Answers to Question 4.1 & 4.2

4.1 - Which 20 words have the highest tf-idf scores in these documents ?

The output file is accessible here :
https://github.com/Thibaultcos/BigData_Homework/tree/master/output_tfidf

The top 20 words are :
Word, text   tf-idf

the,callwild   0.049686562740034475
the,defoe-robinson-103.txt   0.03368342521495419
and,callwild   0.033393383474535895
i,defoe-robinson-103.txt   0.029398002159616504
and,defoe-robinson-103.txt   0.02750670211786081
to,defoe-robinson-103.txt   0.024604042235045417
of,defoe-robinson-103.txt   0.020135774462679994
of,callwild   0.01901961488556192
he,callwild   0.017710925787931514
was,callwild   0.015158982047552219
to,callwild   0.014722752348342083
a   0.014308334134092453
a,defoe-robinson-103.txt   0.012919122037491335
my,defoe-robinson-103.txt   0.012239168246044423
his,callwild   0.01223624306284431
in,callwild   0.01169095593883164
was,defoe-robinson-103.txt   0.011519217172747693
in,defoe-robinson-103.txt   0.011067819277585456
that,defoe-robinson-103.txt 0.010799266099450961
it,defoe-robinson-103.txt   0.0105421407161307

4.2 - Which 10 users have the highest PageRank scores in this social network?

The output file is accessible here :
https://github.com/Thibaultcos/BigData_Homework/tree/master/output_Pagerank/PageRank/
FinalOutput
The top 10 users are :

ID Pagerank
763   0.004751382311249525
634   0.002997655000722514
645   0.0027720979182803665
5232   0.002240818906005603

71399 0.0022371581790451025
4920   0.002079488281140241
3924   0.0018189595100313529
2969   0.001443759677990109
1835   0.0014222231983002775
637   0.0012606686687434848