Q 1.

answer ← ϕ;

while $p_2 \neq$ nil do

If docID(p_1) > docID(p_2) then

~~Add(docID(p_2)~~,

Add(answer, docID(p_2));

$P_2 \leftarrow$ next(P_2);

else:

If docID(p_1) < docID(P_2) then

$P_1 \leftarrow$ skipTo(P_2);

else

$P_2 \leftarrow$ next(P_2)

$P_1 \leftarrow$ skipTo(P_2)

return answer

Q2.

(1). Prove:

To encode a number x.

we need to compute:

$k_d = \lfloor \log_2 x \rfloor$ unary

$k_r = x - 2^{\lfloor \log_2 x \rfloor}$ binary

∴ The above at most take $2\log_2(x) + 1$ bits.

(2).

Q3. (a) Permuterm index $V \cdot L$

bi-gram index: $V \cdot (L+2)$

(b). Use DAAT

Q4. (a).

Assume we merge from I_0

$I_0 \rightarrow I_0$

$I_0, I_1 > I_0$

$I_0, I_1 > I_0$

$I_2 \rightarrow I_2$

$I_0, I_1 > I_0$; $I_2, I_3 > I_0$; $> I_1$

- After dumping the current in-memory index to the disk, the sub-index is I_0, I_1

(b). ~~Index construction time~~

~~The~~ Each posting is merged $O(\log c/m)$

(c) Index construction time is ~~$O(c \log c)$~~ $O(c^2/M)$

Q5

(a). 1. System 1:

Precision (Q_1) = $\frac{2}{8} = \frac{1}{4}$

System 2:

Precision (Q_2) = $\frac{3}{8}$

2. System 1:

When precision = $\frac{1}{3}$, then, ~~Rank~~

Rank at	3	6	9
Recall	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$

System 2:

when precision = $\frac{1}{3}$, then

Rank at	3	9
Recall	$\frac{1}{3}$	1

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

System 1:

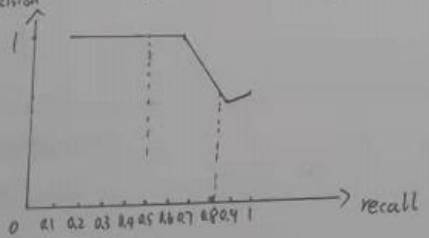
~~MAP = $(1+1+1+1+\frac{5}{9}+\frac{3}{5})/6$~~.

$AP(Q_1) = (1+1+1+1+\frac{5}{9}+\frac{3}{5}) \div 6 = \frac{116}{135} = 0.859$

$AP(Q_2) = (1+\frac{1}{3}+\frac{1}{3}+\frac{2}{5}) \div 4 = \frac{31}{60} = 0.517$

$MAP = \frac{AP(Q_1) + AP(Q_2)}{2} = \frac{0.859+0.517}{2} = 0.688$

(c).



~~The interpolated~~

The 11-point interpolated ~~curve~~ precision-recall is like above.

Then, we can find answer from above graph

Recall	0.5	0.8
precision	1	$\frac{3}{5}$

Q6 $P_1 = P(x_1=1|R,Q) = \frac{1+\frac{1}{2}}{3+1} = \frac{3}{8}$

$r_1 = P(x_1=1|NR,Q) = \frac{H\frac{1}{2}}{2+1} = \frac{1}{2}$

$P_3 = P(x_3=1|R,Q) = \frac{2+\frac{1}{2}}{3+1} = \frac{5}{8}$

$r_3 = P(x_3=1|NR,Q) = \frac{\frac{1}{2}}{2+1} = \frac{1}{6}$

$RSV_1 = \log_{it}(P_1) - \log_{it}(r_1) = \log \frac{P_1(1-r_1)}{r_1(1-P_1)} = \log \frac{3}{5}$

$RSV_3 = \log_{it}(P_3) - \log_{it}(r_3) = \log \frac{P_3(1-r_3)}{r_3(1-P_3)} = \frac{\frac{5}{8}\cdot\frac{5}{6}}{\frac{1}{6}\cdot\frac{3}{8}} = \log \frac{25}{3}$

$RSV_3 > RSV_1$ ∴ order should be x_3, x_1

Q7. (a) $\because P(Q|M_D) = \prod_w P(w|M_D)^{tf_w}$

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

$P(Q|d_2) = \frac{2}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{0}{10} \times \frac{0}{10} = 0$

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

$\therefore$ d_1 and d_2 ranked same.

(b). $P(Q|d_1) = (0.8 \times 0.2 + 0.2 \times 0.8) \times (0.8 \times 0.3 + 0.2 \times 0.1) \times (0.8 \times 0.1 + 0.2 \times 0.025) \times (0.8 \times 0.2 + 0.2 \times 0.025) \times (0.8 \times 0.2 + 0.2 \times 0.025) \times (0.8 \times 0.0 + 0.2 \times 0.025) = 0.32 \times 0.26 \times 0.085 \times 0.165 \times 0.165 \times 0.005$

$= 9.6 \times 10^{-7}$

$P(Q|d_2) = (0.8 \times 0.7 + 0.2 \times 0.8) \times (0.8 \times 0.1 + 0.2 \times 0.1) \times (0.8 \times 0.1 + 0.2 \times 0.025) \times (0.8 \times 0.1 + 0.2 \times 0.025) \times (0.8 \times 0 + 0.2 \times 0.025) \times (0.8 \times 0 + 0.2 \times 0.025) = 0.72 \times 0.1 \times 0.085 \times 0.085 \times 0.005 \times 0.005$

$= 1.3 \times 10^{-8}$

$\therefore P(Q|d_1) > P(Q|d_2)$

$\therefore$ d_1 is ranked higher

Q8.

(a). Duplication is widespread on the web

If the page just fetched is already in the index, do not further process it

This is verified using document fingerprints or shingles

(b). for $h_1(x)$

$h(e) = \{8, 6, 0, 10\}$

for $h_2(x)$

$h(e) = \{0, 2, 4, 9\}$