Your name: ______________________________

B565 MinHash practice (Fall 2023)

1. We will use a toy corpus of documents as an example for this practice. The documents are written in a small alphabet of only three letters $\{a, b, c\}$.

2. Here is the corpus: d1 = abcabcabc, d2 = aaabbbabb, d3 = cacacaca, d4 = abcabcaab.

3. We consider 2-shingles, and a simple hash function that converts a shingle (s) into an integer: $h(s) = idx(s[0]) + idx(s[1]) * 3$ (here idx('a') = 0, idx('b') = 1, and idx('c') = 2). So here are all unique shingles (and their corresponding IDs): aa (0), ba (1), ca(2), ab(3), bb(4), cb(5), ac(6), bc(7), and cc(8).

4. The corpus can be represented as a shingle(word)-document matrix below,

$ID(shingle)$	$d1$	$d2$	$d3$	$d4$
$0(aa)$	0	1	0	1
$1(ba)$	0	1	0	0
$2(ca)$	1	0	1	1
$3(ab)$	1	1	0	1
$4(bb)$	0	1	0	0
$5(cb)$	0	0	0	0
$6(ac)$	0	0	1	0
$7(bc)$	1	0	0	1
$8(cc)$	0	0	0	0

5. Jaccard similarity between the documents: $jaccard(d1, d2) = 1/6$, $jaccard(d1, d4) =$_____, $jaccard(d2, d4) =$_____.

6. Use these three hash functions to compute MinHash values: $h1(x) = (4x + 2)\%9$, $h2(x) = (7x + 5)\%9$, $h3(x) = (5x + 8)\%9$.

7. The signature-document matrix:

$hash$	$d1$	$d2$	$d3$	$d4$
$h1$	1	0	1	1
$h2$	0	3		
$h3$	0	1		

The first signature for d1 is $min(h1(2), h1(3), h1(7)) = min(1, 5, 3) = 1$, and so on.

8. Similarity based on signatures: s(d1, d2) = ,

9. Is h1 a true permutation? How about h2 and h3?