#COMP9313 project1 report

#ZANNING WANG

#zID: z5224151

Q1、Implementation details of your c2lsh(). Explain how your major transform function works.

Q2、Show the evaluation result of your implementation using your own test cases.

Q3、What did you do to improve the efficiency of your implementation?

Answer 1、

手机屏幕截图

描述已自动生成

In order to find the rdd-id which match the query, under the c2lsh()function, I build another two functions called check\_satisfy(), and check\_diff(), the function check\_satisfy() used to check the | data-query | < offset, if the data satisfy this equation, it will return true.; the fuction check\_diff() used to create a new list which store the difference between data and query.

Use data\_hashes.map to change the data\_hashes from [id,data\_hash] to [id, difference, data\_hash

Use data\_hashes.filter to filter the data similar to the query, after filtrate the data, the data\_hashes(rdd) only contain the data satisfy the query.

Use data\_hashes.map to map the data id and return to output

Answer 2、

图片包含 截图, 游戏机

描述已自动生成

In my own test case, I set alpha\_m = 9, beta\_n =100000, the test result show above.

Answer 3、

In order to improve the efficiency of my implementation, I create function check\_diff(), to check the difference of data\_hashes and query\_hashes only one time and save it into the output, in this way, everytime we add offset to get more data satisfy the query, we do not need to compare the data and query again, instead of comparing the data and difference.

And I try to aggregate the data\_hashes(rdd) with the same data to improve the efficiency, so that the same data in data\_hashes(rdd) will be count at the same time, however, the in this progress, the function will create new rdd and merge it together, which will decrease the efficiency.