TDT4305 2021 - Assignment 3

1. k-Shingling and Jaccard

Given four sequences of letters, baaa, accc, abba, and baac, find the *k*-shingles for each sequence, where *k=2*.

a) Fill in the table with the remaining *k*-shingles ordered alphabetically.

b) Fill in which *k*-shingle appears in which sets. Use 0s and 1s.

Characteristics matrix:

id	***k*-shingle**	**baaa**	**accc**	**abba**	**baac**
1	aa	1	0	0	1
2	ab	0	0	1	0
3					
4					
5					
6					

c) Fill in the Jaccard similarity table with the remaining similarities between the sequences using their *k*-shingle sets.

Jaccard similarity:

	baaa	**accc**	**abba**	**baac**
baaa	1	0	1/4	
accc				
abba				
baac				

2. Min-Hashing

Generate signatures using the permutation method presented in MMDS ch. 3 video 2. We want signatures of length *n*=3, so three permutations are needed. They are given below.

Row permutations:

1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
6	4	3	2	1	5
4	1	5	2	3	6
2	3	6	4	1	5

a) Fill in the table with the correct signature values.

Signatures:

s1	***s2***	***s3***	***s4***
2	1	2	2

b) Why is it not always feasible to use permutations and how can the permutations be emulated?

3. LSH

Given signatures for the four sequences, find candidate pairs by hashing the signatures into buckets. To make it more interesting, we now assume n=9. We will use b=3 bands with r=3 rows each. We will assume an identity hash function, i.e. for two signatures ($x1, x2, x3$) and ($y1, y2, y3$) appearing in the same band, the signatures go to the same bucket if and only if $x1=y1$ and $x2=y2$ and $x3=y3$.

a) Fill in the table below with the remaining part of the signatures found in exercise 2a.

Signatures split in bands:

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

b) What candidate pairs can we extract from the buckets?