VAMPOUILLE
MSc in Data Sciences & Business Analytics
ESSEC Business Scholl & CentralSupélec

Question 2.7 Displaying the content of a CSV file

The objective of this exercise is to display the year and the height of each tree of the file arbres.csv.

I obtained the following results:

Annee : ANNEE PLANTATION
Hauteur : HAUTEUR

Annee : 1935
Hauteur : 13.0

Annee : 1854
Hauteur : 20.0

Annee : 1862
Hauteur : 22.0

Annee : 1906
Hauteur : 16.0

…

Question 2.8 Displaying the content of a compact file

I obtained the following results:

USAF code : 007005
Name : CWOS 07005
Country :
Elevation :

USAF code : 007011
Name : CWOS 07011
Country :

Elevation :

USAF code : 007018
Name : WXPOD 7018
Country :
Elevation : +7018.0

USAF code : 007025
Name : CWOS 07025
Country :
Elevation :

USAF code : 007026
Name : WXPOD 7026
Country : AF
Elevation : +7026.0

USAF code : 007034
Name : CWOS 07034
Country :
Elevation :

…

Note that there is a lot of missing values for “Country” and “Elevation” but they are missing in the original isd-history.txt of the NOAA.

Question 4.1 TF-IDF

To write the TF-IDF script I follow the example given in the subject. Hence my code is composed of three parts:

1. Round1.java with the two text as input and the out_round1 as output
2. Round2.java with the out_round1 as input and the out_round2 as output
3. Round3.java with the out_round2 as input and the out_round3 as output

My final output is of the form word/doc ; value(TF-IDF).

The 20 words with the highest TF-IDF are:

buck/callwild.txt	0.006827639335
dogs/callwild.txt	0.002443116951
thornton/callwild.txt	0.001766897080
myself/Crusoe.txt	0.001668513060
sled/callwild.txt	0.001308812652
spitz/callwild.txt	0.001308812652

francois/callwild.txt	0.001134304298
bucks/callwild.txt	0.001025236577
friday/Crusoe.txt	0.001022821362
trail/callwild.txt	0.000894355312
john/callwild.txt	0.000872541768
perrault/callwild.txt	0.000807101135
hal/callwild.txt	0.000654406326
team/callwild.txt	0.000654406326
thoughts/Crusoe.txt	0.000634263526
ice/callwild.txt	0.000610779238
traces/callwild.txt	0.000610779238
solleks/callwild.txt	0.000588965693
around/callwild.txt	0.000567152149
dave/callwild.txt	0.000523525061

Question 4.2 Page Rank

For the Page Rank problem I write a java code that calculate the page rank of documents with the method:

- *Page A has pages T1...Tn which point to it (i.e., are citations).*
- *The parameter d is a damping factor which can be set between 0 and 1. Here we set d to 0.85.*
- *C(A) is defined as the number of links going out of page A.*
- *The PageRank of a page A is given as follows:*

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

All page ranks are initialized at 0.15 as this will be there first value.

I should have like to iterate on the reducer which calculate page rank but was not able to do it with mapreduce. The 10 users that have the highest PageRank scores in this social are, in descending order:

Page	PageRank
136	9.136037732628
4415	8.955841368715
763	8.815291528202
9412	7.820179147648
118	7.712191726016
1179	7.455868229821
2969	6.913863241232
1621	6.886318031670
791	6.847470217101
4416	6.770972316074