MSC-BDT5002, Spring 2020

Knowledge Discovery and Data Mining

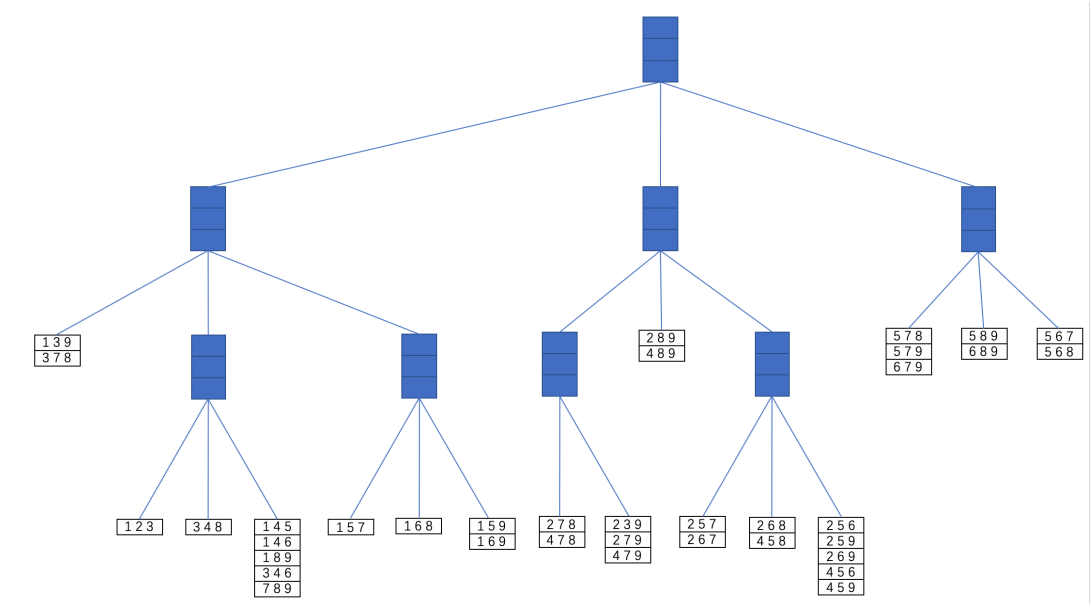
Assignment 1

HUANG XINXIAN 20630198

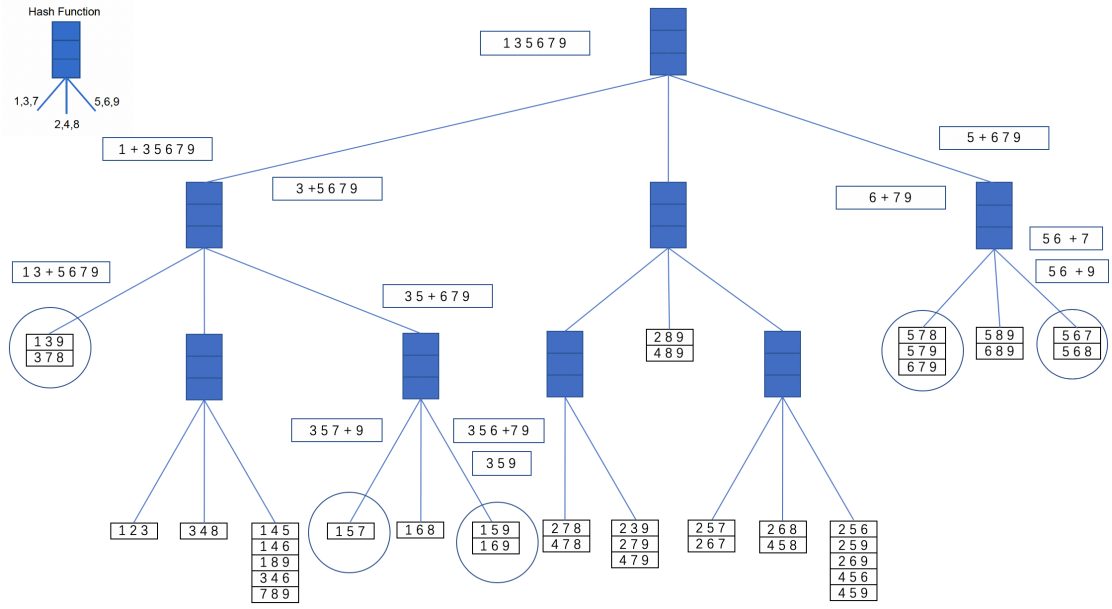
1. (a) The nested list of the hash tree is:

[[[[1, 3, 9], [3, 7, 8]], [[1, 2, 3], [3, 4, 8], [[1, 4, 5], [1, 4, 6], [1, 8, 9], [3, 4, 6], [7, 8, 9]]], [[1, 5, 7], [1, 6, 8], [[1, 5, 9], [1, 6, 9]]]], [[[[2, 7, 8], [4, 7, 8]], [[2, 3, 9], [2, 7, 9], [4, 7, 9]]], [[2, 8, 9], [4, 8, 9]], [[[2, 5, 7], [2, 6, 7]], [[2, 6, 8], [4, 5, 8]], [[2, 5, 6], [2, 5, 9], [2, 6, 9], [4, 5, 6], [4, 5, 9]]]], [[[5, 7, 8], [5, 7, 9], [6, 7, 9]], [[5, 8, 9], [6, 8, 9]], [[5, 6, 7], [5, 6, 8]]]]

The structure of the hash tree is:



(b) As shown in the following picture, 10 comparisons are needed.



2. (a) The program and outputs (pattern-i.txts) are in the folders of Q2.

pattern-0:

1288 data
1163 mining
1087 algorithm
720 graph
560 time
528 pattern
509 tree
488 efficient
416 rule
413 data mining

pattern-1

1488 learning
1050 using
819 model
715 based
582 classification
488 feature
474 clustering
463 network

pattern-2

1226 web
1211 information
1114 retrieval
863 based
757 system
707 search
564 document
490 language
475 information retrieval
421 model
414 semantic

pattern-3

1074 database
928 system
743 knowledge
558 learning
514 data
506 logic
493 reasoning
446 model
426 constraint

pattern-4

1713	query
1163	database
1040	data
767	system
637	processing
567	distributed
528	object
508	efficient
458	xml

(b) There is no FP-conditional tree whose height is larger than 1. But there are some FPconditional trees whose height are exactly 1, shown as following:
Cond.FP-tree on “mining”: ['Null Set 1', ['data 413']]
Cond.FP-tree on “retrieval”: ['Null Set 1', ['information 475']]