PROBLEM 3

To solve the problem it is necessary to do an hypothesis on signature of the set $\boldsymbol{U}$: $sig(U)$.
Since $n \gg k$ we can estimate with an high probability that the elements $x \in U$ will be mapped on all the values of hash function.

Since for definition:
$sig(U) = (\, v_1(U), v_2(U), \dots, v_l(U)\,)$
where $v_i(U) = min_{\,x \in U} h_i(x)$
we can imagine $sig(U) \approx (\,0, 0, \dots, 0)$

Determined $sig(U)$ we can determine the *Signature Matrix*:

	A	B	U
v_1	$v_1(A)$	$v_1(B)$	0
v_2	$v_2(A)$	$v_2(B)$	0
…	…	…	…
v_l	$v_l(A)$	$v_l(B)$	0

Table 1: Signature Matrix

where $v_i(A)$ and $v_i(B)$ are assigned.

1) $\boldsymbol{|A| = n^* SIM(A, U)}$
   $\boldsymbol{|B| = n * SIM(B, U)}$

   where the Jacard similarity $SIM(A, U)$ and $SIM(B, U)$ are determined through *lsh* algorithm, acceding to *Signature Matrix*.

2) $|A \cup B| = |A| + |B| - |A \cap B|$

   The size of $|A|$ and $|B|$ is before established. The problem is to determine the size of the intersection between the two sets $|A \cap B|$.
   It is risolved calculating the Jacard similarity $SIM(A, B)$ through *lsh* algorithm and imposing the system equation:

$$\begin{cases} |A \cup B| = |A| + |B| - |A \cap B| \\ SIM\,(A, B) = \dfrac{|A \cup B|}{|A \cap B|} \end{cases}$$

$$\begin{cases} |A \cup B| = |A| + |B| - |A \cap B| \\ |A \cap B| = \dfrac{|A \cup B|}{SIM\,(A, B)} \end{cases}$$

$$\begin{cases} |A \cap B| = \dfrac{|A \cup B|}{SIM\,(A, B)} \\ |A \cup B| = |A| + |B| - \dfrac{|A \cup B|}{SIM\,(A, B)} \end{cases}$$

$$\begin{cases} |A \cap B| = \dfrac{|A \cup B|}{SIM\,(A,B)} \\ |A \cup B| + \dfrac{|A \cup B|}{SIM\,(A,B)} = |A| + |B| \end{cases}$$

$$\begin{cases} |A \cap B| = \dfrac{|A \cup B|}{SIM\,(A,B)} \\ |A \cup B|(1 + \dfrac{1}{SIM\,(A,B)}) = |A| + |B| \end{cases}$$

$$\begin{cases} |A \cap B| = \dfrac{|A \cup B|}{SIM\,(A,B)} \\ |A \cup B| = \dfrac{|A| + |B|}{(1 + \dfrac{1}{SIM\,(A,B)})} \end{cases}$$

$$\begin{cases} |A \cup B| = \dfrac{|A| + |B|}{(1 + \dfrac{1}{SIM\,(A,B)})} \\ |A \cap B| = \dfrac{|A| + |B|}{SIM\,(A,B)(1 + \dfrac{1}{SIM\,(A,B)})} \end{cases}$$

$$\begin{cases} |A \cup B| = \dfrac{|A| + |B|}{(1 + \dfrac{1}{SIM\,(A,B)})} \\ |A \cap B| = \dfrac{|A| + |B|}{1 + SIM\,(A,B)} \end{cases}$$

$$\boldsymbol{|A \cup B| = \dfrac{|A| + |B|}{(1 + \dfrac{1}{SIM\,(A,B)})}}$$

3) HAMMING SIMILARITY

$$S_H = 1 - \frac{|(A \backslash B) \cup (B \backslash A)|}{n}$$

$|(A \backslash B) \cup (B \backslash A)|$ can be written as:

$|(A \backslash B) \cup (B \backslash A)| = |A \cup B| - |A \cap B|$

where:

$$\begin{cases} |A \cup B| = \dfrac{|A| + |B|}{(1 + \dfrac{1}{SIM\ (A,B)})} \\ |A \cap B| = \dfrac{|A| + |B|}{1 + SIM\ (A,B)} \end{cases}$$

At this point, the hamming similarity can be written $\boldsymbol{S_H}$:

$$S_H = 1 - \frac{|(A\backslash B) \cup (B\backslash A)|}{n}$$

$$S_H = 1 - \frac{|A \cup B| - |A \cap B|}{n}$$

$$S_H = 1 - \frac{\dfrac{|A| + |B|}{(1 + \dfrac{1}{SIM\ (A,B)})} - \dfrac{|A| + |B|}{1 + SIM\ (A,B)}}{n}$$

$$S_H = 1 - \frac{(|A| + |B|) * (\dfrac{SIM\ (A,B)}{1 + SIM\ (A,B)} - \dfrac{1}{1 + SIM\ (A,B)})}{n}$$

$$\boldsymbol{S_H = 1 + \frac{(|A| + |B|)\ \dfrac{1 - SIM\ (A,B)}{1 + SIM\ (A,B)}}{n}}$$