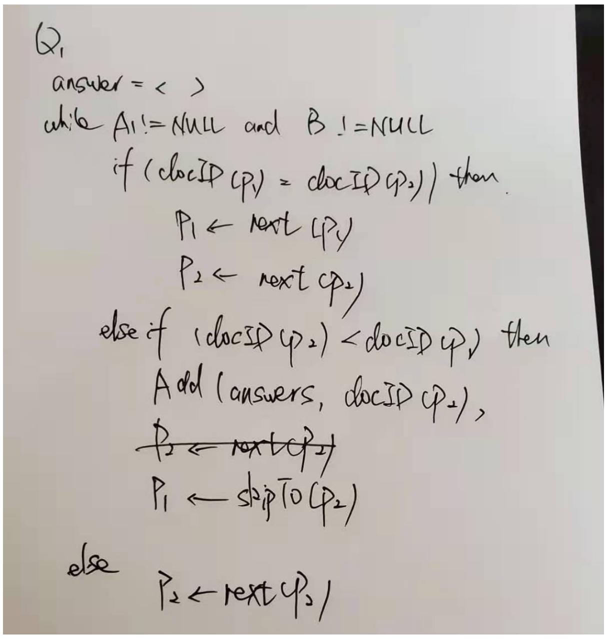
zID: z5224151

Name: Zanning Wang

Q1:

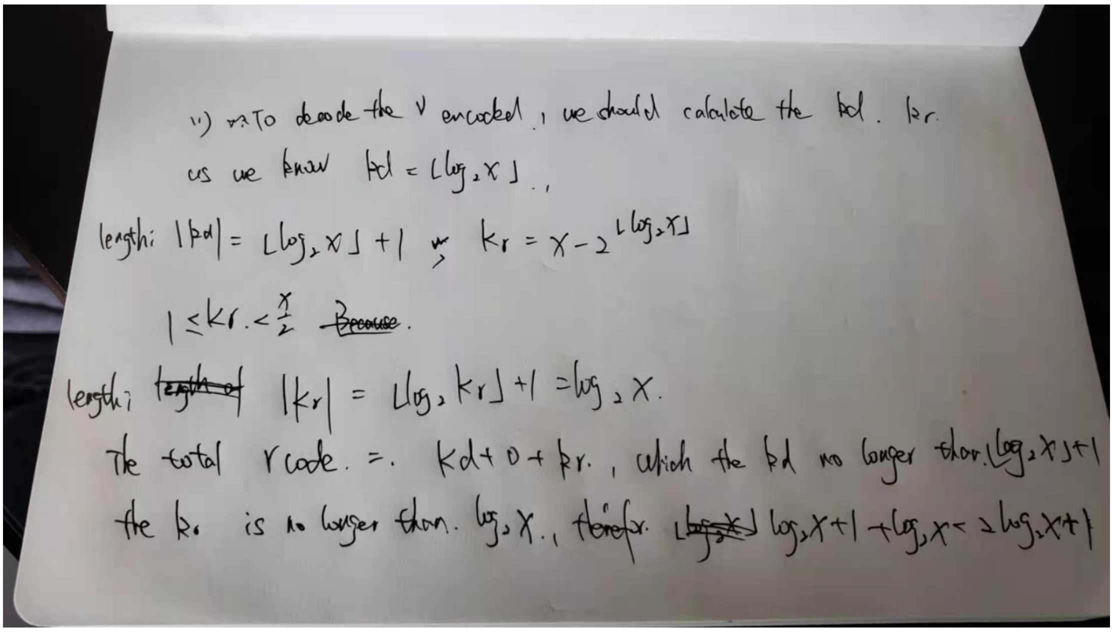


Q1

```
answer = < >
while A1 != NULL and B != NULL
      if (docID (p1) = docID (p2)) then
            P1 ← next (p1)
            P2 ← next (p2)
      else if (docID (p2) < docID (p1)) then
            Add (answers, docID (p2)),
            ~~P2 ← next (p2)~~
            P1 ← skipTo (p2)
   else
            P2 ← next (p2)
```

Q2:

(1)



1) To decode the v encoded, we should calculate the kd. kr.

as we know $kd = \lfloor \log_2 x \rfloor$,

length: $|kd| = \lfloor \log_2 x \rfloor + 1$; $k_r = x - 2^{\lfloor \log_2 x \rfloor}$

$1 \le kr < \frac{x}{2}$ ~~Because~~.

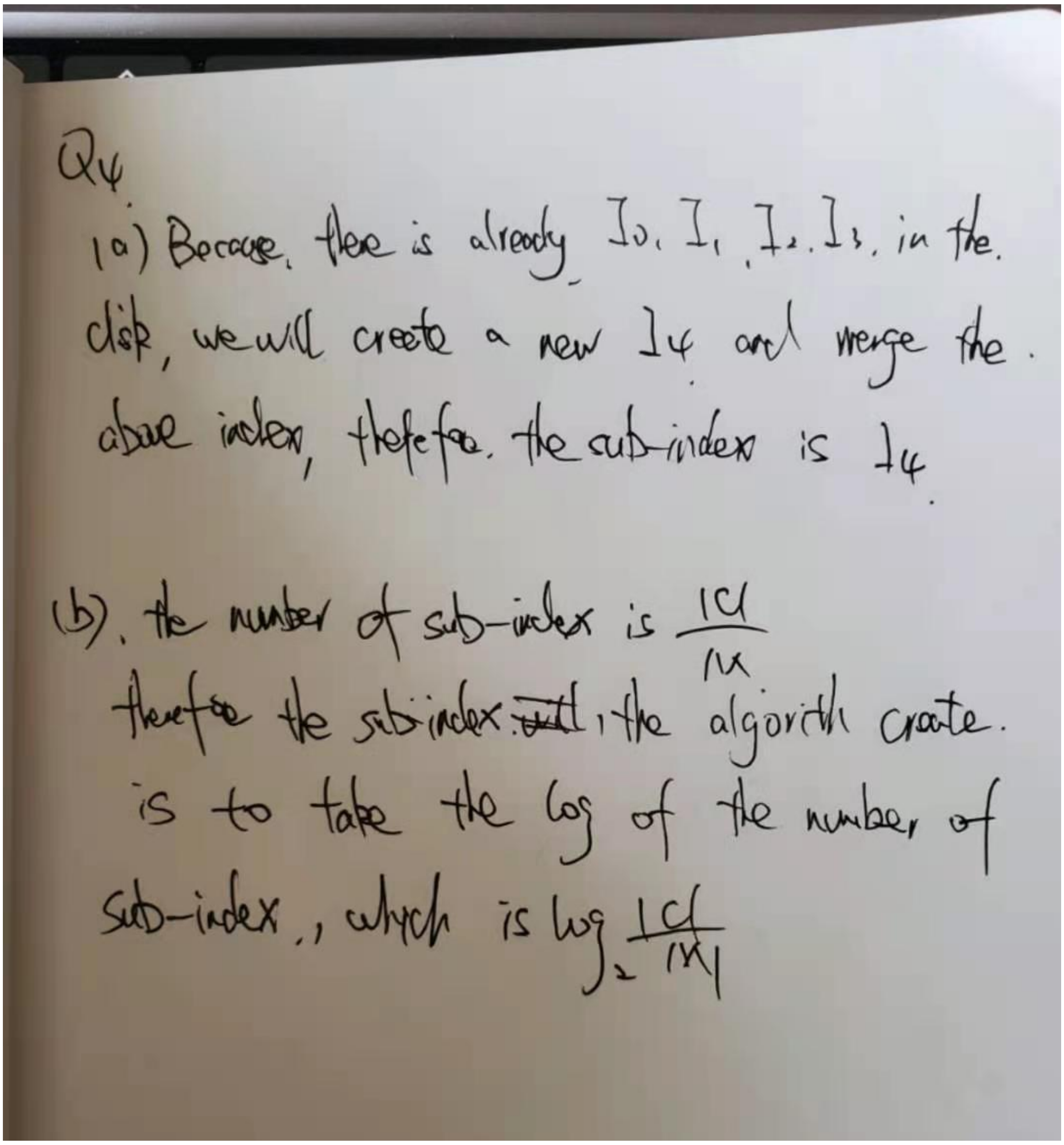
length: ~~length of~~ $|k_r| = \lfloor \log_2 k_r \rfloor + 1 = \log_2 x$.

The total rcode = kd + 0 + kr, which the kd no longer than $\lfloor \log_2 x \rfloor + 1$

the kr is no longer than $\log_2 x$, therefor ~~$\lfloor \log_2 x \rfloor$~~ $\log_2 x + 1 + \log_2 x = 2\log_2 x + 1$

Q3:

Q4:



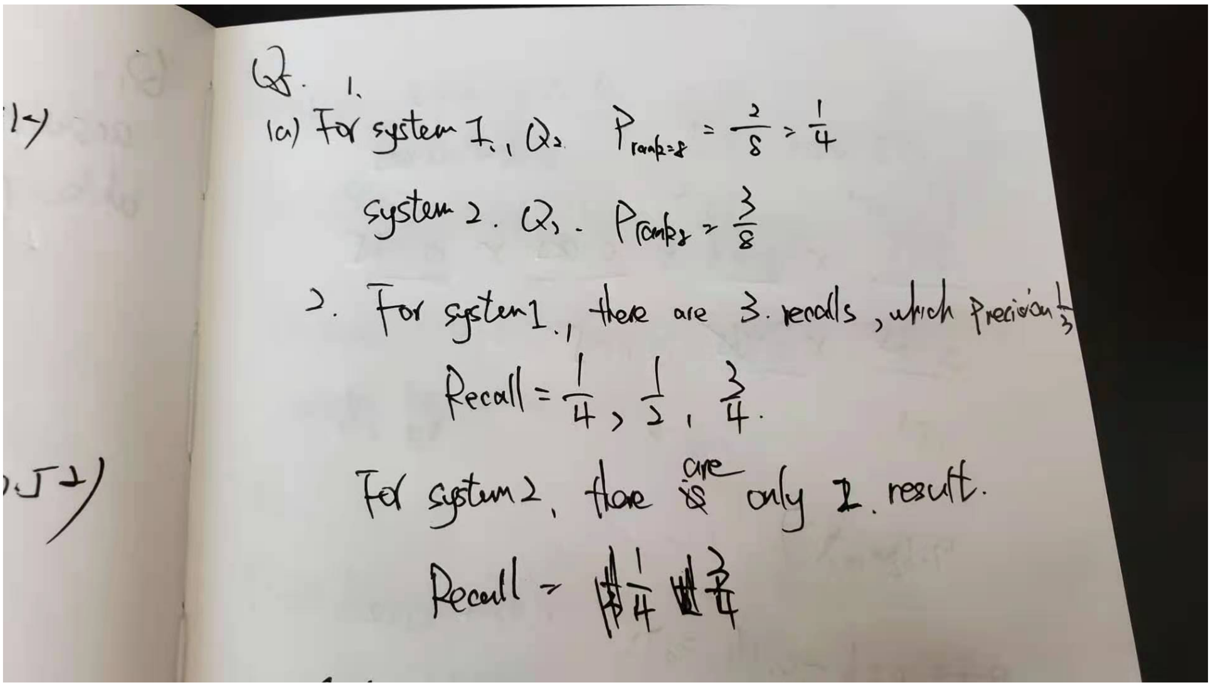
Q4.

1(a) Because there is already I_0, I_1, I_2, I_3, in the disk, we will create a new I_4 and merge the above index, therefore the sub-index is I_4.

(b). the number of sub-index is $\frac{|C|}{|M|}$

therefore the sub-index ~~will~~, the algorith create. is to take the log of the number of sub-index., which is $\log_2 \frac{|C|}{|M|}$

Q5:



Q8. 1.

(a) For system 1., Q2. $P_{rank=8} = \frac{2}{8} = \frac{1}{4}$

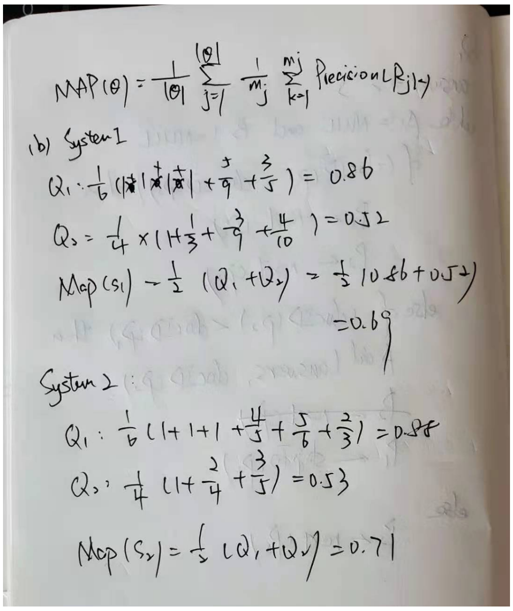
system 2. Q2. $P_{rank8} = \frac{3}{8}$

2. For system 1., there are 3. recalls, which Precision $\frac{1}{3}$

$Recall = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$.

For system 2, there ~~is~~ are only 2. result.

$Recall = \frac{1}{4}\ \frac{3}{4}$



$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

(b) System 1

$Q_1 : \frac{1}{6}(1+1+1 + \frac{5}{9} + \frac{3}{5}) = 0.86$

$Q_2 = \frac{1}{4} \times (1+\frac{1}{3} + \frac{3}{9} + \frac{4}{10}) = 0.52$

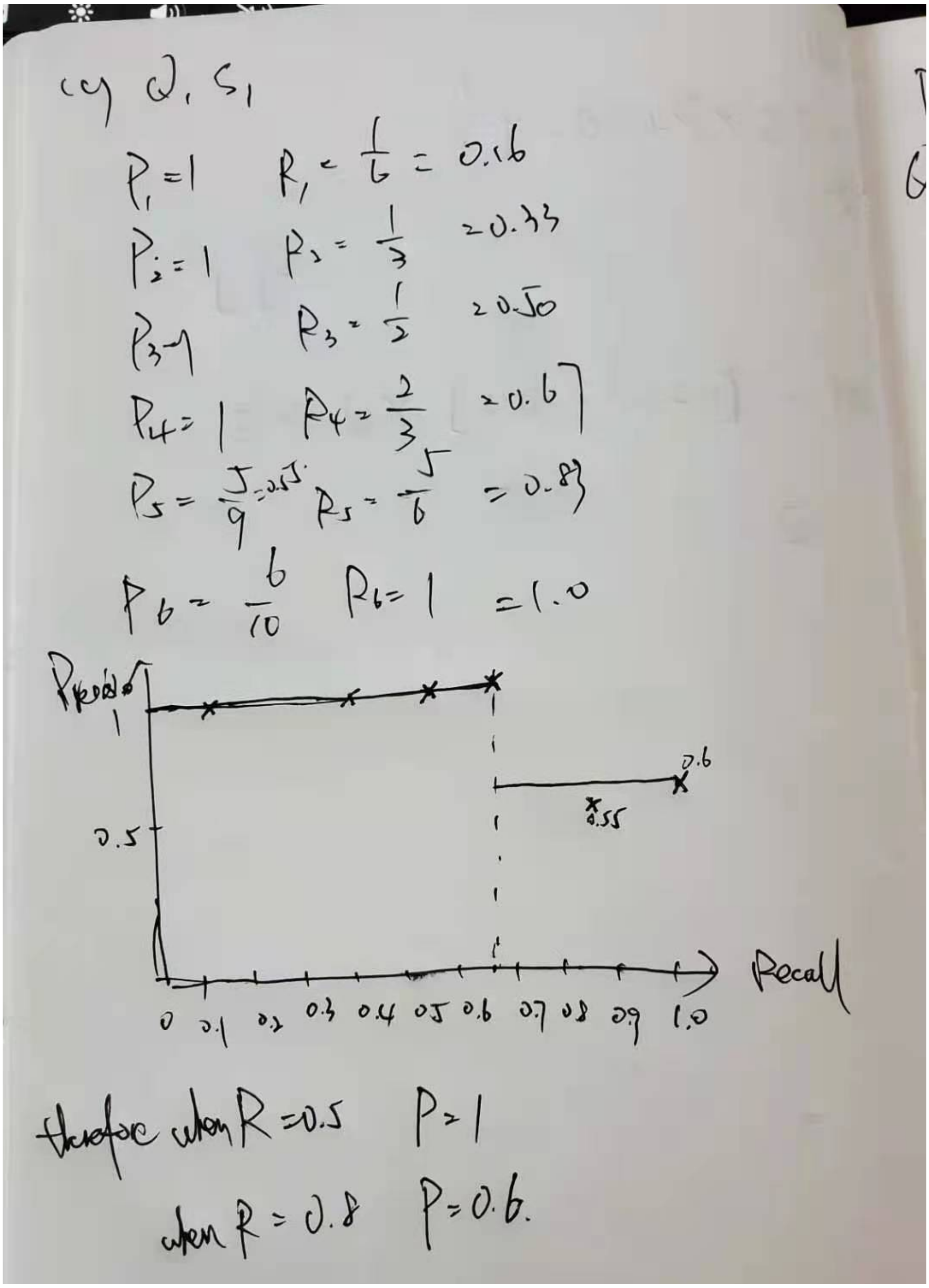
$Map(S_1) = \frac{1}{2}(Q_1+Q_2) = \frac{1}{2}(0.86+0.52)$
$= 0.69$

System 2:

$Q_1 : \frac{1}{6}(1+1+1 + \frac{4}{5} + \frac{5}{6} + \frac{2}{3}) = 0.88$

$Q_2 : \frac{1}{4}(1+\frac{2}{4} + \frac{3}{5}) = 0.53$

$Map(S_2) = \frac{1}{2}(Q_1+Q_2) = 0.71$



(c) Q_1, S_1

$P_1 = 1$ $R_1 = \frac{1}{6} = 0.16$

$P_2 = 1$ $R_2 = \frac{1}{3} = 0.33$

$P_3 = 1$ $R_3 = \frac{1}{2} = 0.50$

$P_4 = 1$ $R_4 = \frac{2}{3} = 0.67$

$P_5 = \frac{5}{9} = 0.55$ $R_5 = \frac{5}{6} = 0.83$

$P_6 = \frac{6}{10}$ $R_6 = 1 = 1.0$

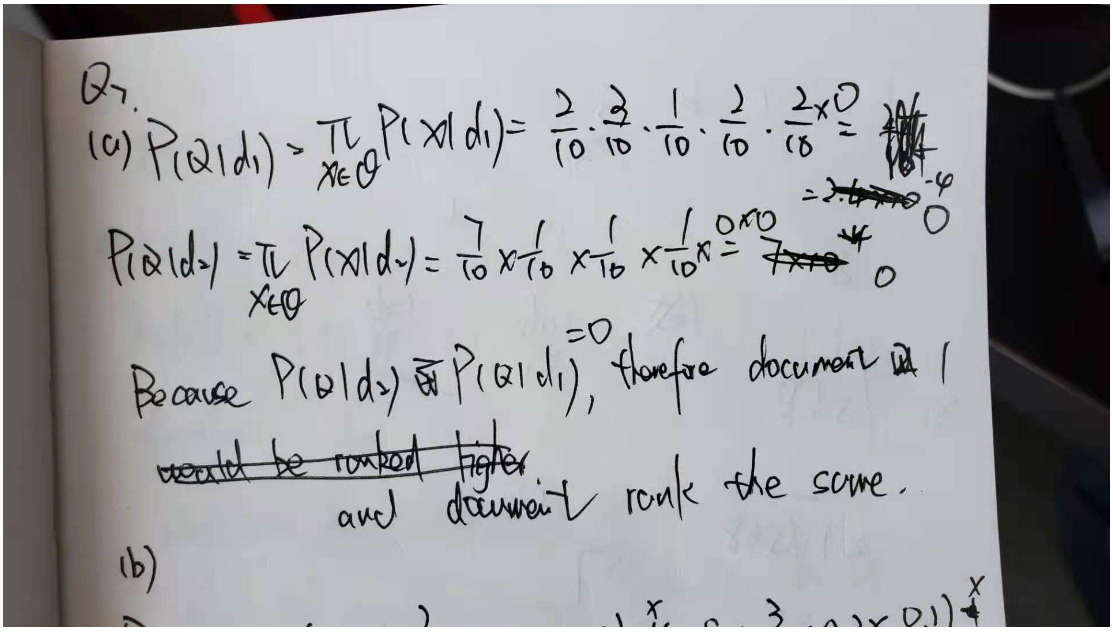
Precision
1
0.5
0.6
0.55
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

therefore when $R = 0.5$ $P = 1$

when $R = 0.8$ $P = 0.6$.

Q6:

Q7:



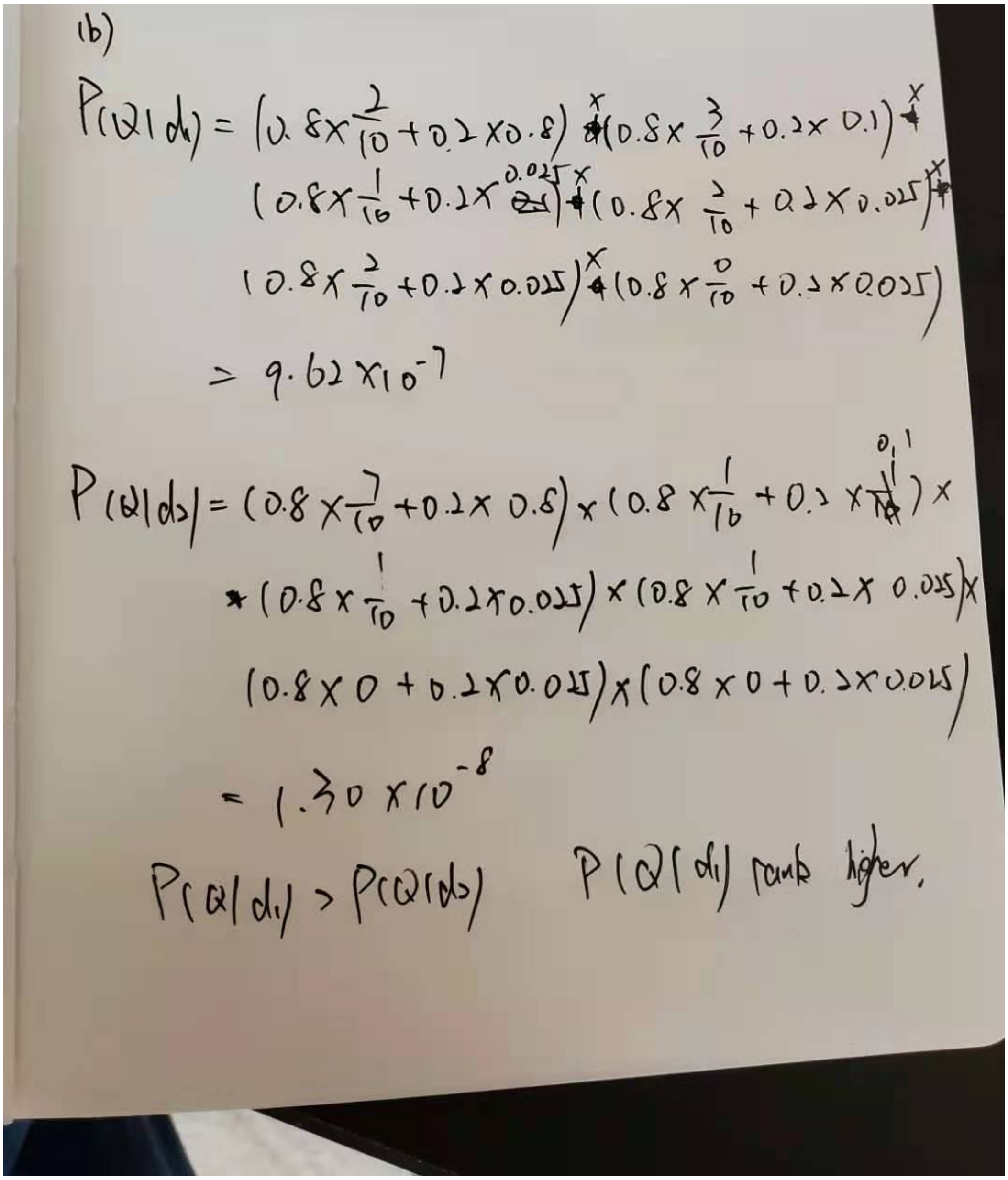
Q7.

(a) $P(Q|d_1) = \prod_{x\in Q} P(x|d_1) = \frac{2}{10}\cdot\frac{3}{10}\cdot\frac{1}{10}\cdot\frac{2}{10}\cdot\frac{2}{10}\times 0 =$ ~~[illegible]~~ $= \cancel{2.4\times10^{-4}}\ 0$

$P(Q|d_2) = \prod_{x\in Q} P(x|d_2) = \frac{7}{10}\times\frac{1}{10}\times\frac{1}{10}\times\frac{1}{10}\times 0\times 0 = \cancel{7\times10^{-4}}\ 0$

Because $P(Q|d_2) = 0\ P(Q|d_1)$, therefore document 1 ~~would be ranked higher~~ and document 2 rank the same.

(b)



(b)

$P(Q|d_1) = (0.8\times\frac{2}{10}+0.2\times0.8)\times(0.8\times\frac{3}{10}+0.2\times0.1)\times$
$(0.8\times\frac{1}{10}+0.2\times\overset{0.025}{\cancel{0.1}})\times(0.8\times\frac{2}{10}+0.2\times0.025)\times$
$(0.8\times\frac{2}{10}+0.2\times0.025)\times(0.8\times\frac{0}{10}+0.2\times0.025)$
$= 9.62\times10^{-7}$

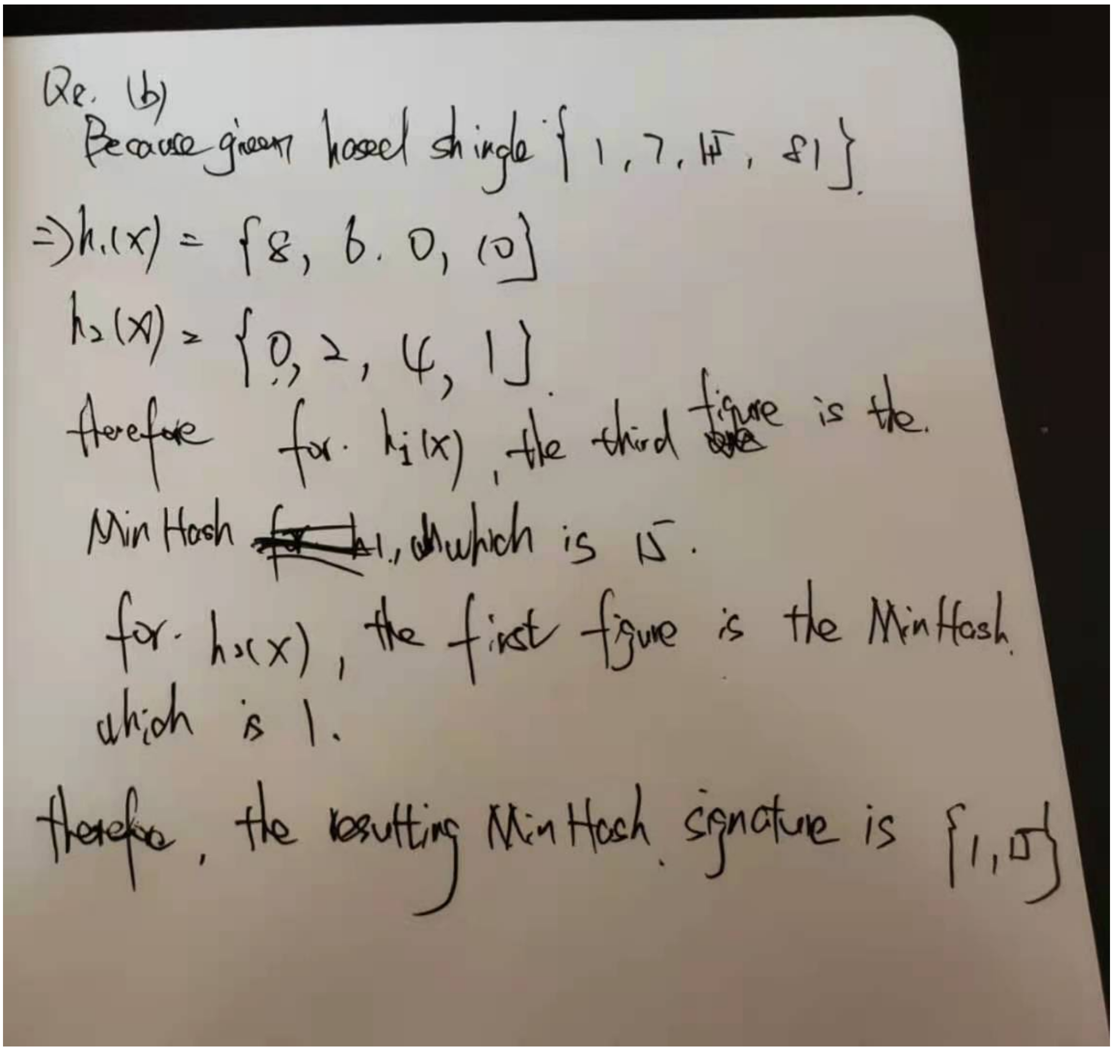
$P(Q|d_2) = (0.8\times\frac{7}{10}+0.2\times0.8)\times(0.8\times\frac{1}{10}+0.2\times\overset{0.1}{\cancel{\frac{1}{10}}})\times$
$(0.8\times\frac{1}{10}+0.2\times0.025)\times(0.8\times\frac{1}{10}+0.2\times0.025)\times$
$(0.8\times0+0.2\times0.025)\times(0.8\times0+0.2\times0.025)$
$= 1.30\times10^{-8}$

$P(Q|d_1) > P(Q|d_2)$ $P(Q|d_1)$ rank higher.

Q8:

(a)

In the content seen module, we will check if the current URL has been crawl, in this process, we will store as an index and it will also pass to URL filter to check the URL valid or not, after that the “Dup url elim can check what URL is duplication or not



Q8. (b)

Because given hased shingle {1, 7, 15, 81}.

$\Rightarrow h_1(x) = \{8, 6, 0, 10\}$

$h_2(x) = \{0, 2, 4, 1\}$.

therefore for $h_1(x)$, the third figure ~~one~~ is the MinHash ~~for h1~~, which is 15.

for $h_2(x)$, the first figure is the MinHash which is 1.

therefore, the resulting MinHash signature is {1, 15}