Dorina Ismaili & Deniz Gaiser
09.12.2024

Mining Massive Datasets
Homework 6

Exercise 1

The Jaccard similarity can be applied to sets of elements. Sometimes, documents (or other objects) may be represented as multi-sets/bags rather than sets. In a multi-set, an element can be a member more than once, whereas a set can only hold each element at most once. Try to define a similarity metric for multi-sets. This metric should take exactly the same values as Jaccard similarity in the special case where both multi-sets are in fact sets.

The Jaccard similarity for sets is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

for the multisets we need to redefine the operations of intersection and union to account for element multiplicities. For each element x in the multi-sets, the multiplicity in the intersection is the minimum of its multiplicities in A and B

$$(A \cap B)(x) = \min(A(x), B(x)).$$

For each element x, the multiplicity in the union is the maximum of its multiplicities A and B

$$(A \cup B)(x) = \max(A(x), B(x)).$$

Therefore the generalized Jaccard similarity for the multisets is

$$J(A,B) = \frac{\sum_{x \in (A \cup B)} \min(A(x), B(x))}{\sum_{x \in (A \cup B)} \max(A(x), B(x))}$$

Exercise 2

See Task2.py

Exercise 3

Representing four sets $S1, S2, S3$ and $S4$ (subsets of 0, 1, 2, 3, 4, 5)

Element	S_1	S_2	S_4	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

(a) Compute the MinHash signature for each set using the following three hash functions:

Element(x)	$h_1(x) = (2x+1) \bmod 6$	$h_2(x) = (3x+2) \bmod 6$	$h_3(x) = (5x+2) \bmod 6$
0	(2*0)+1 mod 6 = 1	(3*0)+2 mod 6 = 2	(5*0) +2 mod 6 = 2
1	(2*1)+1 mod 6 = 3	(3*1)+2 mod 6 = 5	(5*1) +2 mod 6 = 1
2	(2*2)+1 mod 6 = 5	(3*2)+2 mod 6 = 2	(5*2) +2 mod 6 = 0
3	(2*3)+1 mod 6 = 1	(3*3)+2 mod 6 = 5	(5*3) +2 mod 6 = 5
4	(2*4)+1 mod 6 = 3	(3*4)+2 mod 6 = 2	(5*4) +2 mod 6 = 4
5	(2*5)+1 mod 6 = 5	(3*5)+2 mod 6 = 5	(5*5) +2 mod 6 = 3

Hash each element using the hash functions h_1, h_2, h_3. To compute the MinHash-signature, we use the minimum value of the corresponding set.

Set	Element	h_1	h_2	h_3
S_1	{2,5}	min(5,5) = 5	min(2,5) = 2	min(0,3) = 0
S_2	{0,1}	min(1,3) = 1	min(2,5) = 2	min(2,1) = 1
S_3	{3,4}	min(1,3) = 1	min(5,2) = 2	min(5,4) = 4
S_4	{0,2,4}	min(1,5,3) = 1	min(2,2,2) = 2	min(2,0,4) = 0

Final MinHash Signatures

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

(b) To determine whether a hash function is a true permutation, we check if the function produces a one-to-one mapping of the input set onto itself.
For $h_1(x) = (2x+1) \mod 6$, the output is {1,3,5,1,3,5}. It is not a true permutation because there is collision between elements.
For $h_2(x) = (3x+2) \mod 6$, the output is {2,5,2,5,2,5}. It is not a true permutation, because there is a collision between elements.
For $h_3 = (5x+2) \mod 6$, the output is {2,1,0,5,4,3}. It is a true element, as the elements are unique.

(c) To compare the similarity of MinHash signatures with the corresponding Jaccard similarities for the given sets, we proceed in two steps:
1. Calculate the Jaccard Similarity for Each Pair of Sets: The Jaccard similarity between two sets is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Using the sets $S_1 = \{2,5\}$, $S_2 = \{0,1\}$, $S_3 = \{3,4\}$, $S_4 = \{0,2,4\}$.
2. Compare Jaccard Similarity with MinHash Signature Similarity: The MinHash similarity is the fraction of matching values in the MinHash signatures for two sets.

Set	$S_i \cap S_j$	$S_i \cup S_j$	Jaccard Similarity
S_1, S_2	$\emptyset$	{0,1,2,5}	0/4 =0
S_1, S_3	$\emptyset$	{2,3,4,5}	0/4 = 0
S_1, S_4	{2}	{0,2,4,5}	1/4 = 0.25
S_2, S_3	$\emptyset$	{0,1,3,4}	0/4 = 0
S_2, S_4	{0}	{0,1,2,4}	1/4 = 0.25
S_3, S_4	{4}	{0,2,3,4}	1/4 = 0.25

Set	Matches	MinHash Similarity
S_1, S_2	{2}	1/3 =0.333
S_1, S_3	{2}	1/3 = 0.333
S_1, S_4	{0,2}	2/3 = 0.667
S_2, S_3	{1,2}	2/3 = 0.667
S_2, S_4	{1,2}	2/3 = 0.667
S_3, S_4	{1,2}	2/3 = 0.667

The MinHash similarity approximates the Jaccard similarity by comparing the MinHash signatures.

Exercise 4

See Task 4.py with attached results