Exercise 3

resulting permutations

a)

	s_1	s_2	s_3	s_4	h_1	h_2	h_3
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

=>

	s_1	s_2	s_3	s_4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

← Signature Matrix

b) Only h_3 is a true permutation function because it has no collisions (injective). As can be seen from the table above, h_1 and h_2 have multiple collisions:

h_1:
input: 0, 3 -> output: 1
input: 1, 4 -> output: 3
input: 2, 5 -> output: 5

h_2:
input: 0, 2, 4 -> output: 2
input: 1, 3, 5 -> output: 5

c) ~~Jaccard similarity~~

	S1–S2	S1–S3	S1–S4	S2–S3	S2–S4	S3–S4
Jaccard Sim.	0	0	1/4	0	1/4	1/4
h_1	0	0	0	1	1	1
h_2	1	1	1	1	1	1
h_3	0	0	1	0	0	0
Signature fraction	1/3	1/3	2/3	2/3	2/3	2/3

Exercise 4

Assume we have a true permutation h_π and we have the same definitions as on slide 38 of the recent lecture 07.
Then let the Jaccard similarity of C_1, C_2 be 0 $\Rightarrow$ $sim(C_1, C_2) = 0$

$$\Rightarrow \frac{a}{(a+b+c)} = 0$$

Since $a, b, c \geq 0$ and $a+b+c > 0$ because C_1, C_2 have at least one element we get

$\Rightarrow$ $a = 0$

$\Rightarrow$ minhash of C_1, C_2 will be different

thus $P[h(C_1) = h(C_2)] = 0$. $\square$

Exercise 5

a) Jaccard similarity $= \frac{S_1 \cap S_2}{S_1 \cup S_2} = \frac{1}{5}$

b) From the lecture we know that for a random true permutation π we have

$$P[h_\pi(S_1) = h_\pi(S_2)] = Sim(S_1, S_2) = \frac{1}{5}$$

That means that this has to be reflected in the proportion of all permutations (120 in this case) that make the two columns hash to the same value

$\Rightarrow \frac{1}{5} \cdot 120 = 24$ permutations that make the two columns hash to the same value.