

# Big Data Algorithms, Techniques and Platforms (2016–2017 Spring Semester) Tutorial 1

LAU Tsz Kit

Option Mathématiques Appliquées — Data Sciences Track (Exchange Student)

**CentraleSupélec (École Centrale Paris)**

Master of Science in Big Data Technology Student

**The Hong Kong University of Science and Technology**

tsz-kit.lau@student.ecp.fr

May 27, 2017

## Exercise 4

### Question 4.1

The 20 words having the highest tf-idf scores are:

	Word	TF-IDF
1	the@callwild	0.070545973
2	the@defoe-robinson-103.txt	0.048415435
3	and@callwild	0.04744356
4	i@defoe-robinson-103.txt	0.04229064
5	and@defoe-robinson-103.txt	0.03953202
6	to@defoe-robinson-103.txt	0.035394089
7	of@defoe-robinson-103.txt	0.028990148
8	of@callwild	0.027004428
9	he@callwild	0.025239231
10	was@callwild	0.021523025
11	to@callwild	0.020903657
12	a@callwild	0.020439132
13	a@defoe-robinson-103.txt	0.018719212
14	my@defoe-robinson-103.txt	0.017586207
15	his@callwild	0.017373262
16	in@callwild	0.016691958
17	was@defoe-robinson-103.txt	0.016551724
18	in@defoe-robinson-103.txt	0.01590312
19	that@defoe-robinson-103.txt	0.015517241
20	it@defoe-robinson-103.txt	0.015147783

**Question 4.2**

The 10 users having the highest PageRank scores in the social network are:

	User	PageRank
1	18	0.153503835
2	4415	0.151613057
3	737	0.151492923
4	790	0.151298165
5	1753	0.151282519
6	143	0.151276797
7	1719	0.151270792
8	136	0.151115745
9	751	0.151111469
10	118	0.150962278