# TP Hadoop

Please refer to the code at https ://github.com/melisandeteng/BigData/ for details, where I commented the code.

## 1 Displaying the content of a .csv file

I did not need to use a Tree class as specified in the handout, because it is possible to parse the data by choosing the right components (year and tree height) directly. `YearHeightTree.javareturns` the list of year of plantation and height of the trees, separated by a space. Note that when the year was missing, I chose to still print the height of the tree.

## 2 Displaying the content of a compact file

Code is in compact file display folder on the gitHub.

## 3 TF-IDF

I implemented TF-IDF in 3 rounds as suggested in the handout and added one more MapReduce to sort the result in tf-idf score descending order.

```
[cloudera@quickstart TFIDFresults]$ head -c 1000 part-r-00000
buck@callwild   0.006655987848836175
dogs@callwild   0.0021816848557984652
thornton@callwild       0.0018858632023187958
myself@defoe-robinson-103.txt   0.0014316866333092067
spitz@callwild  0.0012017755479541085
sled@callwild   0.00116479784126915
francois@callwild       0.0011093312812417119
friday@defoe-robinson-103.txt   9.070617280137971E-4
trail@callwild  7.580430677346046E-4
john@callwild   7.395542143921253E-4
perrault@callwild       7.21065361049646E-4
hal@callwild    6.840876543646873E-4
team@callwild   6.286210943372494E-4
thoughts@defoe-robinson-103.txt 5.442370448775735E-4
sol@callwild    5.361767872783767E-4
ice@callwild    5.176879339358973E-4
traces@callwild 5.176879339358973E-4
leks@callwild   5.176879339358973E-4
around@callwild 4.8071022725093863E-4
dave@callwild   4.6222137390845933E-4
mates@callwild  4.0675481388102134E-4
```

FIGURE 1 – 20 words with highest tf-idf scores

# 4 PageRank

I implemented Page Rank using 3 MapReduce jobs. Results were computed on 150 iterations. This time, to sort the result in descending order, I simply computed minus the pagerank score. This leads to a display of the pagerank score (times -1) followed by the link number (page ID).

```
[cloudera@quickstart PRank]$ head part-r-00000
-0.0032505099661648273  18
-0.0022580698132514954  737
-0.0015209423145279288  118
-0.0014895503409206867  1719
-0.0014242453034967184  136
-0.0014112156350165606  790
-0.0014025804121047258  143
-0.00130797631572932    40
-0.0011011117603629827  1619
-0.0010722678853198886  725
```

FIGURE 2 – 10 highest Page rank links (-pagerank and link number)

# 5 The trees of Paris

For both questions, I considered "Type" to be "Espece".

## 5.1 Number of trees per type and Height of highest tree of each type

Results are reported in the figure below.

(a) Type - Number of tree per type

```
araucana        1
atlantica       2
australis       1
baccata 2
bignonioides    1
biloba  5
bungeana        1
cappadocicum    1
carpinifolia    4
colurna 3
coulteri        1
decurrens       1
dioicus 1
distichum       3
excelsior       1
fraxinifolia    2
giganteum       5
giraldii        1
glutinosa       1
grandiflora     1
hippocastanum   3
ilex    1
involucrata     1
japonicum       1
kaki    2
libanii 2
monspessulanum  1
nigra   3
nigra laricio   1
opalus  1
orientalis      8
papyrifera      1
petraea 2
pomifera        1
pseudoacacia    1
sempervirens    1
serrata 1
stenoptera      1
suber   1
sylvatica       8
tomentosa       2
tulipifera      2
ulmoides        1
virginiana      2
x acerifolia    11
```

(b) Type - Highest tree per type

```
[cloudera@quickstart hightree]$ cat part-r-00000
araucana        9.0
atlantica       25.0
australis       16.0
baccata 13.0
bignonioides    15.0
biloba  33.0
bungeana        10.0
cappadocicum    16.0
carpinifolia    30.0
colurna 20.0
coulteri        14.0
decurrens       20.0
dioicus 10.0
distichum       35.0
excelsior       30.0
fraxinifolia    27.0
giganteum       35.0
giraldii        35.0
glutinosa       16.0
grandiflora     12.0
hippocastanum   30.0
ilex    15.0
involucrata     12.0
japonicum       10.0
kaki    14.0
libanii 30.0
monspessulanum  12.0
nigra   30.0
nigra laricio   30.0
opalus  15.0
orientalis      34.0
papyrifera      12.0
petraea 31.0
pomifera        13.0
pseudoacacia    11.0
sempervirens    30.0
serrata 18.0
stenoptera      30.0
suber   10.0
sylvatica       30.0
tomentosa       20.0
tulipifera      35.0
ulmoides        12.0
virginiana      14.0
x acerifolia    45.0
```