Asg 1 COMPSCI 753

Student ID: 964150056

Name: Haolin Feng

UPI: hfen690

Part 1:

(a)

bitvector\_all.csv: 534 articles, 14927 features

bitvector\_query.csv: 4 articles, 14927 features

//features are the number of 1 and 0 in each articles

(b)total 20 hash function for level 1

(5351x + 11510) mod 14929 mod 14927

(6236x + 2407) mod 14929 mod 14927

(12257x + 12857) mod 14929 mod 14927

(304x + 80) mod 14929 mod 14927

(14208x + 7186) mod 14929 mod 14927

(8428x + 13132) mod 14929 mod 14927

(6118x + 11694) mod 14929 mod 14927

(10900x + 3032) mod 14929 mod 14927

(195x + 12596) mod 14929 mod 14927

(2634x + 6517) mod 14929 mod 14927

(13323x + 1537) mod 14929 mod 14927

(12464x + 163) mod 14929 mod 14927

(3645x + 14323) mod 14929 mod 14927

(9569x + 9875) mod 14929 mod 14927

(8254x + 2566) mod 14929 mod 14927

(135x + 2144) mod 14929 mod 14927

(2479x + 12979) mod 14929 mod 14927

(2185x + 12333) mod 14929 mod 14927

(1391x + 6049) mod 14929 mod 14927

(2299x + 6457) mod 14929 mod 14927

(c)

图表, 条形图

描述已自动生成

m= number of documents

图表

描述已自动生成

m = number of documents / 5 (for better value display)

Part 2:

(a)

Query 1:

['531', 1.0, 'entertainment']

['422', 0.0, 'tech']

['428', 0.0, 'tech']

['526', 0.0, 'tech']

['174', 0.0, 'entertainment']

Query 2

['532', 1.0, 'entertainment']

['26', 0.1, 'tech']

['267', 0.1, 'entertainment']

['86', 0.1, 'entertainment']

['301', 0.1, 'entertainment']

['203', 0.05, 'tech']

['314', 0.05, 'entertainment']

['12', 0.0, 'entertainment']

['89', 0.0, 'entertainment']

['391', 0.0, 'entertainment']

Query 3

['533', 1.0, 'tech']

['51', 0.2, 'tech']

['242', 0.2, 'tech']

['414', 0.15, 'entertainment']

['19', 0.1, 'tech']

['47', 0.1, 'tech']

['149', 0.1, 'tech']

['229', 0.1, 'tech']

['289', 0.05, 'entertainment']

['360', 0.05, 'tech']

Query 4

['534', 1.0, 'tech']

['64', 0.1, 'tech']

['306', 0.1, 'tech']

['183', 0.05, 'tech']

['240', 0.05, 'entertainment']

['371', 0.05, 'tech']

['91', 0.05, 'tech']

['79', 0.05, 'entertainment']

['184', 0.0, 'entertainment']

['160', 0.0, 'entertainment']

(b)

Query 1

('531', 1.0)

('145', 0.19718309859154928)

('70', 0.16778523489932887)

('314', 0.16538461538461538)

('130', 0.11458333333333333)

('412', 0.07239819004524888)

('317', 0.06451612903225806)

('503', 0.062146892655367235)

('35', 0.062111801242236024)

('321', 0.06030150753768844)

Query 2

('532', 1.0)

('419', 0.09411764705882353)

('42', 0.09401709401709402)

('520', 0.09227467811158799)

('524', 0.09205020920502092)

('197', 0.09117647058823529)

('251', 0.09117647058823529)

('221', 0.09047619047619047)

('501', 0.09047619047619047)

('125', 0.08924485125858124)

Query 3

('533', 1.0)

('64', 0.16666666666666666)

('306', 0.1662269129287599)

('117', 0.1391509433962264)

('484', 0.1391509433962264)

('275', 0.1372093023255814)

('408', 0.1372093023255814)

('495', 0.135632183908046)

('132', 0.12709832134292565)

('455', 0.12654320987654322)

Query 4

('534', 1.0)

('211', 0.12272727272727273)

('324', 0.11072664359861592)

('330', 0.11061946902654868)

('183', 0.11059907834101383)

('371', 0.11059907834101383)

('525', 0.1076923076923077)

('83', 0.09120521172638436)

('440', 0.09090909090909091)

('255', 0.08849557522123894)

(c)

Part II(a): 506.357 ms

Part II(b): 6888.975 ms

The query time in (b) is much longer than (a) which is around 13.6 times of (a) ‘s query time.

Part 3  
(a)

K=2:

F1-score = 0.167

Query time = 84.032 ms

K=4

F1-score = 0.195

Query time = 207.496 ms

K=8

F1-score = 0.163

Query time = 343.997 ms

Average query time for k=2,4,8: 211.842 ms

(b)

From the result of part III(a), the difference of F1-score between different hash size k is not far apart. However the query time is differ a lot. I would tune the hash size(k) between 2-4, and k=2 is more preferrable.