Name : Pawan Suresh Bathe **Roll No.**: CS18M519

CS6745: Mining Massive DataSets

Tutorial 4

September 27, 2019

1. (2 marks) 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 as n.

Answer:

Assume size of alphabet is a , so $a^k >= n$, since it is given that number of possible strings of length k is at least as n.

So, $a = k\ th\ root\ of\ n$.

So, the total number of possible shingles of size k, are $a^k = (k\ th\ root\ of\ n)^k = n$ if the alphabet size is a.

Document size is n bytes, assuming character is of 1 bytes , the document can have n characters or words.

Largests number of $k - shingles = n - k + 1$ where n is the number of words/characters in the document.

Hence, the largest number of k-shingles a document of n bytes can have is $n - k + 1$

2. (5marks) Figure1 is a matrix with six rows.

$Element$	S_1	S_2	S_3	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

Figure 1: Matrix for Q2

(a) Compute the minhash signature for each column if we use the following three hash functions:

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

Answer:

Hash Functions:

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

For x= 0 :

$h_1(x) = 1$ $h_2(x)= 2$ $h_3(x)=2$

For x= 1 :

$h_1(x) = 3$ $h_2(x)= 5$ $h_3(x)=1$

For x= 2 :

$h_1(x) = 5$ $h_2(x)= 2$ $h_3(x)=0$

For x= 3 :

$h_1(x) = 1$ $h_2(x)= 5$ $h_3(x)=5$

For x= 4 :

$h_1(x) = 3$ $h_2(x)= 2$ $h_3(x)=4$

For x= 5 :

$h_1(x) = 5$ $h_2(x)= 5$ $h_3(x)=3$

Element 0					
		S1	S2	S3	S4

	h1	NA	1	NA	1
	h2	NA	2	NA	2
	h3	NA	2	NA	2
Element 1					
		S1	S2	S3	S4
	h1	NA	1	NA	1
	h2	NA	2	NA	2
	h3	NA	1	NA	2
Element 2					
		S1	S2	S3	S4
	h1	5	1	NA	1
	h2	2	2	NA	2
	h3	0	1	NA	0
Element 3					
		S1	S2	S3	S4
	h1	5	1	1	1
	h2	2	2	5	2
	h3	0	1	5	0
Element 4					
		S1	S2	S3	S4
	h1	5	1	1	1
	h2	2	2	2	2
	h3	0	1	4	0
Element 5					
		S1	S2	S3	S4
	h1	5	1	1	1
	h2	2	2	2	2
	h3	0	1	4	0

Min has signature matrix is :

S1	S2	S3	S4
5	1	1	1
2	2	2	2
0	1	4	0

(b) Which of these hash functions are true permutations?

Ans: $h_3(x)=5x + 2 \bmod 6$ hash function is true permutation.

(c) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

Similarities	1-2	1-3	1-4	2-3	2-4	3-4
Col-Col	0	0	0.25	0	0.25	0.25
Sig-Sig	0.33	0.33	0.67	0.67	0.67	0.67

Estimated Jaccard similarities are not close to true Jaccard similarities.

3. (3 marks) Suppose we want to use a MapReduce framework to compute minhash signatures. If the matrix is stored in chunks that correspond to some columns, then it is quite easy to exploit parallelism. Each Map task gets some of the columns and all the hash functions, and computes the minhash signatures of its given columns. However, suppose the matrix were chunked by rows, so that a Map task is given the hash functions and a set of rows to work on. Design Map and Reduce functions to exploit MapReduce with data in this form.

Ans:

In map function multiply each element from hash function to each element of rows.

Create a mapping hash function element -> result of element * multiplied by each element of row (in each input row)

In reduce function , group elements by hash function elements from each chunk.

And sum all elements mapping to the same hash function element.

For all hash elements with same column corresponding to hash function element in previous element, sum them to and map it to corresponding column.