Exercise Sheet 3 - Shingling and near duplicate detection

1 Setup parameters

hash function = MD5
k = 4
number of documents = 826;
$\frac{826*825}{2}$ = 340725 document pairs

2 Similar Documents

Similar Documents for various threshold θ:

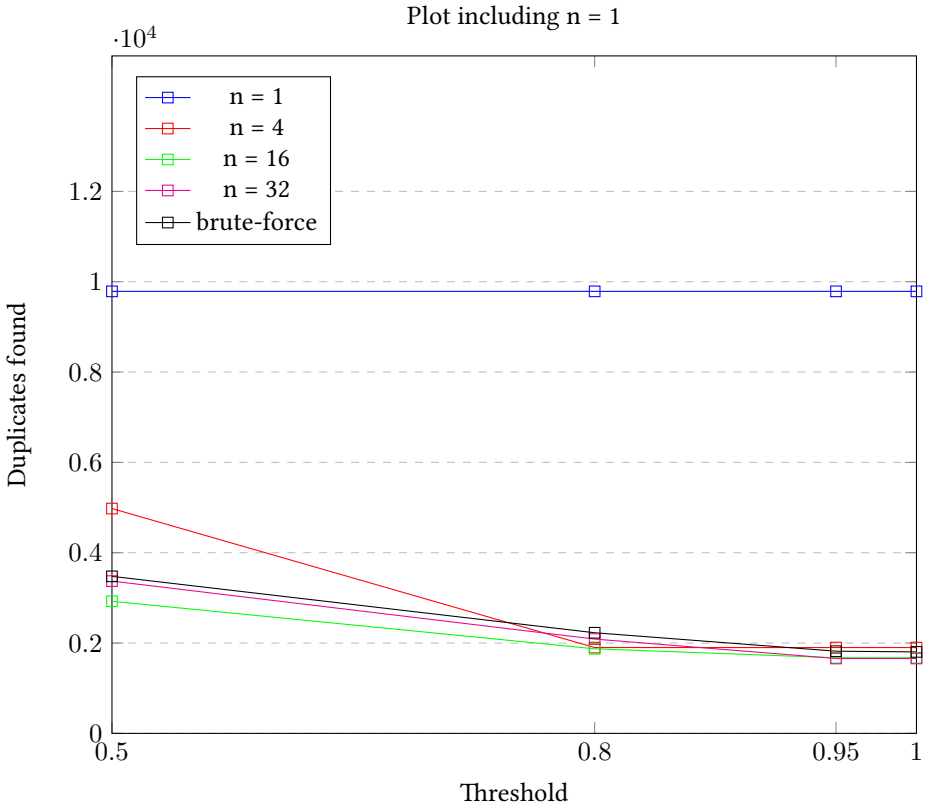
n	$\theta = 0.5$	$\theta = 0.8$	$\theta = 0.95$	$\theta = 1$
1	9787	9787	9787	9787
4	4978	1900	1900	1900
16	2925	1873	1674	1674
32	3372	2088	1658	1658
brute-force	3477	2228	1820	1804

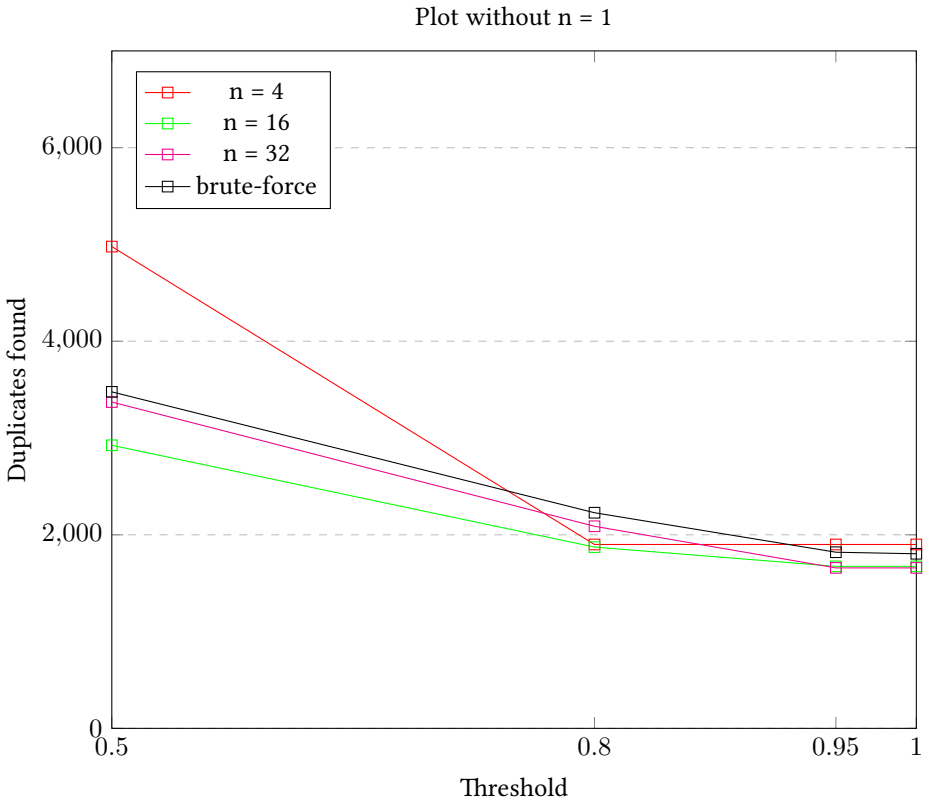
3 Errors

Median , first and third quartile error are 0 for all n since the majority of Jaccard-values is 0 for both minHash and BruteForce. This means the error for most values between them is 0.

n	**Average Error**
1	0.01922101252452837
4	0.00845905841308058
16	0.004916355280523232
32	0.0035869980112459205

4 Plots


Plot including n = 1
$\cdot 10^4$
n = 1
n = 4
n = 16
n = 32
brute-force
1.2
1
0.8
0.6
0.4
0.2
0
Duplicates found
0.5
0.8
0.95
1
Threshold


Plot without n = 1
n = 4
n = 16
n = 32
brute-force
6,000
4,000
2,000
0
Duplicates found
0.5
0.8
0.95
1
Threshold