Intelligent Data Mining - Exercise 4

Michael Debono Mrđen

14 November 2017

1 Assignment 1: Minhashing

a. Compute the minhash signature for each column if we use the following three has functions:

- $h_1(x) = 2x + 1 \mod 6$
- $h_2(x) = 3x + 2 \mod 6$
- $h_3(x) = 5x + 2 \mod 6$

Element	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

b. Which of these hash functions are true permutations?

Only h_3 is a true permutation as it defines different hashes for all available elements.

c. How close are the estimated Jaccard similarities (based on the minhashes) for the six pairs of columns to the true Jaccard similarities.

Signature matrix computation:

	S_1	S_2	S_3	S_4
h_1	∞	∞	∞	∞
h_2	∞	∞	∞	∞
h_3	∞	∞	∞	∞

(1)

	S_1	S_2	S_3	S_4
h_1	∞	1	∞	1
h_2	∞	2	∞	2
h_3	∞	2	∞	2

(2)

	S_1	S_2	S_3	S_4
h_1	∞	1	∞	1
h_2	∞	2	∞	2
h_3	∞	1	∞	2

(3)

	S_1	S_2	S_3	S_4
h_1	5	1	∞	1
h_2	2	2	∞	2
h_3	0	1	∞	0

(4)

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	5	2
h_3	0	1	5	0

(5)

	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

(6)

	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

(7)

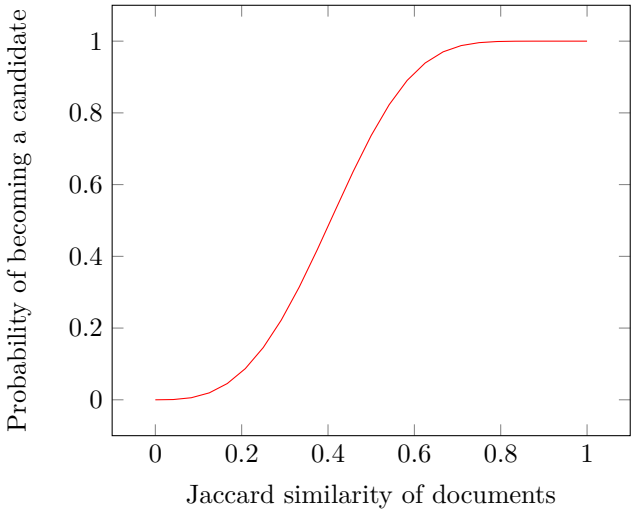
Jaccard similarities:

	Estimated	True	|Difference|
S_1, S_2	1/3	0	0.33
S_1, S_3	1/3	0	0.33
S_1, S_4	2/3	1/4	0.42
S_2, S_3	2/3	0	0.67
S_2, S_4	2/3	1/4	0.42
S_3, S_4	2/3	1/4	0.42

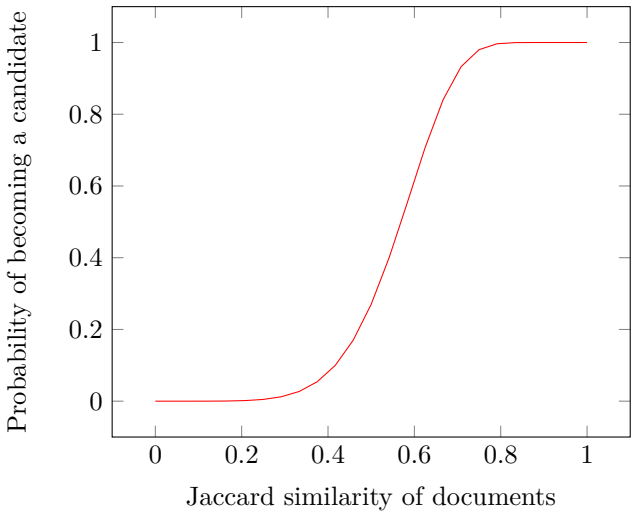
2 Assignment 2: Locality-sensitive hashing

a. Provide plots of the S-curve $1 - (1 - s^r)^b$ for the following values of r and b:

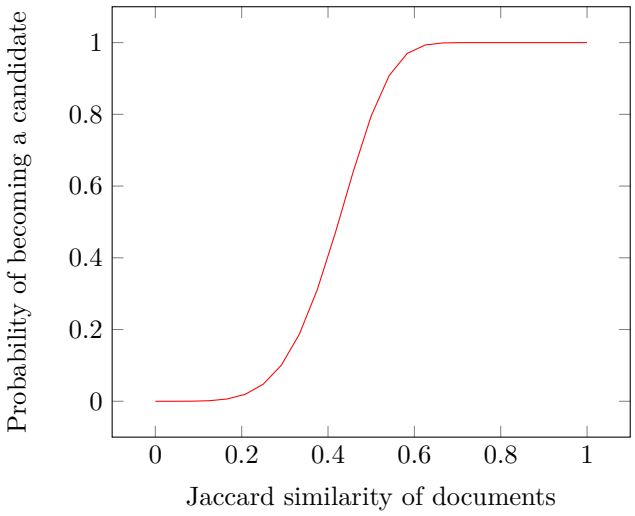
- $r = 3$ and $b = 10$



- $r = 6$ and $b = 20$



- $r = 5$ and $b = 50$



b. For each of the (r, b) pairs in (a), compute the threshold, that is, the value of s for which the value of $1 - (1 - s^r)^b$ is exactly $1/2$. How does this value compare with the estimate of $(1/b)^{1/r}$ that was suggested in Section 3.4.2?

	s when $1 - (1 - s^r)^b = 1/2$	$(1/b)^{1/r}$	|Difference|
$r = 3, b = 10$	0.40609	0.46415	0.058
$r = 6, b = 20$	0.56935	0.60696	0.038
$r = 5, b = 50$	0.42439	0.45730	0.033

3 Assignment 3: Minhashing in Java

This assignment is done separately.