Q-1

(a) Brosnan AND Sean NOT ROGER

← Assuming AND here

{d2} and {d2, d5, d6}

⇒ {d2} Ans

(b) David OR George AND (NOT Daniel)

⇒ {d1, d2, d3, d4, d5, d7} And (d3, d7)

⇒ {d3, d7} Ans

(c) Timothy AND Sean OR Brosnan OR Daniel

⇒ ∅ OR {d1, d2, d3, d4, d5, d6, d7}

⇒ {d1, d2, d3, d4, d5, d6, d7} Ans

Q-2 Dictionary = Fiji, is, an, island, country, in, the, south, Seoul, closer, to, Japan, a, having, more, than, one, reachable, from

N = 4

Next Page

| term	n_i	$\log_{10}\frac{N}{df_i}$	$w_{i,1}$	$w_{i,2}$	$w_{i,3}$	$w_{i,4}$	$w_{i,4}$
fiji	2	0.301	0.301	-	0		0.301
is	4	0	0	0	0		0
an	1	0.602	0.602	-	-		
island	2	0.301	0.301	-	0.301		
country	3	0.1249	0.124	-	0.124		0.124
in	1	0.602	0.602	-	-		
the	1	0.602	0.602	-	-		
south	1	0.602	0.602	-	-		
Seoul	1	0.602	-	0.602	-		
closer	1	0.602	-	0.602	-		
to	1	0.602	-	0.602	-		
Japan	3	0.1249	-	0.1249	0.124		0.124
a	1	0.602	-	-	0.602		
to having	1	0.602	-	-	0.602		
more	1	0.602	-	-	0.602		
than	1	0.602	-	-	0.602		
one	1	0.602	-	-	0.602		
reachable	1	0.602	-	-	0.602		
from	1	0.602	-	-	-		0.602
		0.602	-	-	-		0.602

Assuming and taking log freq $(1 + \log_{10} tf)$

query = is japan an island or country

is
japan
an
island
or
country
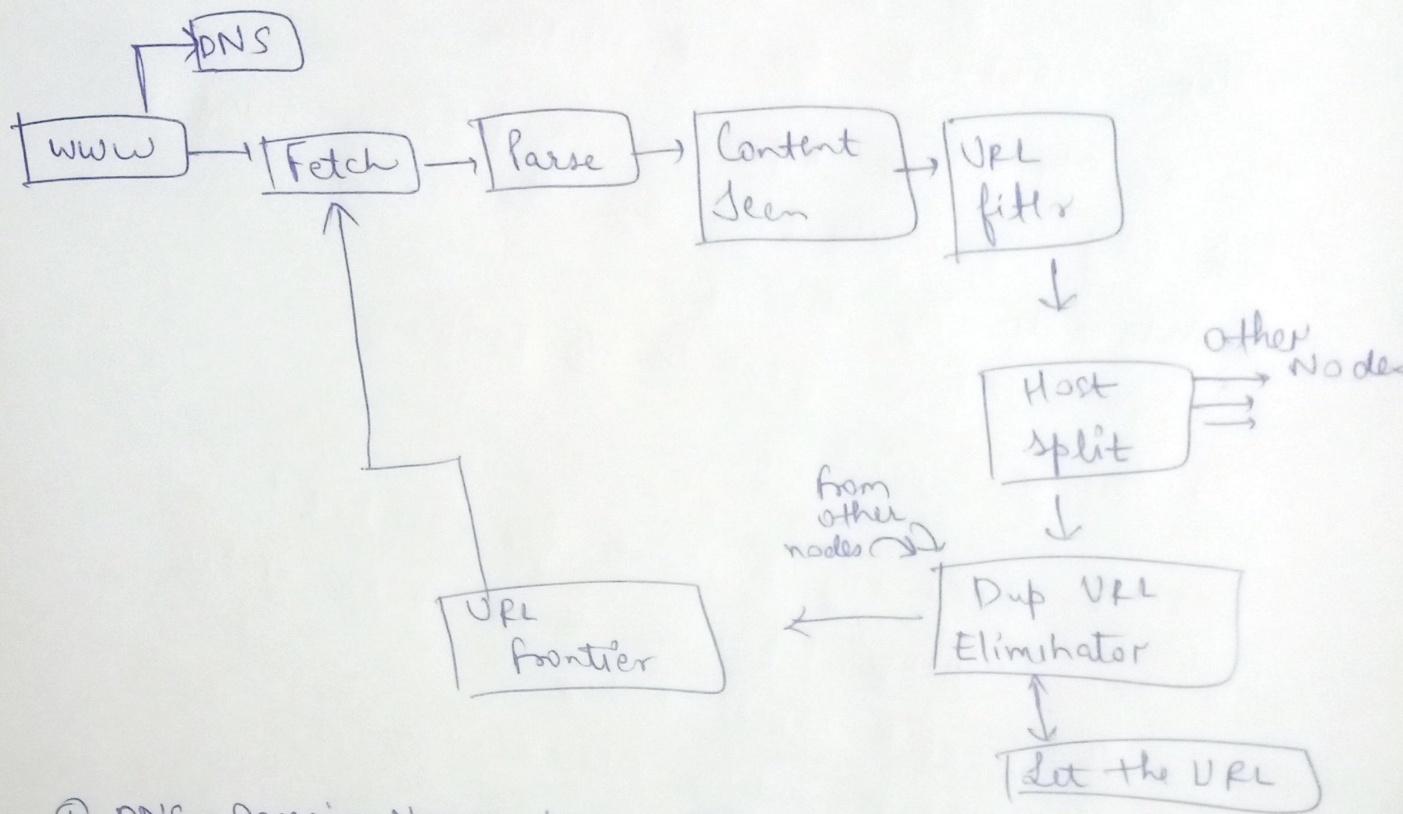
① DNS = Domain Name Server. Helps in
locating IPs of domains

② Fetch the data

③ Parse and tokenise the data

④ Apply IR to check if content is there

⑤ Filter out the spam urls

⑥ Check Relative URLs

⑦ Duplicate URLs and fetch their content

⑧ Reiterate and perform indexing/IR at end.

Q) Query Expansion,

It means addition of more words (synonyms or in context) or finding query relevant phrases for better retrieval for IR system

We can do query expansion in following ways →

① Adding synonyms (manually thesaurus)

② Automatically derived
       (from gramatically positions)

③ Query equivalence.

Example Let's say I search for "cricket",
then query expansion terms can be either

  ① cricket scores today

  ② cricket matches / tomorrow

  ③ cricket rules

  ④ cricket ground   etc. etc.

Q-4

Rank	Tyke	Recall	Precision	R	inter p
1	NR			0	0.5
2	R	1/16	1/2	0.1	0.5
3	R	2/16	2/3	0.2	0.5
4	NR			0.3	0.57
5	NR			0.4	0
6	R	3/16	1/2	0.5	0
7	R	4/16	4/7	0.6	0
8	NR			0.7	0
9	R	5/16	5/9	0.8	0
10	R	6/16	6/10	0.9	0

average precision

Average Precision

$$= \frac{1}{16}\left(\frac{1}{2} + \frac{2}{3} + \frac{1}{2} + \frac{4}{7} + \frac{5}{9} + \frac{6}{10} \right)$$
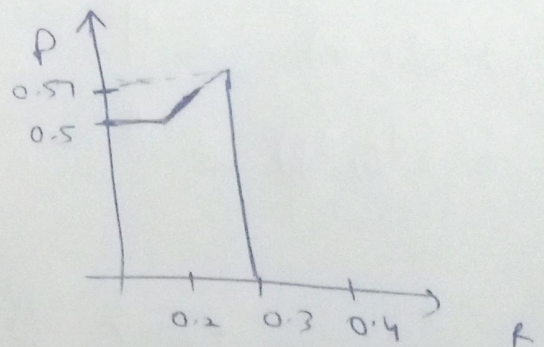
$$= 0.211$$

Let's say scores assigned

$G = \langle 0,3,3,0,0,3,3,0,3,3 \rangle$

$CG = 6 \times 3 = 18 = \sum_{j=1}^{i} G(i)$

Cummulative Gain $= 18$

     N=5

$q = 110001$

taking 1st, 2nd and 6th column

$d_1 = 1\ 1\ 1$ $P_i = 0.5$ top 2

$d_2 = 1\ 0\ 1$

$d_3 = 0\ 0\ 1$

$d_4 = 1\ 1\ 0$

$d_5 = 1\ 0\ 1$

(a)

	t_1	t_2	t_3
n_i	4	2	4
u_i	.4	.2	0.4

$SC(d_1, q) = \cancel{log\ p_3} + p_3 \quad \sum_{i=1}^{3}\left(log\ \dfrac{p_i}{1-p_i} + log\ \dfrac{1-u_i}{u_i}\right) = \cancel{14}\ 10$

$SC(d_2, q) = \left(\dfrac{log\ p_1}{1-p_1} + log\ \dfrac{1-u_1}{u_1}\right) + log\dfrac{p_3}{1-p_3} + log\left(\dfrac{1-u_3}{u_3}\right) = \cancel{10}\ 6$

$SC(d_3, q) = 3/2$

$SC(d_4, q) = 1/2$

$SC(d_5, q) = 6$

S_0
$d_2 > (d_1 = d_4) > d_5 > d_3$

Ans

$d_1 > (d_5 = d_2) > d_4 > d_3$ Ans