M1522.001400 Introduction to Data Mining

Solution of Midterm

Question 1.

(a) If we use only TF, then general words that appear in most documents will become important for all the documents.

(b) 13 is better, since 13 is a prime number which will map x uniformly into bins. Using 12 will not map x uniformly to bins.

(c) index

Question 2.

(a) 10

(b) 1) MapReduce replicates data to multiple machines, so that data in a failed machine are stored in other machines. 2) MapReduce reexecutes the failed tasks in another machine.

(c) Shorten job completion time.

(d) Decrease the intermediate data.

(e) (One of the followings; other answers are possible as well) 1) Computing average, 2) Computing median.

Question 3.

(a) 1/6

(b) Can take the order into consideration.

(c) N-k+1

(d) Compress the original data into short vectors (which can be used for similarity estimation).

(e) Decrease both false positive and negative.

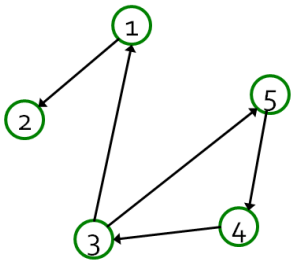
Question 4.

(a) Given a stream, estimate its k-th moments efficiently (with small memory space)
(b) The sampled data size grows continuously, eventually exceeding the memory limit.
(c) P(a) = P(j) = 3/10
(d) N
(e) [1010101] 0 [1010101][101]0[1]
(f) Decrease false positive
(g) (Any of the followings will get full score. Other answers about `distinct item count' are possible, too.) 1) counting the number of different words found among the Web pages crawled at a site . 2) counting the number of different web pages each customer request in a week. 3) counting the number of distinct products sold in the last week.
(h) We can say whether the data distribution is skewed or not.

Question 5.

(a) (Dead end) the PageRanks leak out. (Spider trap) The spider traps absorb all the pageranks.

(b) Allow teleport to a random node, occasionally.

(c) 

. We also give full scores to the solution which adds all possible edges from node 2 (the 'dead end' node).

(d) A =

0.04	0.2	0.44	0.04	0.04
0.84	0.2	0.04	0.04	0.04
0.04	0.2	0.04	0.84	0.04
0.04	0.2	0.04	0.04	0.84
0.04	0.2	0.44	0.04	0.04