BDA

Assignment-1 group- C7

(Prince, Bhautik, Lyon)

1. Shingles:-

(a) Here 2-shingles for ABRACADABRA are;

$$D_1 = \{AB, AC, AD, BR, CA, DA, RA\}$$

so $|D_1| = 7$

(b) 2-shingles for BRICABRAC are,

$$D_2 = \{AB, AC, BR, CA, IC, RA, RI\}$$

so $|D_2| = 7$

(c) The common 2-shingles are;

$$[D_1 \cap D_2] = \{AC, AB, BR, CA, RA\}$$

$$\therefore |D_1 \cap D_2| = 5$$

(d) The Jaccard similarity is defined as;

~~$SIM(D_1 \cap D_2)$~~

$$SIM(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} = \frac{5}{9}$$

$$\boxed{SIM(D_1, D_2) = \frac{5}{9}}$$

2. Minhashing :-

(a) Permuting rows of characteristic matrix

① using permutation vector.
$(4, 5, 0, 2, 3, 1)^T$

A characteristic matrix of four sets (C_1, C_2, C_3, C_4) over universal set $\{0, 1, 2, 3, 4, 5\}$ and permutation of it's rows,

$0 \to 4, 1 \to 5, 2 \to 0, 3 \to 2, 4 \to 3, 5 \to 1$
is given as;

Row	C_1	C_2	C_3	C_4
0	0	1	1	0
1	1	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	1	0
5	0	1	0	0

→

Row	C_1	C_2	C_3	C_4
4	1	0	1	0
5	0	1	0	0
0	0	1	1	0
2	0	1	0	1
3	0	0	1	0
1	1	0	1	1

The minhash function h_π on C is defined by,

$$h_\pi(c) = \min_{i \in \{1, \ldots m\}} \{\pi(i) / c[i] = 1\}$$

where we are given;

- a characteristic matrix with m rows = 6 and column c.
- a permutation π on the rows, that is $\pi: \{1, \ldots m\} \to \{1, \ldots m\}$ is a bijection. (1 corresponds to 0)

π: $4 \to 0, 5 \to 1, 0 \to 2, 2 \to 3, 3 \to 4, 1 \to 5$

so; $h_\pi(C_1) = 0$, $h_\pi(C_2) = 1$, $h_\pi(C_3) = 0$, $h_\pi(C_4) = 3$

(2) Using the permutation vectors; $(3,1,0,5,2,4)^T$

π: $3 \to 0, 1 \to 1, 0 \to 2, 5 \to 3, 2 \to 4, 4 \to 5$

Permutations of it's row is given by:

Row	C_1	C_2	C_3	C_4
0	0	1	1	0
1	1	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	1	0
5	0	1	0	0

$\longrightarrow$

Row	C_1	C_2	C_3	C_4
3	0	0	1	0
1	1	0	1	1
0	0	1	1	0
5	0	1	0	0
2	0	1	0	1
4	1	0	1	0

so; $h_\pi(C_1) = 1$, $h_\pi(C_2) = 2$, $h_\pi(C_3) = 0$, $h_\pi(C_4) = 1$.

(counting from 0).

(b)

Row	S_1	S_2	S_3	S_4	$h_1(x)$ $2x+1 \bmod 7$	$h_2(x)$ $3x+2 \bmod 7$	$h_3(x)$ $5x+2 \bmod 7$
0	0	1	0	1	1	2	2
1	0	0	0	1	3	5	0
2	1	1	0	0	5	1	5
3	0	0	1	0	0	4	3
4	0	1	1	0	2	0	1
5	1	0	0	0	4	3	6
6	1	0	1	0	6	6	4

First iteration gives:

$$SIG_{12} = SIG_{14} = \min(\infty, h_1(0)\} = 1$$

$$SIG_{22} = SIG_{24} = \min\{\infty, h_2(0)\} = 2$$

$$SIG_{32} = SIG_{34} = \min(\infty, h_3(0)\} = 2$$

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

Similarly second iteration:

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

Third iteration;

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

fourth iteration

	S_1	S_2	S_3	S_4
h_1	5	1	0	1
h_2	1	1	4	2
h_3	5	2	3	0

~~fourth~~ fifth iteration

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

~~fifth~~ Sixth iteration

	S_1	S_2	S_3	S_4
h_1	4	1	0	1
h_2	1	0	0	2
h_3	5	1	1	0

Seventh / final iteration.

	S_1	S_2	S_3	S_4
h_2	4	1	0	1
h_2	1	0	0	2
h_3	4	1	1	0

(b) here h_1 and h_3 are permutations.

(c) Jaccard similarities.

$$SIM(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = 0 = SIM(S_1, S_3)$$

$$= SIM(S_1, S_4)$$

$$= SIM(S_3, S_4)$$

$SIM(S_2, S_3) = \frac{2}{3}$,

$SIM(S_2, S_4) = \frac{1}{3}$.

True Jaccard Similarity:

$$SIM(S_1, S_2) = \frac{x}{x+y} = \frac{1}{4+1} = \frac{1}{5}.$$

$SIM(S_1, S_3) = \frac{1}{5}$, $SIM(S_2, S_4) = \frac{1}{4}$

$SIM(S_1, S_4) = 0$ $SIM(S_3, S_4) = 0$

$SIM(S_2, S_3) = \frac{1}{5}$

for pair (S_1, S_4) & (S_3, S_4) it's equal

Pair (S_2, S_3) has a difference of $\frac{1}{12}$

which is small, so this pair has close estimated and True Jaccard Similarity.

(d) Benifits of using hash functions instead of permutations:

- Requires less memory.
- easy to generate
- No memorization required.

- Permutation require iteration over whole Problem space while for hash function we can apply iteratively

- Some hash functions come with the guarantee of uniformality, randomness

- Permutation is time consuming for large number of rows while hash function is easier as it is not a bijection