Student Name : WANG ZHUORAN

Student ID : 3521225

Question 1 :

I draw,

this is not A and B

Algorithm 1 :

```
answer ← ∅ ;
B ← P1 ;
A ← P2 ;
while P1 ≠ nil and P2 ≠ nil do
    if docID(P1) = docID(P2) then
        P1 ← skipto(P1);
        P2 ← skipto(P2);
    else if docID(P1) < docID(P2) then
        Add(answer, docID(P1));
        P1 ← skipto(P1);
    else
        P2 ← skipto(P2);
if P1 ≠ nil then
    while P1 ≠ nil do
        Add(answer, docID(P1));
        P1 ← skipto(P1)
return answer;
```

Explain: For not A and B, when ~~B ← A~~, docID(B) < docID(A) means this docID(B) will not show in A, also, when A is nil, B is not nil, also the rest of docID(B) should add to answer.

Question 2

For γ-encode: if encode a number k,

the unary is $k_d = \lfloor \log_2 k \rfloor$

the binary is $k_r = k - 2^{\lfloor \log_2 k \rfloor}$

so, $k = 2^{k_d} + k_r$

Question 3.

(a)

Question 4:

(a) ~~Before~~ Original: Disk: I_3, I_2, I_1, I_0

after dumping the current in-memory index to the disk.

Disk: I_3, I_2, I_1, I_0, I_0, and there cannot be two same sub-index.

so. $I_0, I_0 \rightarrow I_1$, Disk: I_3, I_2, I_1, I_1

$I_1, I_1 \rightarrow I_2$, Disk: I_3, I_2, I_2

$I_2, I_2 \rightarrow I_3$, Disk: I_3, I_3

$I_3, I_3 \rightarrow I_4$, Disk: I_4.

So, the sub-index is I_4.

(b) I draw a table

time	1	2	3	4	5	6	2^3	2^k
disk	I_0	$\{I_0, I_0\}$ $= I_1$	~~$\{I_0, I_0, I_0\}$~~ ~~$= \{I_1, I_0\}$~~ $\{I_0, I_1\}$	$\{I_0, I_1, I_0\}$ $= \{I_2\}$	$\{I_2, I_0\}$	$\{I_2, I_0, I_0\}$ $= \{I_2, I_1\}$	$\{I_2, I_1, I_0, I_0\}$ $= \{I_3\}$	$\{I_k\}$
size	M	2M	3M	4M	5M	6M	8M	2^kM

So, it will create k sub-index.

Question 5

(a) 1. Precision at rank 8

For system 1, when rank 8, there are 2 relevant documents for Q2

So, Q2 : $\frac{2}{8} = 0.25$

For system 2, when rank 8, there are 3 relevant documents for Q2, so, Q2 : $\frac{3}{8} = 0.375$

So, System 1 Q2 : 0.25, System 2 Q2 : 0.375

2. Recall at precision $\frac{1}{3}$.

For system 1, at rank 3, there is one relevant document, so, the precison is $\frac{1}{3}$.

also, at rank 6 and 9, the precison is $\frac{1}{3}$.

So, Q2 : $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ at rank 3, 6, 9.

For system 2, at rank 3 and 9. the precision is $\frac{1}{3}$.

So. Q2 : $\frac{1}{3}$, 1 at rank 3, 9.

So. system 1 : Q2 : $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ at rank 3, 6, 9

system 2 : Q2 : $\frac{1}{3}$, 1 at rank 3, 9

(b): MAP formula:

$$MAP = \frac{1}{|Q|} \sum_{|Q|} \frac{\text{number of Relevant documents}}{RANK}$$

For system1:

Q1: $MAP_{Q1} = \frac{1}{6}(1+1+1+\frac{5}{9}+\frac{6}{10}) = \frac{118}{135}$

Q2: $MAP_{Q2} = \frac{1}{4}(1+\frac{2}{6}+\frac{3}{9}+\frac{4}{10}) = \frac{31}{15}$

$MAP_1 = \frac{1}{2}(MAP_{Q1} + MAP_{Q2}) = \frac{397}{270}$

For system2:

Q1: $MAP_{Q1} = \frac{1}{6}(1+1+1+\frac{4}{5}+\frac{5}{6}+\frac{6}{9}) = \frac{53}{60}$

Q2: $MAP_{Q2} = \frac{1}{2}(1+\frac{2}{4}+\frac{3}{5}) = \frac{7}{10}$

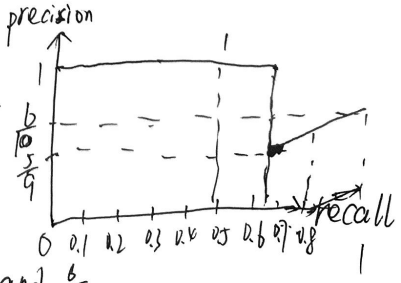
$MAP_2 = \frac{1}{2}(MAP_{Q1} + MAP_{Q2}) = \frac{19}{24}$

(c) calculat recall, precision.

recall: $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}$

precision: $1, 1, 1, 1, \frac{5}{9}, \frac{6}{10}$

so # 2 draw a graph



As we can see from graph,
when recall = 0.5, precision = 1.
when recall = 0.8, precision is between $\frac{5}{9}$ and $\frac{6}{10}$

Question 7

(a) $P(Q|d_1) = \frac{2}{10} \times \frac{3}{10} \times \frac{1}{10} \times \frac{2}{10} \times \frac{2}{10} \times 0 = 0$

$P(Q|d_2) = \frac{7}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times 0 \times 0 = 0$

$P(Q|d_1) = P(Q|d_2)$

So, d_1, d_2 rank same

(b) For d_1:

$P(w_1|d_1) = 0.8 \times \frac{2}{10} + (1-0.8) \times 0.8 = 0.32$

So,

$P(w_2|d_1) = 0.8 \times \frac{3}{10} + 0.2 \times 0.1 = 0.26$

$P(w_3|d_1) = 0.8 \times \frac{1}{10} + 0.2 \times 0.025 = 0.085$

$P(w_4|d_1) = 0.8 \times \frac{2}{10} + 0.2 \times 0.025 = 0.165$

$P(w_5|d_1) = 0.8 \times \frac{2}{10} + 0.2 \times 0.025 = 0.165$

$P(w_6|d_1) = 0.8 \times \frac{0}{10} + 0.2 \times 0.025 = 0.005$

So, $P(Q|d_1) = P(w_1|d_1) \times P(w_2|d_1) \times P(w_3|d_1) \times P(w_4|d_1) \times P(w_5|d_1) \times P(w_6|d_1) = 0.32 \times 0.26 \times 0.085 \times 0.165 \times 0.165 \times 0.005 = 9.6 \times 10^{-7}$

using Jelinek-Mercer smoothing method with $\lambda = 0.8$

$P(w|d) = \lambda \cdot P_{mle}(w|M_d) + (1-\lambda) \cdot P_{mle}(w|M_c)$

For d_2:

$P(w_1|d_2) = 0.8 \times \frac{7}{10} + 0.2 \times 0.8 = 0.72$

$P(w_2|d_2) = 0.8 \times \frac{1}{10} + 0.2 \times 0.1 = 0.1$

$P(w_3|d_2) = 0.8 \times \frac{1}{10} + 0.2 \times 0.025 = 0.085$

$P(w_4|d_2) = 0.8 \times \frac{1}{10} + 0.2 \times 0.025 = 0.085$

~~$P(w_5|d_2) = 0.8 \times \frac{1}{10} + 0.2 \times 0.025 = 0.085$~~

$P(w_5|d_2) = 0.8 \times \frac{0}{10} + 0.2 \times 0.025 = 0.005$

$P(w_6|d_2) = 0.8 \times \frac{0}{10} + 0.2 \times 0.025 = 0.005$

So, $P(Q|d_2) = P(w_1|d_2) \times P(w_2|d_2) \times P(w_3|d_2) \times P(w_4|d_2) \times P(w_5|d_2) \times P(w_6|d_2)$

$= 0.72 \times 0.1 \times 0.085 \times 0.085 \times 0.005 \times 0.005 = 1.3 \times 10^{-8}$

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

Question 8:

(a) Explain: if the page just fetched is already in the index, it is not necessary to further process it.

(b) hash shingles = {1, 7, 15, 81}

For $h_1(x) = (7x+1 \bmod 31) \bmod 13$.

when $x=1$, $h_1(x) = (7\times1+1 \bmod 31) \bmod 13 = 6$

when $x=7$, $h_1(x) = (7\times7+1 \bmod 31) \bmod 13 = 6$

when $x=15$, $h_1(x) = (15\times7+1 \bmod 31) \bmod 13 = 0$

when $x=81$, $h_1(x) = (81\times7+1 \bmod 31) \bmod 13 = 10$

For $h_2(x) = (18x+26 \bmod 31) \bmod 13$

when $x=1$, $h_2(x) = (18\times1+26 \bmod 31) \bmod 13 = 0$

$x=7$, $h_2(x) = (18\times7+26 \bmod 31) \bmod 13 = 2$

$x=15$, $h_2(x) = (18\times15+26 \bmod 31) \bmod 13 = 11$

$x=81$, $h_2(x) = (18\times81+26 \bmod 31) \bmod 13 = 1$

Choose $h=0$. So.

when $x=15$, $h_1(x) = 0$

when $x=1$, $h_2(x) = 0$

So, the signatures is 1, 15.