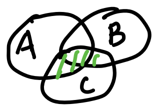
Q1.

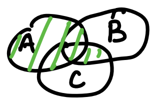
According to the question, the result set of "A OR B AND C" is
((A OR B) AND C) ∪ (A OR (B AND C)).

To help visualise,

(A OR B) AND C :



A OR (B AND C) :



Therefore ((A OR B) AND C) ∪ (A OR (B AND C))
= A OR (B AND C)

Therefore, the most efficient way to solve it is to find intersection of B and C firstly. Then we merge it with all postings from A.

Q2.

(a) By using permuterm,

we need to expand the terms, { hello\$, world\$, word\$, hi\$, lord\$ }

And each expansion will have a pointer.

For example,

hello\$ will be expanded to
- hello\$ → 6 bytes
- ello\$h
- llo\$he
- lo\$hel
- o\$hell
- \$hello

∴ totally, it will cost:

$6 \times (6+4) + 6 \times (6+4) + 5 \times (5+4) + 3 \times (3+4) + 5 \times (5+4)$

$= 231$ bytes

By using bi-gram.

we need encode these terms: { \$h, he, el, ll, lo, o\$, \$w, wo, or, rl, ld, d\$, rd, hi, i\$, \$l }, and each has a pointer.

∴ It will cost.

$16 \times (2+4) = 96$ bytes

∵ $231 > 96$

∴ permuterm is larger.

Q2. (b).

For query $*e*lo$, in permuterm index, it can be rewriten as $*e*lo\$$. $\Rightarrow$ $lo\$*e*$

Therefore, we only need to look at terms start with $lo\$$ in B-tree, and verify that it has e in the following

(C) For query $*e*lo$, in bi-gram index,

1° we find the intersection posting lists of 'lo' and 'o\$'.

2° we need to verify that it has 'lo\$'

3° we find all postings that have 'e' in front of 'lo\$'.

Q3.

(a) Given threshold of 2.

{weird, windy, wired, wiry, word, wordy} will be considered.

(b) By using the formula. AND $\sum_{i=0}^{2} \exp(-\beta \cdot i) = 1 + 0.37 + 0.135$
$= 1.505$

P(err) of ware $= \frac{\exp(-3)}{1.505} = 0.033$

P(err) of weird $= \frac{\exp(-2)}{1.505} = 0.090$

P(err) of windy $= \frac{\exp(-1)}{1.505} = 0.244$

P(err) of wired $= \frac{\exp(-2)}{1.505} = 0.090$

P(err) of wiry $= \frac{\exp(-1)}{1.505} = 0.244$

P(err) of word $= \frac{\exp(-2)}{1.505} = 0.090$

P(err) of wordy $= \frac{\exp(-1)}{1.505} = 0.244$

Then we compare the result of P(err) · P(w).

the largest one is the correct word.

P(ware) $= 2.145 \times 10^{-3}$

P(weird) $= 4.05 \times 10^{-3}$

P(windy) $= 1.1 \times 10^{-2}$

P(wired) $= 9 \times 10^{-3}$

P(wiry) $= 7.1 \times 10^{-3}$

P(word) $= 2.25 \times 10^{-3}$

P(wordy) $= 9 \times 10^{-3}$

$\Rightarrow$ windy will be choosen as correct.

Q4.

(a) Elias-γ code is $2\log G + 1$ bits

3 bytes = 24 bits

$\therefore$ G is $2^{12}-1 = 4095$

(b) Elias-δ code is $2\log_2\log_2 k + \log_2 k$ bits

3 bytes = 24 bits

$\therefore$ k is $2^{16}-1 = 65535$.

(c). After decoding.

A = 1111 1111 1011 1100 110 $\Rightarrow$ 998.

100 $\Rightarrow$ 2

11111 0 00000 $\Rightarrow$ 32

B : 1111 1111 11 00 0000 0011 $\Rightarrow$ 515

0 $\Rightarrow$ 1

11111 0 00000 $\Rightarrow$ 32

Assume A and B all start from posting ID 1,

$\therefore$ A AND NOT B $\Rightarrow$ $\{999, 1001, 1032\}$

Q5.

Term	Query tf	df	idf	wt	D_1 tf	D_2 tf	D_3 tf	D_4 tf	D_5 tf
do	0			0	1	1	1	1	0
you	0			0	1	0	0	0	0
like	0			0	1	1	1	1	0
green	1	2	1.14	1.14	1	0	0	0	1
eggs	1	3	0.97	0.97	1	0	1	1	0
and	0			0	1	0	0	0	0
ham	0				1	0	1	0	0
not	0				0	1	1	1	0
them	0				0	1	0	0	0
or	0				0	0	1	0	0
why	0				0	0	0	0	1
are	0				0	0	0	0	1
they	0			0	0	0	0	0	1
i	0			0	0	1	1	1	0

By using $W_{t,d} = (1 + \log_{10} tf) \times \log_{10}(N/df)$ for query, we can get a vector V_{query}

for docs, we do not need consider idf,

so we can get normalised vector for each doc. V_{di}, $i \in [1, 2, 3, 4, 5]$

$\therefore$ final weight is computer cosine similarity of (V_{query}, V_{di}). Ranks: $D_1 > D_5 > D_4 > D_3$

Q6.

(a) YES, it is possible that there is a horizontal line segment in recall-precision graph.
Example is that all docs are relevant.

(b).

1. System 1, Precision at rank 8 $= \frac{1}{4}$

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

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

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

System 2, Recall at precision $\frac{1}{3}$:

Rank	3	9
Recall	$\frac{1}{4}$	$\frac{3}{4}$.

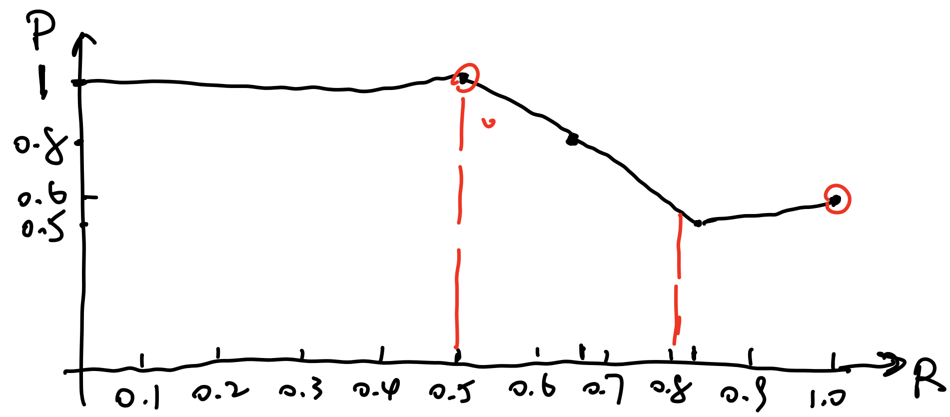
(c) MAP for System 1,

$$\begin{cases} Q1: \frac{1}{6} \times \left(1+1+1+\frac{4}{5}+\frac{5}{9}+\frac{6}{10}\right) = 0.826 \\ Q2: \frac{1}{4} \times \left(1+\frac{2}{5}+\frac{3}{9}+\frac{4}{10}\right) = 0.53 \end{cases}$$

System 2

$$\begin{cases} Q1: \frac{1}{6} \times \left(1+1+1+\frac{4}{6}+\frac{5}{7}+\frac{6}{9}\right) = 0.84 \\ Q2: \frac{1}{4} \times \left(1+\frac{2}{4}+\frac{3}{6}\right) = 0.5. \end{cases}$$

(d)



As shown in plot.

interpolated precisions for recall level 0.5 is 1.
recall level 0.8 is 0.6

Q7.

(a) By definition A Markov chain is ergodic if we have a path from any state to any other.
But in given chain, Z has no path to other states
Therefore, it is NOT ergodic.

(b)

	V	W	X	Y	Z
V	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
W	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
X	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
Y	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$
Z	0	0	0	0	1

Q7.

(C)

$$\text{final Matrix} = 0.5 \times \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + 0.5 \times \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{10} & \frac{4}{15} & \frac{4}{15} & \frac{4}{15} & \frac{1}{10} \\ \frac{7}{20} & \frac{7}{20} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \\ \frac{7}{20} & \frac{1}{10} & \frac{7}{20} & \frac{1}{10} & \frac{1}{10} \\ \frac{9}{40} & \frac{9}{40} & \frac{9}{40} & \frac{1}{10} & \frac{9}{40} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{3}{5} \end{pmatrix}$$

(d)

$$(0.2, 0.2, 0.2, 0.2, 0.2) \begin{pmatrix} \frac{1}{10} & \frac{4}{15} & \frac{4}{15} & \frac{4}{15} & \frac{1}{10} \\ \frac{7}{20} & \frac{7}{20} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \\ \frac{7}{20} & \frac{1}{10} & \frac{7}{20} & \frac{1}{10} & \frac{1}{10} \\ \frac{9}{40} & \frac{9}{40} & \frac{9}{40} & \frac{1}{10} & \frac{9}{40} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{3}{5} \end{pmatrix} = \begin{pmatrix} \frac{9}{40} \\ \frac{5}{24} \\ \frac{5}{24} \\ \frac{2}{15} \\ \frac{9}{40} \end{pmatrix} \text{ for } \begin{pmatrix} V \\ W \\ X \\ Y \\ Z \end{pmatrix}$$