CMPT741 - Data Mining
Assignment 2

Topic: Similar pairs
Submitted to: Professor Ke Wang
Submitted by: Raquel Yuri da Silveira Aoki
Date: November 17, 2017

Details of the assignment

1. M will be given to you as a text file, with n=6000 and m=1000.
2. Generate the signature matrix using p=100 random permutations through minhash
3. Determine the proper b and r for t=0.3 (i.e., 30%), where b is the number of bands and r is the number of rows per band, and b*r=p. Show the S curve to justify your choice.
4. Find candidate pairs of signatures using LHS, by choosing a hash function h with k=10,000 buckets.
5. Determine FP and FN of the result in 4, i.e., the number of dissimilar signature pairs that are candidate pairs, and the number of similar signature pairs that are not candidate pairs.
6. Find similar pairs of signatures by removing FP.
7. Find similar pairs of objects from the remaining candidate pairs in 6 and determine the FP and FN of this result.

Question 1

Report the result in the table below
Solution

Table 1: Results

	3(b,r)	4 (Number of pairs)	5(FP,FN)	6(number of pairs)	7 (FP,FN)
p = 100	(25,4)	34426	(29117,6679)	5309	(2350,4595)
p = 500	(100,5)	303113	(295916,1680)	7197	(2129,2486)

General comments about implementation of results:

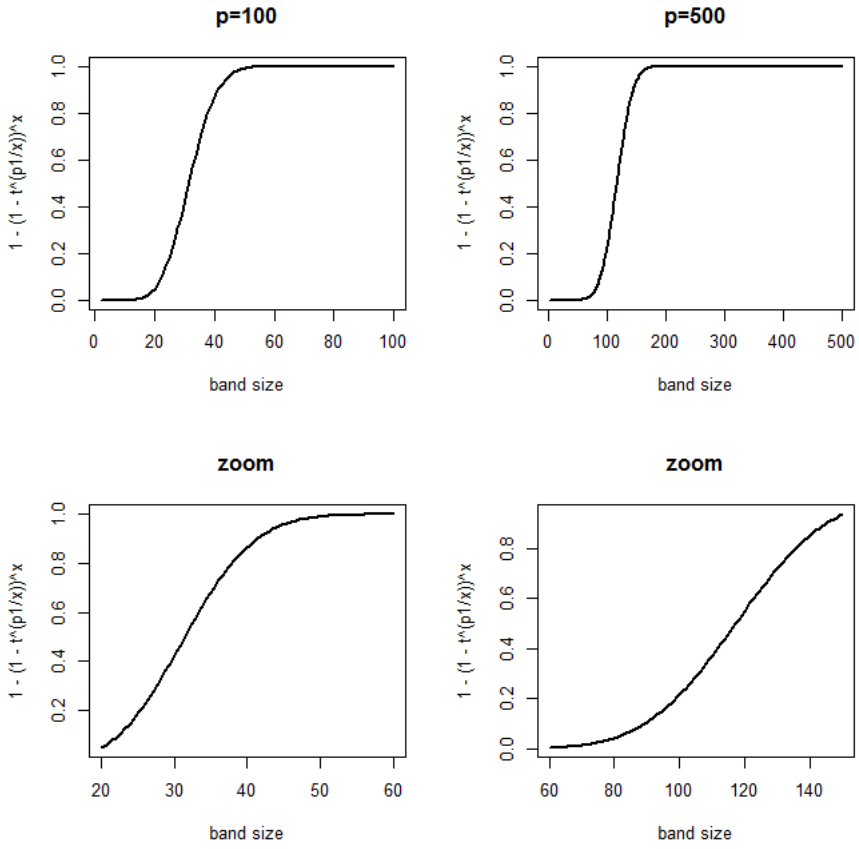

Figure 1: Item 3

1. File read

2. The signature matrix was generate using 100 and 500 permutations. In this stage it was used as hash function the position of the first element 1 in the column.

3. Figure 1 shows the results. The best b is on the transition of probability from 0 to 1. For $p = 100$, the best b is 25 ($r = 4$); for $p = 500$, the best b is 100 ($r = 5$).

4. For each set of r rows, it was create a string $(r_1 - r_2 - r_r)$ and applied the hash function $MD5$. The MD5 created is unique for each different string. To distribute those values among 10000 buckets, the strings were ordered and transformed into numbers, and then it was applied x $mod(10000)$. As Table 1 shows, with more permutations we have more candidate pairs in this step.

5. The pairs False Positive (FP) or False Negative (FN) were obtained as follow: first, it was calculated on the signature matrix the Jaccard Correlation between all pairs. Then, using $t = 0.3$, it was found all

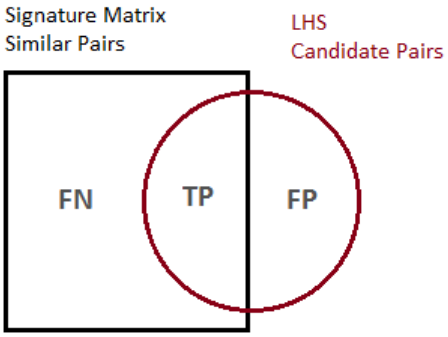

Figure 2: Part 5

similar pairs on *signature matrix*(SM). Those pairs can change according with the permutations. If a candidate pair is not present at similar pairs list of SM, then this pair is a FP; if similar pair of SM is not among the candidate pairs, then this pair is a FN; if a pairs is candidate pair and it is present on similar pairs of SM, then is a true positive. Figure 2 shows a scheme to this interpretation.

6. It was excluded from candidate pairs the pairs not present on similar pairs list from signature matrix.

7. It was used the same logic of Item 5, but now the candidate pairs were compared with similar pairs from object file. There are 7554 true similar pairs on this matrix. The similarity between the pairs were calculated once and saved in a file. Figure 3 illustrate the logic behind FN and FP.

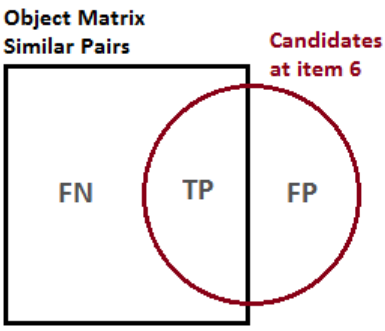

Figure 3: Part 7

For $p = 100$ permutations

Question 2

Draw the figure 1 for the FP and FN in 5 for different choices of b.

Question 3

Draw the figure 1 for the FP and FN in 7 for different choices of b.
Solution

Figure 4 shows the graphics for Questions 2 and 3, with FP and FN for different values of b (bands). The number of false positive(FP) pairs increases when b is increased (and r is decreased) in both questions 2 (compared with signature similar pairs) and 3(compared with true similar pairs). On the other hand, the number of false negative (FN) pairs decreases when b is increased. Once that more pairs are considered, it is less likely that some similar pairs are excluded.

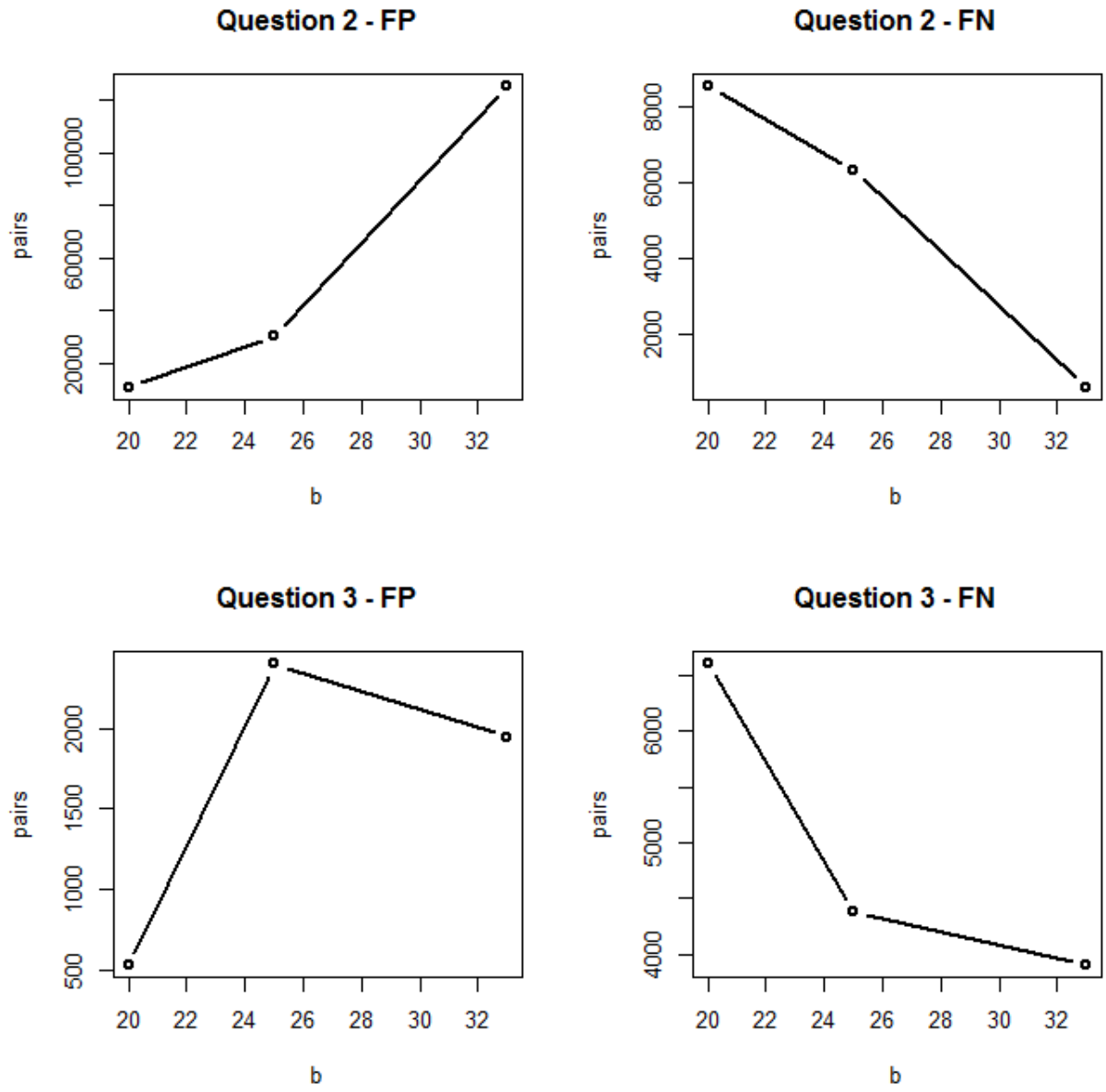

Figure 4: Questions 2 and 3