TDT4305 Big Data-arkitektur

Frå signaturar til LSH

Signaturmatrise

s1	*s2*	*s3*	*s4*
2	1	2	2
1	5	1	1
4	2	1	2
5	4	4	5
3	3	3	1
1	1	1	1
4	2	4	4
1	2	5	1
2	3	5	2

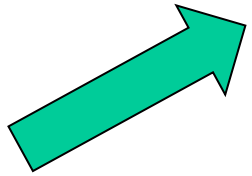
Signaturmatrise splitta i band

	s1	*s2*	*s3*	*s4*
	2	1	2	2
b1	1	5	1	1
	4	2	1	2
	5	4	4	5
b2	3	3	3	1
	1	1	1	1
	4	2	4	4
b3	1	2	5	1
	2	3	5	2

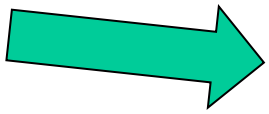
- Signaturmatrise med k=9 rader (dvs. *k* hash-funksjonar brukt på kvart sett, eks. *k* permutasjonar med minhash)
- Vel b=3 band, som her gjev r=k/b=9/3=3 rader per band

Signaturmatrise splitta i band

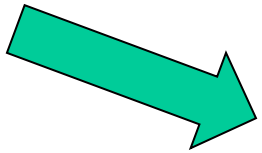
	s1	*s2*	*s3*	*s4*
b1	2	1	2	2
	1	5	1	1
	4	2	1	2
b2	5	4	4	5
	3	3	3	1
	1	1	1	1
b3	4	2	4	4
	1	2	5	1
	2	3	5	2



Bucket 0	S4
Bucket 1	S2
Bucket 2	
Bucket 3	S1
Bucket 4	S3
Bucket 5	



Bucket 0	
Bucket 1	
Bucket 2	S2, S3
Bucket 3	
Bucket 4	S4
Bucket 5	S1



Bucket 0	S2
Bucket 1	S3
Bucket 2	
Bucket 3	
Bucket 4	S1, S4
Bucket 5	

Hash-funksjon vert brukt på kvar "del-signatur", eks. $H([2,1,4]^T)$ og gjev bøtte settet skal ende opp i

- Sett som endar opp i same bøtte er kandidat-par, og likheit sjekkast ved å bruke Jaccard, dvs. J(S2,S3) og J(S1,S4)
- Tal på bøtter? Ideelt ei for kvar mogleg "del-signatur" men dette vert altfor mange i praksis, så meir realistisk er t.d. 1 million