

TDT4305 2021 - Assignment 3 solutions

1. k -Shingling and Jaccard

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

- Fill in the table with the remaining k -shingles ordered alphabetically.
- 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	ac	0	1	0	1
4	ba	1	0	1	1
5	bb	0	0	1	0
6	cc	0	1	0	0

- 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	2/3
accc	0	1	0	1/4
abba	1/4	0	1	1/5
baac	2/3	1/4	1/5	1

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.

- Using the method presented in Video 2 at time 6:15, the following signatures are obtained.

<i>s1</i>	<i>s2</i>	<i>s3</i>	<i>s4</i>
2	1	2	2
1	5	1	1
4	2	1	2

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

- See [MMDS] section 3.3.5.

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:

	<i>s1</i>	<i>s2</i>	<i>s3</i>	<i>s4</i>	
<i>b1</i>	2	1	2	2	
	1	5	1	1	
	4	2	1	2	
<i>b2</i>	5	4	4	5	
	3	3	3	1	
	1	1	1	1	
<i>b3</i>	4	2	4	4	
	1	2	5	1	
	2	3	5	2	

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

- $(s2, s3)$ from the second band and $(s1, s4)$ from the third band. The similarity for these pairs can now be calculated, to check for actual match, given a certain similarity threshold. E.g., if the threshold is $2/3$, $(s1, s4)$ will be a found match.