IR - Assignment - 2

SRUJAN.R
BT18CSE041

1. (a) Relevant documents = {3, 4}
documents retrieved = {2, 5, 6, 7, 8}
∴ precision = $\frac{0}{5}$ = 0

recall = $\frac{0}{2}$ = 0

(b) Relevant documents = {3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
documents retrieved = {2, 5, 6, 7, 8}

∴ precision = $\frac{4}{4+1}$ = $\frac{4}{5}$ = 0.8

recall = $\frac{4}{10}$ = 0.4

(c) Average precision in case (a) = 0

Average precision in case (b)
= $\frac{1}{4}\left(\frac{1}{2} + \frac{2}{3} + \frac{3}{4} + \frac{4}{5}\right)$ = 0.679

(d) Kappa measure is a common measure per argument between the judges. It is designed for categorical judgements and corrects a simple argument rate for the rate of chance of arguments.

		J_2 Relevance		
		Y	N	Total
J_1 Relevance	Y	2	5	7
	N	3	2	5
	Total	5	7	12

P(A) = observed proportion of times the judges agreed.

$$= \frac{YY + NN}{total} = \frac{4}{12} = 0.33$$

$$P(NR) = \frac{7+5}{12+12} = \frac{12}{24} = 0.5$$

$$P(R) = \frac{7+5}{12+12} = \frac{12}{24} = 0.5$$

Probability that 2 judges agreed by chance

$$P(E) = P(NR)^2 + P(R)^2 = (0.5)^2 + (0.5)^2$$
$$= 0.5$$

$$\text{Kappa} = K = \frac{(P(A) - P(E))}{(1 - P(E))}$$
$$= \frac{0.33 - 0.5}{1 - 0.5} = \frac{-0.17}{0.5}$$
$$= -0.34 < 0.67$$

(e) As the Kappa measure is less than 0.67, it is not acceptable.

2. The F1 is defined as a Harmonic mean because Arithmetic and Geometric mean both have high values even if the corpus has only a few relevant documents and thousands of non-relevant document whereas harmonic means are closes to the small value and thus give a realistic picture.

The intuition is to balance precision and recall. F1 score can't be high if either of precision of recall is low.

So, there must be a point where precision = reall which could be potentially an optimal point. Such point is called break-even point.

∴ At break-even point, $F_1 = P = R$.

We derived above from,

$$F_1 = \frac{2PR}{P+R}$$

$$= \frac{2P.P}{2P} \quad \{\because P = R\}$$ at break-even point

$$= \frac{2P^2}{2P}$$

$$\therefore \boxed{F_1 = P = R}$$

3. (a) If no skip list is used then comparisons made are:

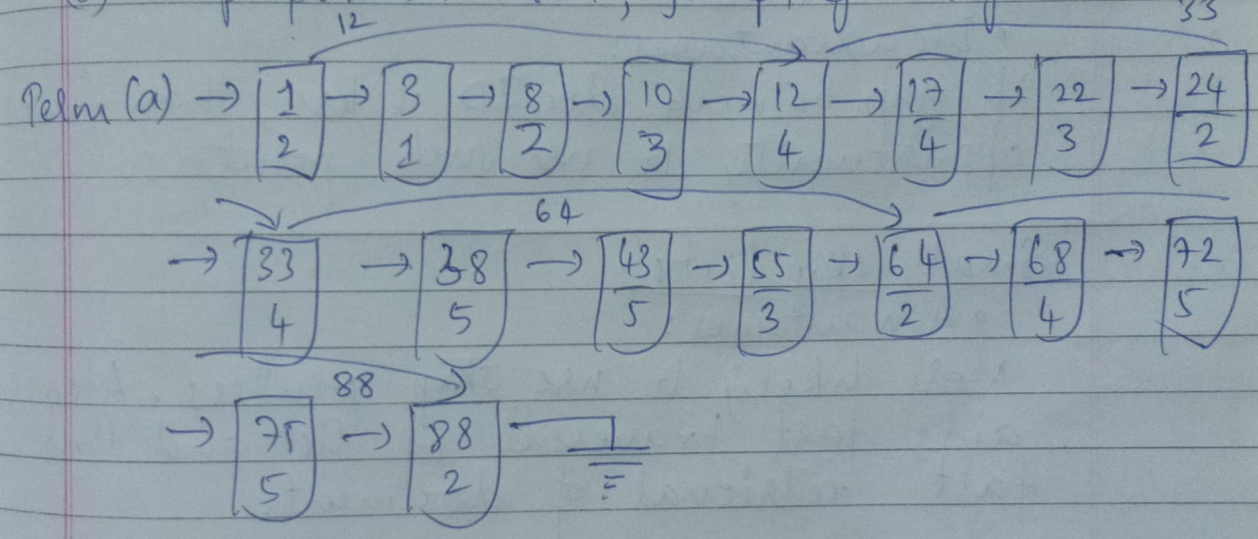
term a	term b
1	10
3	10
8	10
10	10
12	56
17	56
22	56
24	56
33	56
38	56
43	56
55	56
64	56

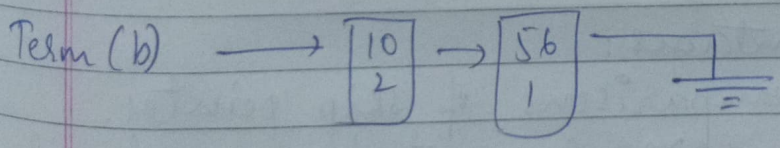
At this point the docID for Term 'a' is greater than doc ID for term 'b' and then the list for term 'b' ends so the algo -rithm terminates returning only

doc ID = 10

so, no of comparisons made = 13

(b) skip pointer used, jumping length = 4





Comparisons made = 12

term a	term b
1	10
12	10
3	10
8	10
10	10
12	56
33	56
64	56
38	56
43	56
55	56
64	

(c) Large skip lists

- Advantages:

1) Can skip the list with less comparisons for large gap between doc ID.
2) Space needed to store the skip pointers will be less as no of such pointers will be less.

- Disadvantages:

if gap between doc ID is less, than fewer oppurtunity to use such pointers

Small skip lists:

- Advantages:

More likely to use skip pointers. Also aids fast traversal of list and thus fast retrieval of document.

- Disadvantages:

1) More comparisons of skip pointer.
2) More space requirement because of more no of pointers.

4. Suppose, Q consists of 2 query term M and N, and M is frequent in D but rare in E, and N is frequent in E but rare in D.

Now suppose, M is rare in documents of A and frequent in documents of B, and N is frequent in documents of A but rare in documents of B.

Now, in context of A, idf value of M is more and idf value of N is less. And as D has frequent no of A, the cosine score of D will be more than the cosine score of E with respect to Q in context of A.

∴ D will be more relevant than E. In context of B, idf value of N is more and idf value of M is less. As E has more no of N as compared to D, the cosine score of E will be more than that of D with respect to Q in context of B.

∴ E will be more relevant than D.