

BIG DATA COMPUTING 2020/21 – HOMEWORK 2

Do the tests with the parameters indicated in the first 3 columns of the following table. Fill the table reporting, for each test, 4 values: the approximate average silhouette coefficient of the input clustering (**approxSilhFull**); the time to compute approxSilhFull; the exact silhouette coefficient of the clustering sample (**exactSilhSample**); the time to compute exactSilhSample. The running times must not include the time to load the input file and to compute the sample. The values in the table for Uber_3_small.csv must be the average of 3 runs of the same experiment.

Table to be filled

Dataset	k	t	approxSilhFull	Time (ms) to compute approxSilhFull	exactSilhSample	Time (ms) to compute exactSilhSample
Uber_10_large.csv	10	2000	0.6284781034778061	140525ms	0.5284811736171426	5142ms
Uber_3_large.csv	3	2000	0.49262204165251927	17583ms	0.4777229036647352	631ms
Uber_3_small.csv	3	40	0.8782979000141207	80ms	0.808918819124161	4ms
Uber_3_small.csv	3	80	0.916279244530017	50ms	0.8204577684646315	3ms
Uber_3_small.csv	3	1000	0.9148194891306463	110ms	0.9148194891306459	33ms