**COMP9313 FINAL**

**Name: Yu ZHANG**

**Student ID: z5238743**

**Q1. HDFS**

**1.**

X = G’-1 \* P

G = P \* X-1

X =

P =

∴ G =

The code is Figure 1:

A screenshot of a cell phone

Description automatically generated

Figure 1

**2.**

The same maximum toleration as (6,3)-Reed-Solomon coding is 3, because when we lose 3 parities, we cannot recover the data from any 6 internal block.

**Q2. Spark and MapReduce**

**1.**

The code is shown below(figure 2)

**A screenshot of a cell phone

Description automatically generated**

Figure2

**2.**

This implementation is not correct, in the while loop, offset can not put in the last, it must put in the beginning.

Such like that

While cand\_num < beta\_n:

offset += 1

Candidates = data\_hashes.flatMap(lambda x:

[x[0] ] if collision\_count(x[1], query\_hashes, offset)>=alpha\_m else [])

cand\_num = candidates.count()

return candidates

Because of lazy evaluation, the data inside RDD is not available or transformed until an action is executed that triggers the execution.

In this codes, when in the last loop, because the action not trigger, the flatMap do not transformed. Thus, this implementation is not correct.

**Q3: LSH**

**1.**

Because Jaccard similarity between o and q is 0.8, R = 4, S =5

Thus pq,o =0.8

Because it is OALSH

First, using OR operation: Pr[HOR(O1)= HOR(O2)] = (1-pq,o)R

Second, using AND operation: Pr[HAND(O1)= HAND(O2)] = (1-(1-pq,o)R)S

Thus, the probability of o is a nearest neighbor candidate of q is (1-(1-pq,o)R)S

Thus, (1-(1-pq,o)R)S = (1-(1-0.8)4)5) = 0.9921

**2.**

Because threshold t = 0.5

So Jaccard similarity between o and q >= 0.5, R = 2, S =5

Thus pq,o >=0.5

From Q1, in OALSH, the probability of o is a nearest neighbor candidate of q is (1-(1-pq,o)R)S

Thus, (1-(1-pq,o)R)S = （1-(1-0.5)2)5) <= 0.2373

Because LSH schema with k = 5, in AOLSH, the probability of o is a nearest neighbor candidate of q is 1-(1-pkq,o)l

Thus, 1-(1-p5q,o)l = 1-(1-0.55)l > 0.2373

The minimal of l is 9

**3.**

Assume threshold t, Jaccard similarity between o and q equal to t, k,l,R,S for LSH and OALSH scheme,

Recall = number of returned positive/ total number of positive

Recall(OALSH) > Recall(LSH)

According to OALSH, the probability of o is a nearest neighbor candidate of q is (1-(1-pq,o)R)S and in AOLSH, the probability of o is a nearest neighbor candidate of q is 1-(1-pkq,o)l

Thus , (1-(1-t)R)S > 1-(1-tk)l

**Q4: Spark SQL**

The answer is shown above, from Figure 3(Figure 3)

**A screenshot of a cell phone

Description automatically generated**

Figure 3

**Q5: Stacking**

**1.**

Type A： 3 labels

5 group

1 metaclassifier

Type B: In the end, 3 base classifiers train the whole data respectively

Thus 3\*5+3+1= 19

**2.**

Because we have 3 base classifiers and the instance are trained in the 4 loops,So we have 3\*4 12 times, then In the end, 3 base classifiers train the whole data, so we have 3 times, finally，the metaclassifer train the data one times.

Thus, 3\*4 +3 +1 =16 times

**3.**

3 base classifiers train the training set 3 times, metaclassifier do not predict the training the data 1 time

Thus, 3 times

**Q6: Mining Data Streams**

**1.**

Because h1(str) = Pc∈str(c − ’a’) mod 8

|  |  |  |  |
| --- | --- | --- | --- |
| word | Hello | map | reduce |
| h1(str) | 7 | 3 | 2 |
| h2(str) | 5 | 3 | 6 |

The processes is shown in code(Figure 4):

A screenshot of a social media post

Description automatically generated

Figure 4

Sh1(str) and Sh2(str)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| h1(str) | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| h2(str) | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |

S:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| h1(str) | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |

**2.**

Spark is not contained in S.

Because，according to the code in Q4.1 h1(spark) = 4, h5(spark) = 5.

Although h2(spark) is in table h2(str), h1(spark) is not in h1(str)

They must both in S h1(str) and S h2(str).

Thus Spark is not contained in S.

**3.**

Because we have 2 independent hash function and the size of Bloom filter of size is 8,Besides,the number of elements in S is 5

Thus k =2, m=8, n= 5

According to False positive probability = (1-e-km/n)k

P = (1-e-2\*8/5)2 = 0.92013

**Q7: Recommender System**

**1.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | users | | | | |
| movies | 3 | 5 | ? | 0 | 2 |
| 0 | 4 | 0 | 1 | ? |
| 4 | ? | 5 | 2 | 0 |

According to baseline estimator = u + bx + bi

for movie 1: u = 3.25

bx = 5 - 3.25 = 1.75

bi = 3/10 – 3.25 =0.08333

Thus, baseline estimator = 3.25 +1.75 +0.08333=5.0833

for movie 2: u = 3.25

bx = 2 – 3.25 = -1.25

bi = 2.5– 3.25 = -0.75

Thus, baseline estimator = 3.25 -1.25 -0.75 = 1.25

for movie 3: u = 3.25

bx = 4.5 – 3.25 = 1.25

bi = 3.66667– 3.25 = 0.4166667

Thus, baseline estimator = 3.25 +1.25 + 0.75 = 4.9166667

**2.**

Rxi = qi \* PT

PT  =

for movie 1: Rxi = = 3.93

for movie 2: Rxi == 4.01

for movie 3: Rxi == 5

**3.**

For question 1’result:

RMSE = (1/|R|)\*((i,x) ∈R(Rxi^- Rxi))(1/2)

= (1/3)\*((5.0833-3)2 + (1.25-4)2 +(4.9166667-4))(1/2)

= 1.1896

For question 2’result:

RMSE = (1/|R|)\*((i,x) ∈R(Rxi^- Rxi))(1/2)

= (1/3)\*((3.93-3)2 + (4.01-4)2 +(5-4))(1/2)

= 0.4552

Because question 1’result > question 2’result:

Thus, using baseline estimator is better than using matrix factorization