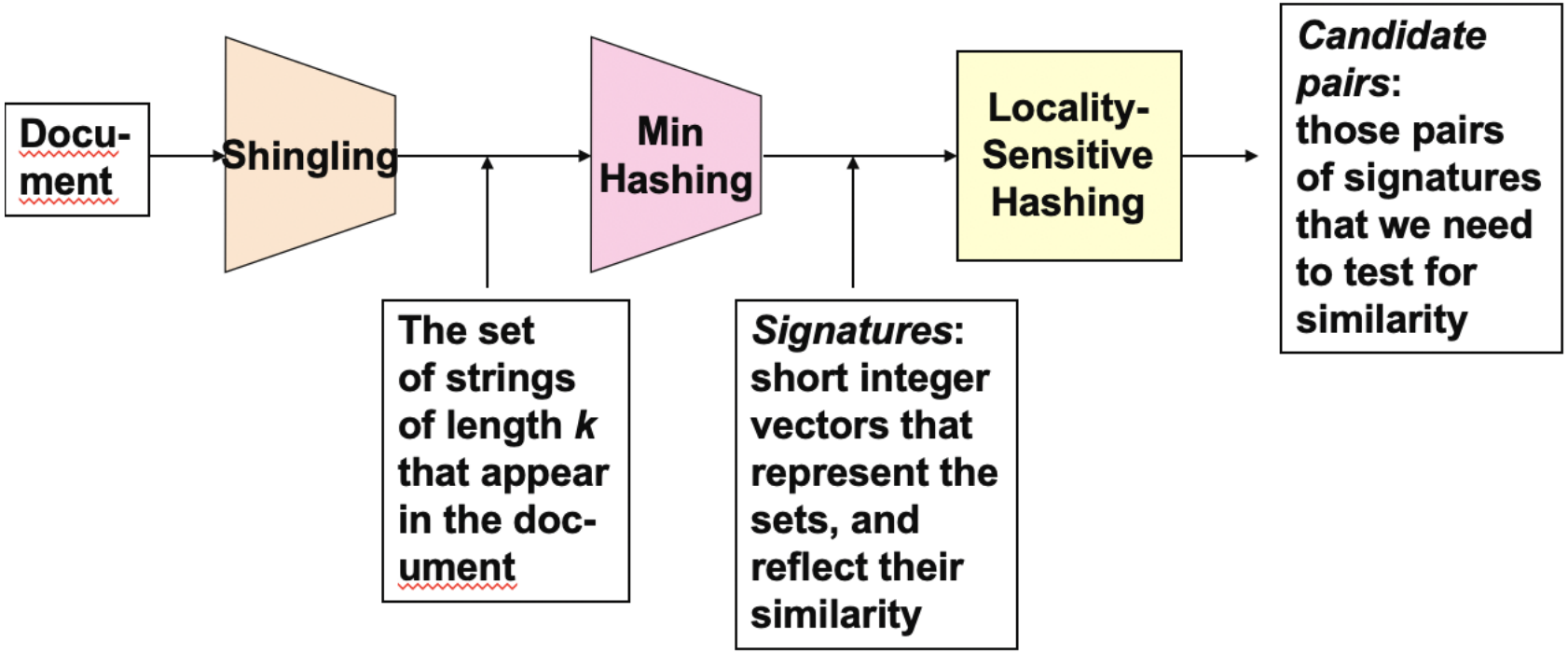
Tutorial 4:MapReduce – data mining

Li Yuan

li.yuan@u.nus.edu


Docu-
ment
Shingling
Min
Hashing
Locality-
Sensitive
Hashing
Candidate
pairs:
those pairs
of signatures
that we need
to test for
similarity
The set
of strings
of length k
that appear
in the doc-
ument
Signatures:
short integer
vectors that
represent the
sets, and
reflect their
similarity

Question 1

1. Find the set of 2-shingles for the "document":

ABRACADABRA

and also for the "document":

BRICABRAC

Answer the following questions:

i) How many 2-shingles does ABRACADABRA have?

ii) How many 2-shingles does BRICABRAC have?

iii) How many 2-shingles do they have in common?

iv) What is the Jaccard similarity between the two documents?

Solution 1a

ABRACADABRA

⇒{AB,BR,RA,AC,CA,AD,DA,AB,BR,RA}

⇒{AB,BR,RA,AC,CA,AD,DA}

7 2-shingles!

Solution 1b

BRICABRAC

- **{BR,RI,IC,CA,AB,BR,RA,AC}**
- **{BR,RI,IC,CA,AB,RA,AC}**

7 2-shingles!

Solution 1c

S1 = {AB, BR, RA, AC, CA, AD, DA}
S2 = {BR, RI, IC, CA, AB, RA, AC}

Common shingles:
S1∩S2 = {AB, BR, RA, AC, CA}

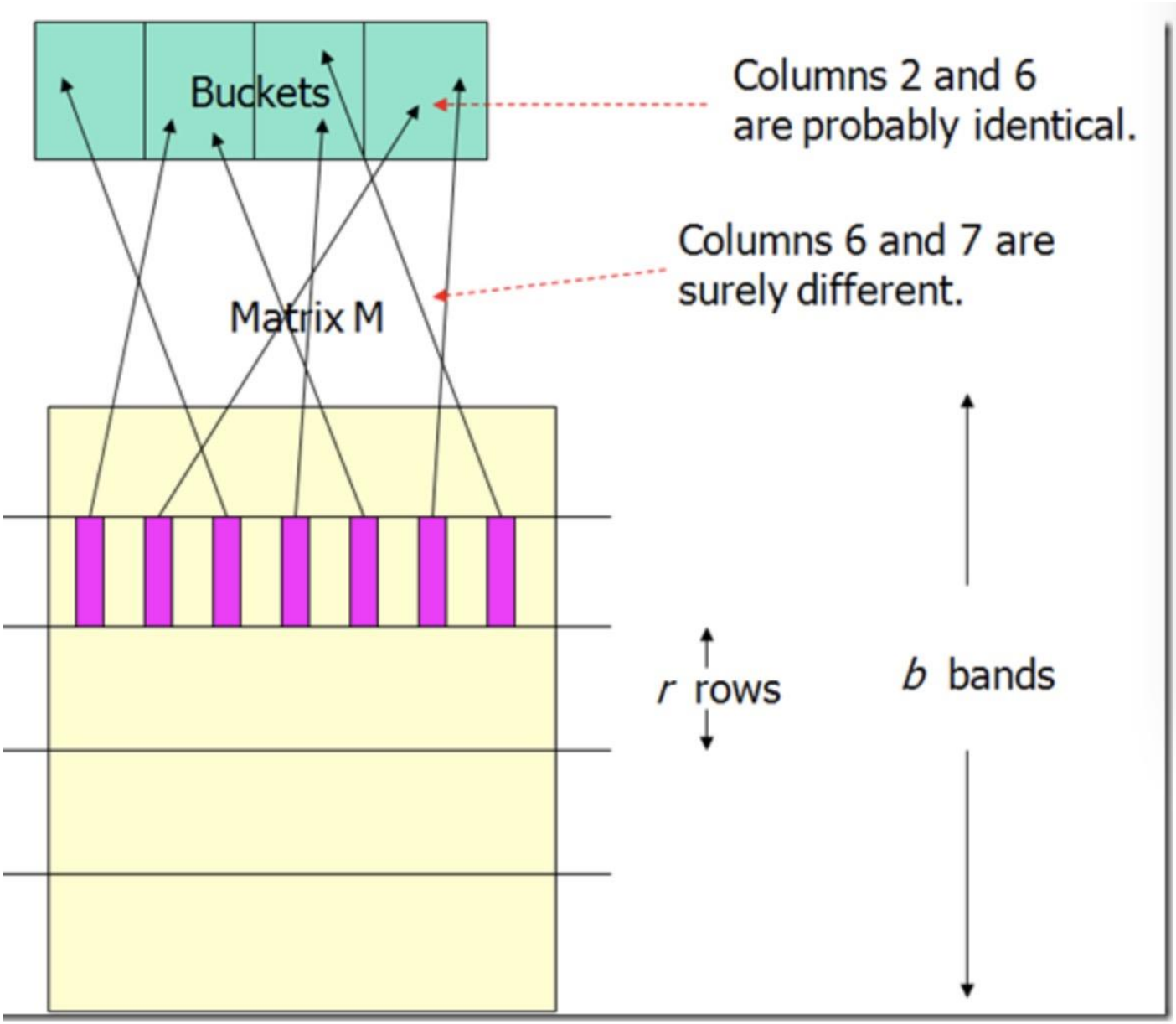
5 2-shingles!

Solution 1d

S1 = {AB, BR, RA, AC, CA, AD, DA}
S2 = {BR, RI, IC, CA, AB, RA, AC}

S1∩S2 = {AB, BR, RA, AC, CA}
S1∪S2 = {AB, BR, RA, AC, CA, AD, DA, RI, IC}

$$\text{Jaccard similarity} = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{5}{9}$$


Buckets
Columns 2 and 6
are probably identical.
Columns 6 and 7 are
surely different.
Matrix M
r rows
b bands

Question 2

Here is a matrix representing the signatures of seven columns, C1 through C7.

C1	C2	C3	C4	C5	C6	C7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Suppose we use locality-sensitive hashing with three bands of two rows each. Assume there are enough buckets available that the hash function for each band can be the identity function (i.e., columns hash to the same bucket if and only if they are identical in the band). Find all the candidate pairs.

C1	C2	C3	C4	C5	C6	C7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Pair 1: C1 and C4

Pair 2: C2 and C5

Pair 3: C1 and C6

Pair 4: C1 and C3

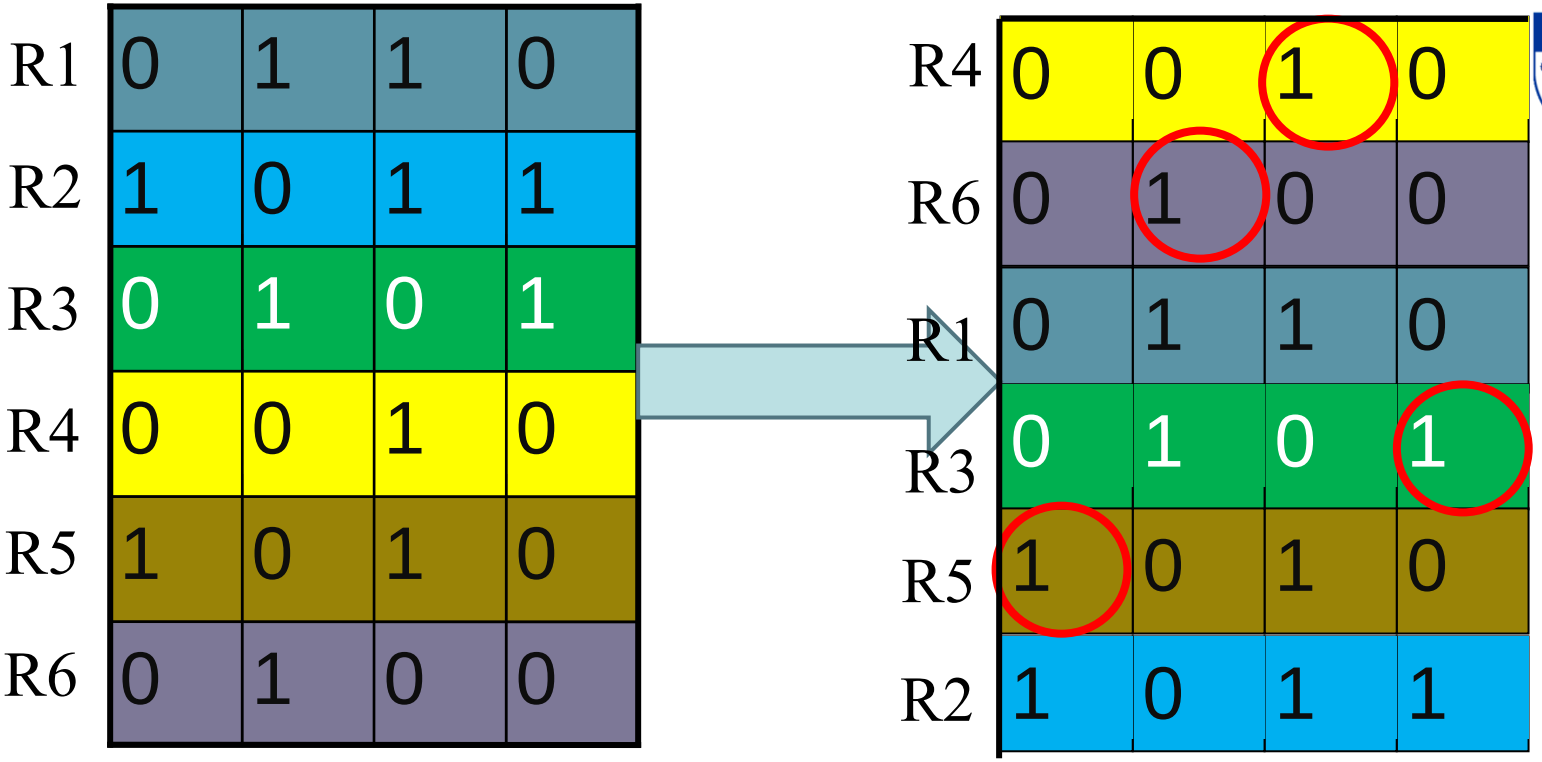
Pair 5: C4 and C7

Question 3

Consider the following matrix:

	$C1$	$C2$	$C3$	$C4$
$R1$	0	1	1	0
$R2$	1	0	1	1
$R3$	0	1	0	1
$R4$	0	0	1	0
$R5$	1	0	1	0
$R6$	0	1	0	0

Perform a minhashing of the data, with the order of rows: R4, R6, R1, R3, R5, R2. Compute the signature values of the four columns.



NUS National University of Singapore

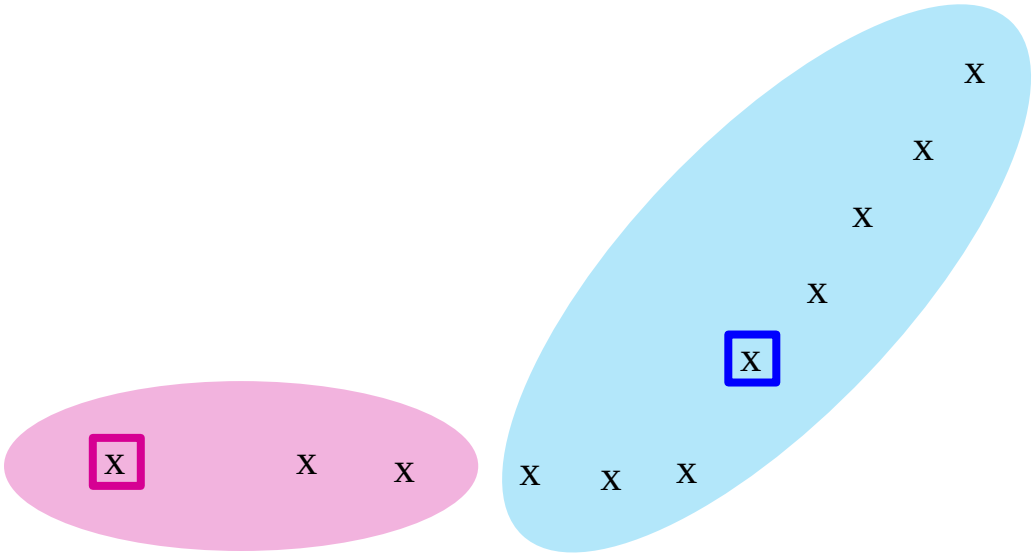
The signature value for C1 is 5 (R5)

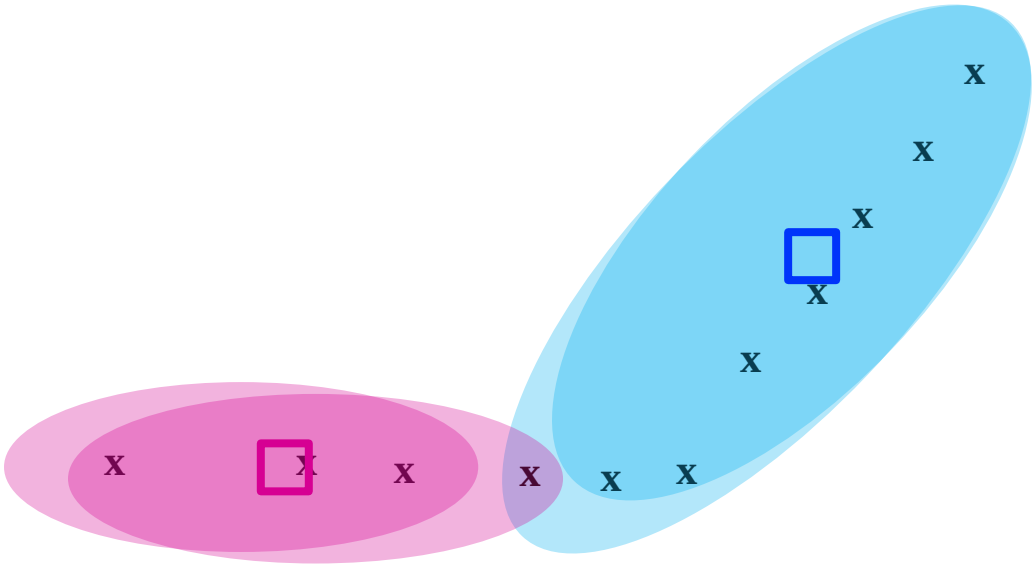
The signature value for C2 is 2 (R6)

The signature value for C3 is 1 (R4)

The signature value for C4 is 4 (R3)

K-means Algorithm





Question 4

We can cluster in one dimension as well as in many dimensions. In this problem, we are going to cluster numbers on the real line. The particular numbers (data points) are 1, 4, 9, 16, 25, 36, 49, 64, 81, and 100, i.e., the squares of 1 through 10. We shall use a k-means algorithm, with two clusters. You can verify easily that no matter which two points we choose as the initial centroids, some prefix of the sequence of squares will go into the cluster of the smaller and the remaining suffix goes into the other cluster. As a result, there are only nine different clusterings that can be achieved, ranging from {1}{4,9,...,100} through {1,4,...,81}{100}.

We then go through a reclustering phase, where the centroids of the two clusters are recalculated and all points are reassigned to the nearer of the two new centroids. For each of the nine possible clusterings, calculate how many points are reclassified during the re-clustering phase. List five pairs of initial centroids that results in exactly one point being reclassified.

Solution 4

Consider this case in the initial clustering:
{1,4,9,16,25}{36,49,64,81,100}

How many points are reclassified during the re-clustering phase?

The centroid for {1,4,9,16,25}: 11

The centroid for {36,49,64,81,100}: 66

(11+66)/2 = 38.5

36 will be reclassified to the first cluster.

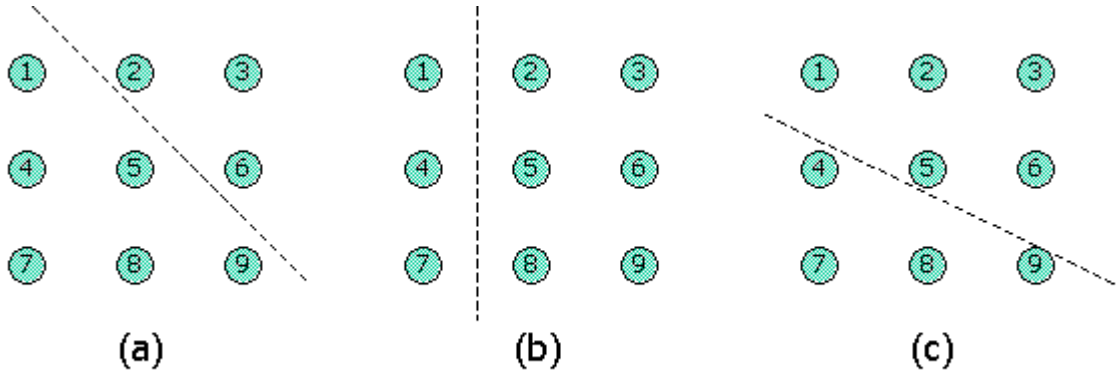
five pairs of initial centroids: (25,36),(16,36),(16,49),(9,49),(4,49)

Question 5

The Bisecting k-Means algorithm starts by dividing the points into two clusters. It may consider several bisections and pick the best one. Let us take "best" to mean the lowest SSE (Sum Squared Error). The SSE is defined to be the sum of the squares of the distances between each of the points of the cluster and the centroid of the cluster.

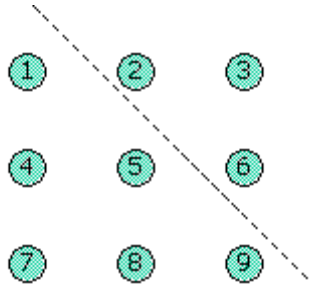
Suppose that the data set consists of nine points arranged in a square grid, as suggested by the figure below:

Question 5



Although it doesn't matter for this question, you may take the grid spacing to be 1 (i.e., the squares are 2-by-2) and the lower-left corner to be the point (0,0). We see in the figure three possible bisections. (a) would be the bisection if we chose the two initial centroids to be 3 and 7, for example, and broke ties in favor of 7. (b) would be the split if we chose initial centroids 1 and 2. (c) would be the split for initial choice 2 and 7. Rank these three options from the best to the worse choice.

Solution 5



(a)

The right cluster: (1,2),(2,1),(2,2)

The centroid: ((1+2+2)/3,(2+1+2)/3)=(5/3,5/3)

SSE: $(1-\frac{5}{3})^2+(2-\frac{5}{3})^2+\left(2-\frac{5}{3}\right)^2+\left(1-\frac{5}{3}\right)^2+\left(2-\frac{5}{3}\right)^2+\left(2-\frac{5}{3}\right)^2=\frac{4}{3}$

For the left cluster, the SSE is 20/3.

The total SSE is 8.

Solution 5

For (b), the centroids are (0,1) and (1.5,1), and the SSE's for these clusters are 2 and 5.5, respectively. Thus, this bisection has an SSE of 7.5.

For (c), the centroids are (0.75,0.25) and (1.2,1.6). The SSE's for these clusters are 3.5 and 4, respectively. Thus, this bisection also has an SSE of 7.5.

We conclude that (b) and (c) are equally good, and better than (a).

Question 6 (Enhance)

Suppose we have computed signatures for a number of columns.

There are N pairs of signatures that are 50% similar (i.e., they agree in half of the rows). There are M pairs that are 20% similar.

We can try to find 50%-similar pairs by using Locality-Sensitive Hashing (LSH), and we can do so by choosing b bands with r rows for each band. Calculate approximately the number of false positive and the number of false negatives, in terms of N, M, b and r.

FP & FN

False positive error:
A false positive error, or in short a false positive, commonly called a "false alarm", is a result that indicates a given condition exists, when it does not.

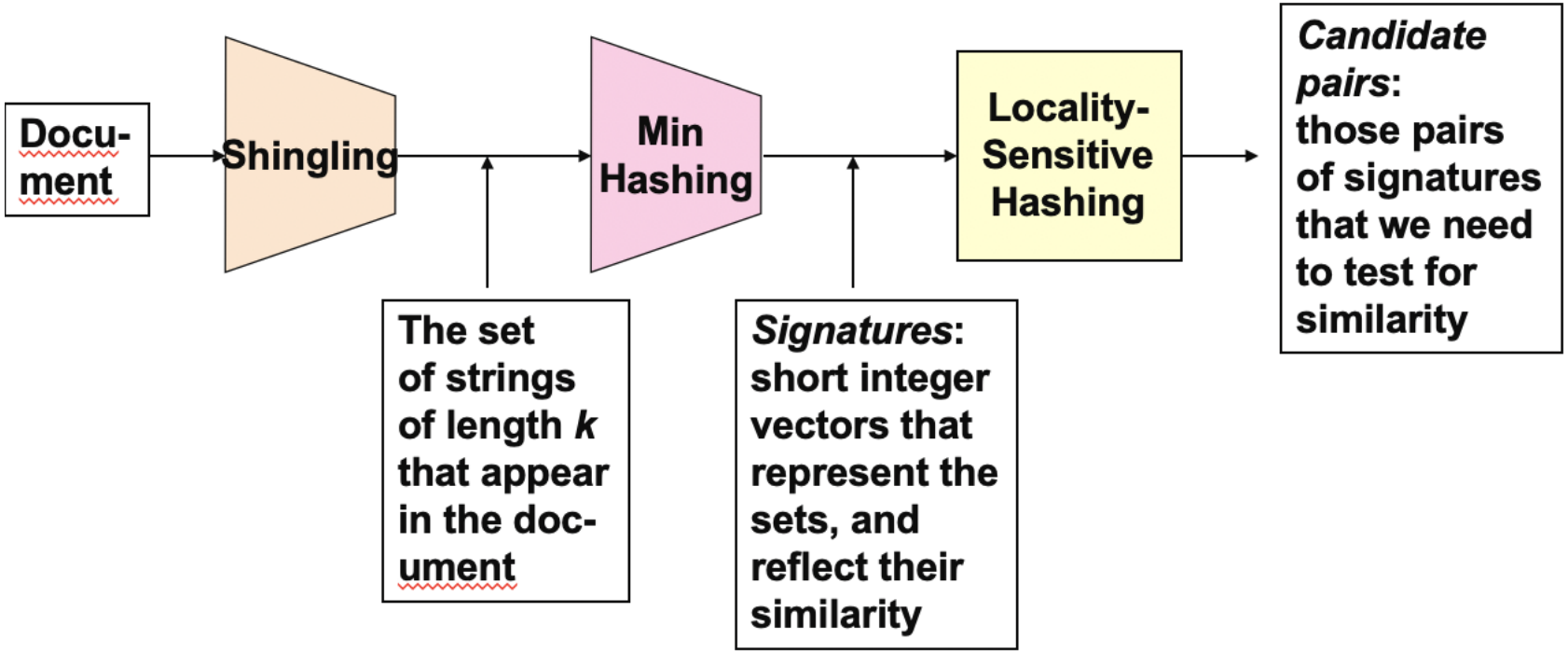
False negative error:
A false negative error, or in short a false negative, is a test result that indicates that a condition does not hold, while in fact it does.

Solution 6

- *Calculate False Positive*
  - As the similarity threshold $s = 50\%$, the M pairs of 20% similarity are actually NOT similar
  - Use these pairs to calculate no. of false positive (FP)
  - Set c_1, c_2 as 2 columns, $sim(c_1, c_2) = 20\% < s$
  - Then the probability that c_1, c_2 are identical in one band is: 0.2^r
  - The probability that c_1, c_2 are identical in at least 1 of b bands is: $1 - (1 - 0.2^r)^b$
- Therefore, no. of false positive is: $\mathrm{FP} = \mathrm{M} \cdot \left[1 - (1 - 0.2^{\mathrm{r}})^{\mathrm{b}}\right]$

Solution 6

- Calculate False Negative
  - As the similarity threshold $s = 50\%$, the N pairs of 50% similarity are actually similar
  - Use these pairs to calculate no. of false negative (FN)
  - Set c_1, c_2 as 2 columns, $sim(c_1, c_2) = 50\% \geq s$
  - Then the probability that c_1, c_2 are identical in one band is: 0.5^r
  - The probability that c_1, c_2 are not identical in all b bands is: $(1 - 0.5^r)^b$
- Therefore, no. of false negative is: $\mathrm{FN} = \mathrm{N} \cdot (1 - 0.5^{\mathrm{r}})^{\mathrm{b}}$


Docu-
ment
Shingling
Min
Hashing
Locality-
Sensitive
Hashing
Candidate
pairs:
those pairs
of signatures
that we need
to test for
similarity
The set
of strings
of length k
that appear
in the doc-
ument
Signatures:
short integer
vectors that
represent the
sets, and
reflect their
similarity

Acknowledgement

Thanks to Li Qinbin for making these slides.

liqinbin@u.nus.edu