DS 702 - ASSIGNMENT 2

Shingling

1. What are the first ten 3-shingles in the first sentence of Section 3.2? [1]

   Since shingles can be a set of words or characters so here I have shown both.

   I. The most effective
   II. most effective way
   III. effective way to
   IV. way to represent
   V. to represent documents
   VI. represent documents as
   VII. documents as sets
   VIII. as sets for
   IX. sets for the
   X. for the purpose

   1) “The”
   2) “he “
   3) “e m”
   4) “ mo”
   5) “mos”
   6) “ost”
   7) “st “
   8) “t e”
   9) “ ef”
   10) “eff”

2. If we use the stop-word-based shingles of Section 3.2.4 [1], and we take the stop words to be all the words of three or fewer letters, then what are the shingles in the first sentence of Section 3.2 [1]?

    I. The most effective
    II. way to represent
    III. to represent documents
    IV. as sets for
    V. for the purpose
    VI. the purpose of
    VII. of identifying lexically
    VIII. is to construct
    IX. to construct from
    X. the document the
    XI. the set of
    XII. set of short
    XIII. of short strings

'It' wont be considered as it is at the end of the sentence thus can't be a start for a 3-shingle unless I include the next sentence.

3. What is the largest number of k-shingles a document of n bytes can have? You may assume that the size of the alphabet is large enough that the number of possible strings of length k is at least n.

    Each character is one byte thus N bytes in a document refers to N characters. Let's take an example of 5 characters "abcde" as a document, if we take 3-shingles from then we get 'abc', 'bcd', 'cde'. Thus, we can see that the largest number of k-shingles a document of N bytes can have is = $N - K + 1$

Hash Functions

Row	S_1	S_2	S_3	S_4	$x+1 \mod 5$	$3x+1 \mod 5$
0	1	0	0	1	1	1
1	0	0	1	0	2	4
2	0	1	0	1	3	2
3	1	0	1	1	4	0
4	0	0	1	0	0	3

a) $H_3(x) = 2x + 4 \% 5$

b) $H_4(x) = 2x + 4 \% 5$

Row	2x + 4%5	3x - 1% 5
0	4	4
1	1	2
2	3	0
3	0	3
4	2	1

MinHash Signatures

a) Compute the minhash signature for each column if we use the following three hash functions: h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x) = 5x + 2 mod 6.

Row	S1	S2	S3	S4	2x+1%6	3x+2%6	5x+2%6
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

		S1	S2	S3	S4
Row 0	h1(0)	INF	1	INF	1
	h2(0)	INF	2	INF	2
	h3(0)	INF	2	INF	2
Row 1	h1(1)	INF	1	INF	1
	h2(1)	INF	2	INF	2
	h3(1)	INF	1	INF	2
Row 2	h1(2)	5	1	INF	1
	h2(2)	2	2	INF	2
	h3(2)	0	1	INF	0
Row 3	h1(3)	5	1	1	1
	h2(3)	2	2	5	2
	h3(3)	0	1	5	0
Row 4	h1(4)	5	1	1	1
	h2(4)	2	2	2	2
	h3(4)	0	1	4	0
Row 5	h1(5)	5	1	1	1
	h2(5)	2	2	2	2
	h3(5)	0	1	4	0
Final MinHash Signature		5 2 0	1 2 1	1 2 4	1 2 0

b) Which of these hash functions are true permutations?

h3 minhash function is the only true permutation as there are no collisions

Part 2 Distances, Part 3 Shingles and Part 4 Misleading Profile Section all on the attached notebook.