Name: XIE, MINGLANG
Student ID: z5228006

Name: MINGLANG XIE ZID: z5228006

Q1: let A be P_1, B be P_2

Algorithm 1: Q1(P_1, P_2)

answer ← ∅;
while P_1 != NULL and P_2 != NULL do
~~if docID(P_1) > docID(P_2) then~~
if docID(P_1) == docID(P_2) then
~~P_1 ← skipTo(P_2)~~
~~P_2~~
P_2 ← skipTo(docID(P_2));
P_1 ← skipTo(docID(P_2));
else if docID(P_1) > docID(P_2) then
~~Add conv~~
Add(answer, docID(P_2));
P_2 ← ~~nex~~ skipTo(docID(P_2));
else
P_1 ← skipTo(docID(P_1));
end if
end while
// add the remaing
if P_2 != NULL then
while P_2 != NULL do
Add(~~co~~answer, docID(P_2));
P_2 ← next(P_2);
end while
return answer

Q2:

For applying γ-encoding compute

$$\cancel{k_d = \lfloor \log_2 k \rfloor}$$

$$k_d = \lfloor \log_2 k \rfloor \quad \text{unary}$$

$$k_r = k - 2^{\lfloor \log_2 k \rfloor} \quad \text{binary}$$

$$k_r = k - 2^{k_d}$$

Prove

- For a value x, its γ-encoded value takes at most $2\log_2(x)+1$ bits.

  For unary part, ~~the~~ γencoded takes

  $k_d = \lfloor \log_2 k \rfloor$ ~~for~~ bits.

  For binary part, γ-encoded takes

  $\cancel{\log_2(k_r = k - 2^{\lfloor \log_2 k \rfloor})}$ ~~pack~~ bits

  $\log_2(k_r)$

  Since k_r is a ~~d~~ integer, and we need to convert it to binary for ~~end~~ encoding. ~~Th.~~

  $\therefore$ $k_d + \log_2(k_r)$ is at most $2\log_2(x)$, ~~plus a zero (log~~

  ~~So~~ plus a zero between unary and binary, which takes 1 bit

  γ-encoded takes at most $2\log_2(x)+1$ bits.

- The compressed posting list (using γ codes on the gaps) takes at most $n \cdot \log_2(\frac{2N^2}{n^2})$ bits

Q3:

(b) P*Q*R

R$PQ* is the query ~~for~~ build by permuterm index

(c)
for Permuterm query processing is rotate query wild-card to the righ, so that *'s occur at the end. However, Bigram indexes ~~is~~ enumerate all k-grams (sequence of k chars) occuring in any term, and finds terms based on a query consisting of k-grams.

Q4:

(a) the sub-indexes after dumping the current in-memory index to the disk is I_4.

(b) The size is $\lceil \log_2 \frac{|C|}{M} \rceil$

(c) $|C| = 14 \cdot M$,

there will be 3 sub-indexes: 2M, 4M, 8M

which mean the total times are merged is 11.

Q5:

(a)

1. $\text{Precision} = \dfrac{\#\text{ relevant doc in result}}{\#\text{ total retrieved doc in } \cancel{\text{system}} \text{ query}}$

query Q 2 for System 1, precision at rank 8 $= \frac{2}{8} = \frac{1}{4}$

for System 2, Precision at rank 8 $= \frac{3}{8}$

2. $\text{Recall} = \dfrac{\#\text{ relevant doc in result}}{\#\text{ total relevant docs in system}}$

answer's ~~format~~ form

(Rank#, Recall)

~~System 1: (Rank 3, Rank 6, Rank 9)~~ Sys 1: $(3, \frac{1}{4}), (6, \frac{1}{2}), (9, \frac{3}{4})$

~~System 2: (Rank 3, Rank 9)~~ Sys 2: $(3, \frac{1}{3}), (9, 1)$

~~System 1: (no result recall)~~

~~System 2: (Rank 1, Rank 2, Rank 3)~~

~~(Rank 1,~~

(b)

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

Sys 1 $\begin{cases} Q1: MAP_1 = \frac{1}{6}\left(\frac{1}{1}+\frac{2}{2}+\frac{3}{3}+\frac{4}{4}+\frac{5}{9}+\frac{6}{10}\right) \\ Q2: MAP_2 = \frac{1}{4}\left(\frac{1}{1}+\frac{2}{6}+\frac{3}{9}+\frac{4}{10}\right) \end{cases}$

Sys 1: $MAP = \frac{1}{2}(MAP_1 + MAP_2) = \frac{1}{2} \times 1.376 = 0.688$

Sys 2 $\begin{cases} Q1: MAP_3 = \frac{1}{6}\left(\frac{1}{1}+\frac{2}{2}+\frac{3}{3}+\frac{4}{5}+\frac{5}{8}+\frac{6}{9}\right) \\ Q2: MAP_4 = \frac{1}{3}\left(\frac{1}{1}+\frac{2}{4}+\frac{3}{5}\right) \end{cases}$

Sys 2: $MAP = \frac{1}{2}(MAP_3 + MAP_4) = \frac{1}{2} \cdot 1.583 = 0.792$

(c) For Q1 system 1:

precision	recall	interpolated precision
1	~~1/6~~	1
1	2/6	1
1	3/6	1
1	4/6	1
4/5	4/6	4/5
4/6	4/6	4/6
4/7	4/6	6/10
4/8	5/6	6/10
5/9		6/10
6/10	6/6	6/10

∴

Recall	0.5	0.8
Ans	1	0.6

Q6:

X_1:

Doc	Relavant	Non-Relevant	Total
$X_1=1$	$1+\frac{1}{2}=\frac{3}{2}$	$1+\frac{1}{2}=\frac{3}{2}$	$2+1=3$
$X_1=0$	$2+\frac{1}{2}=\frac{5}{2}$	$1+\frac{1}{2}=\frac{3}{2}$	$3+1=4$
Total	$3+1=4$	$2+1=3$	$5+2=7$

X_3:

Doc	Relavant	Non-Relevant	Total
$X_3=1$	$2+\frac{1}{2}=\frac{5}{2}$	$0+\frac{1}{2}=\frac{1}{2}$	$2+1=3$
$X_3=0$	$1+\frac{1}{2}=\frac{3}{2}$	$2+\frac{1}{2}=\frac{5}{2}$	$3+1=4$
Total	$3+1=4$	$2+1=3$	$5+2=7$

$p_i \approx \frac{s}{S} \quad r_i \approx \frac{(n-s)}{(N-S)}$ $\qquad c_i \approx K(N,n,S,s) = \log \frac{s/(S-s)}{(n-s)/(N-n-S-s)}$

$p_1 = \frac{\frac{3}{2}}{4} = \frac{3}{8}$, $r_1 = \frac{\frac{3}{2}}{3} = \frac{1}{2}$ $\qquad c_1 = \log \frac{\frac{3}{2}/\frac{5}{2}}{\frac{3}{2}/\frac{3}{2}} = \log 0.6 = \log\frac{3}{5}$

$p_3 = \frac{\frac{5}{2}}{4} = \frac{5}{8}$, $r_3 = \frac{\frac{1}{2}}{3} = \frac{1}{6}$ $\qquad c_3 = \log \frac{\frac{5}{2}/\frac{3}{2}}{\frac{1}{2}/\frac{5}{2}} = \log\frac{25}{3}$

~~the order is~~

D_1, $RSV = c_1 + c_3 = \log\frac{3}{5} + \log\frac{25}{3} = 0.69897$

D_2, $RSV = 0$

D_3, $RSV = c_1 = \log\frac{3}{5} = -0.22185$

D_4, $RSV = c_3 = \log\frac{25}{3} = 0.92082$

D_5, $RSV = 0$

$\therefore$ The order is D_4, D_1, D_3, D_2, D_5.

Q7:

(a)

$$P(Q|d_1) = \prod_{x \in Q} P(x|d_1) = \frac{2}{10} \cdot \frac{3}{10} \cdot \frac{1}{10} \cdot \frac{2}{10} \cdot \frac{2}{10} \cdot \frac{0}{10} = 0$$

$$P(Q|d_2) = \prod_{x \in Q} P(x|d_2) = \frac{7}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{0}{10} \cdot \frac{0}{10} = 0$$

∴ They are ~~equal~~ equal

(b)

$$P(Q|d_1) = (0.8 \cdot \tfrac{2}{10} + 0.2 \cdot 0.8) \cdot (0.8 \cdot \tfrac{3}{10} + 0.2 \cdot 0.1) \cdot (0.8 \cdot \tfrac{1}{10} + 0.2 \cdot 0.025) \cdot (0.8 \cdot \tfrac{2}{10} + 0.2 \cdot 0.025)$$
$$\cdot (0.8 \cdot \tfrac{2}{10} + 0.2 \cdot 0.025) \cdot (0.8 \cdot \tfrac{0}{10} + 0.2 \cdot 0.025)$$

$$= 9.62676 \times 10^{-7}$$

$$P(Q|d_2) = (0.8 \cdot \tfrac{7}{10} + 0.2 \cdot 0.8) \cdot (0.8 \cdot \tfrac{1}{10} + 0.2 \cdot 0.1) \cdot (0.8 \cdot \tfrac{1}{10} + 0.2 \cdot 0.025)$$
$$\cdot (0.8 \cdot \tfrac{1}{10} + 0.2 \cdot 0.025) \cdot (0.8 \cdot \tfrac{0}{10} + 0.2 \cdot 0.025) \cdot (0.8 \cdot \tfrac{0}{10} + 0.2 \cdot 0.025)$$

$$= 1.3005 \times 10^{-8}$$

∴ Document 1 is ranked higher

Q8:

(a) · Duplication is widespread on the web

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

(b)

hashed shingles: $\{1, 7, 15, 81\}$

$h_1(x) = \{(7+1 \bmod 31) \bmod 13, (49+1 \bmod 31) \bmod 13, (105+1 \bmod 31) \bmod 13, (567+1 \bmod 31) \bmod 13\}$

$= \{8, 6, \underline{\underline{0}}, 10\}$

$h_2(x) = \{(18+26 \bmod 31) \bmod 13, (126+26 \bmod 31) \bmod 13, (270+26 \bmod 31) \bmod 13, (1458+26 \bmod 31) \bmod 13\}$

$= \{\underline{\underline{0}}, 2, 4, 1\}$