LEARNING WITH MASSIVE DATA
Approximate Near-Duplicate Detection with SimHash

Serena Zanon 887050

Complete analysis

In this addition material the full tables are presented.

Dataset	Num bits	Block len	Num Similar Docs	Execution time
small	64	4	137	305.318 s
small	64	8	189	12.5736 s
small	64	16	186	3.7835 s
small	64	32	105	3.89521 s
small	128	4	169	491.452 s
small	128	8	167	27.1291 s
small	128	16	111	5.19636 s
small	128	32	107	4.48202 s
small	256	4	112	782.649 s
small	256	8	109	72.4538 s
small	256	16	107	8.22829 s
small	256	32	170	7.63725 s
medium	64	4	658	1000.81 s
medium	64	8	697	40.437 s
medium	64	16	606	7.01535 s
medium	64	32	636	6.26071 s
medium	128	4	499	1687.22 s
medium	128	8	509	188.221 s
medium	128	16	676	9.64004 s
medium	128	32	643	8.6298 s
medium	256	4	597	2964.12 s
medium	256	8	513	401.189 s
medium	256	16	516	14.6804 s
medium	256	32	493	14.1854 s
big	64	4	1443	2014.73 s
big	64	8	1338	189.289 s
big	64	16	1356	9.43681 s
big	64	32	1088	8.76112 s
big	128	4	1354	3776.14 s
big	128	8	1369	330.238 s
big	128	16	1319	14.2566 s
big	128	32	1291	12.8303 s
big	256	4	1230	6387.67 s
big	256	8	1229	756.601 s
big	256	16	1219	24.5221 s
big	256	32	1209	23.9868 s

Num cores	Num bits	Block len	Num Similar Docs	Execution time
4	64	4	135	301.959 s
4	64	8	132	13.6419 s
4	64	16	132	3.64316 s
4	64	32	119	4.69728 s
4	128	4	119	475.486 s
4	128	8	121	26.9089 s
4	128	16	116	4.7221 s
4	128	32	111	4.61967 s
4	256	4	127	794.488 s
4	256	8	105	72.7951 s
4	256	16	106	7.23101 s
4	256	32	128	7.60397 s
8	64	4	151	309.782 s
8	64	8	178	13.4914 s
8	64	16	152	3.73769 s
8	64	32	105	3.64644 s
8	128	4	117	491.778 s
8	128	8	114	27.352 s
8	128	16	124	4.9396 s
8	128	32	114	4.63238 s
8	256	4	137	784.897 s
8	256	8	129	73.9764 s
8	256	16	128	8.06258 s
8	256	32	142	8.29716 s
12	64	4	137	305.318 s
12	64	8	189	12.5736 s
12	64	16	186	3.7835 s
12	64	32	105	3.89521 s
12	128	4	169	491.452 s
12	128	8	167	27.1291 s
12	128	16	111	5.19636 s
12	128	32	107	4.48202 s
12	256	4	112	782.649 s
12	256	8	109	72.4538 s
12	256	16	107	8.22829 s
12	256	32	170	7.63725 s