Intelligent Data Management - Exercise 4

Name: Triet Ho Anh Doan
Date: November 12, 2017

Assignment 1: Minhashing

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

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

Element	S1	S2	S3	S4	$h_1(x)$	$h_2(x)$	$h_3(x)$
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. $h_3(x)$ gives true permutation.

c. Signature matrix and similarity:

	S1	S2	S3	S4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

	Estimated Similarity	True Similarity
$(S1, S2)$	1/3	0
$(S1, S3)$	1/3	0
$(S1, S4)$	2/3	1/4
$(S2, S3)$	2/3	0
$(S2, S4)$	2/3	1/4
$(S3, S4)$	2/3	1/4

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. Compute and estimate the threshold of the S-curve:

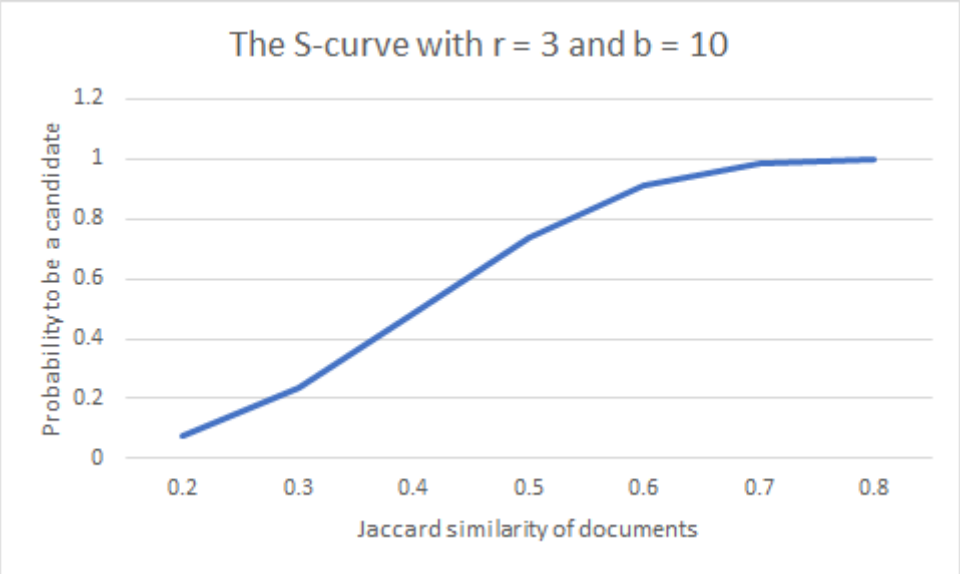

Figure 1: $r = 3$ and $b = 10$

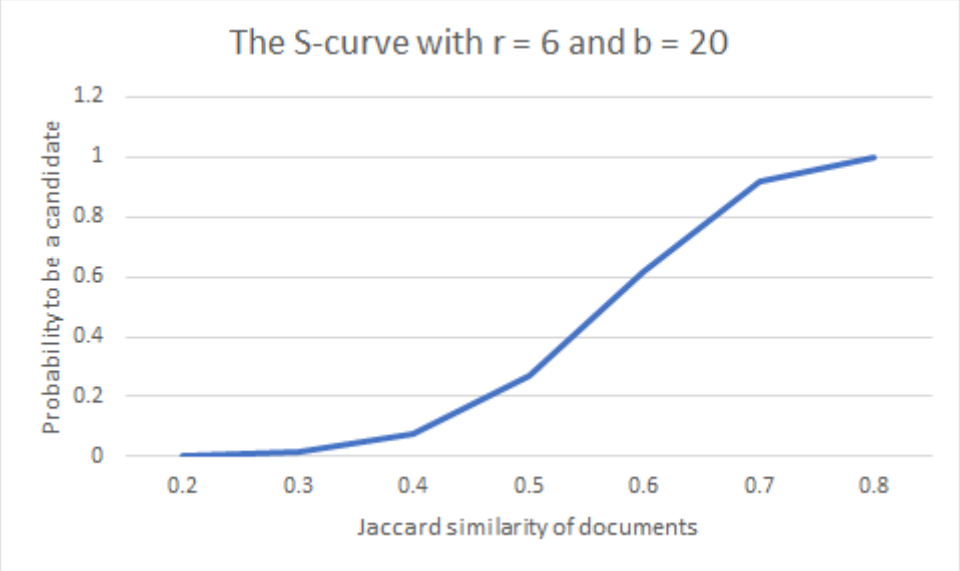

Figure 2: $r = 6$ and $b = 20$

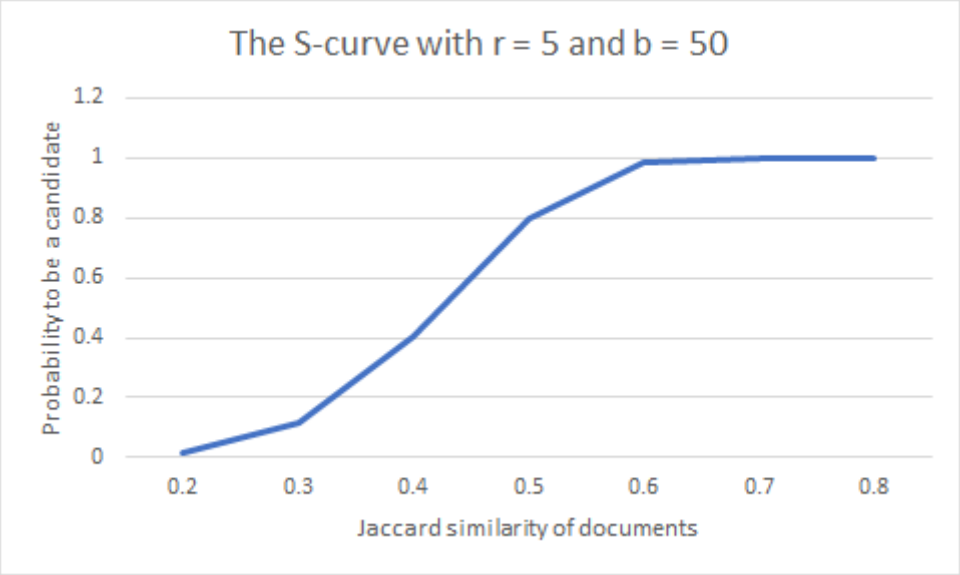

Figure 3: $r = 5$ and $b = 50$

	Computed Threshold	Estimated Threshold
$r = 3$ and $b = 10$	0.406	0.464
$r = 6$ and $b = 20$	0.569	0.607
$r = 5$ and $b = 50$	0.424	0.457