Q1) 5 Q1: D1, D2, D3, D4, D5, D6, D7, D8, D9, D10
4 Q2: D1, D2, D3, D4, D5, D6, D7, D8, D9, D10

a) R-Precision for Q1 & Q2?

$$\text{R-Precision} = \frac{\#\text{ of relevant docs in the top R retrieved docs}}{R}$$

R-Precision for $Q1 = \frac{3}{5} = 0.6$

R-Precision for $Q2 = \frac{2}{4} = 0.5$

b) P@10 & R@10 for Q1 & Q2?

$$P@10 = \frac{TP}{TP+FP} @10 \qquad R@10 = \frac{TP}{TP+FN} @10$$

Q1: TP: 5
FP: 5
TN: 0
FN: 0

Q2: TP: 4
FP: 6
TN: 0
FN: 0

P@10 for Q1 = 5/10 = 0.5   R@10 for Q1 = 5/5 = 1
P@10 for Q2 = 4/10 = 0.4   R@10 for Q2 = 4/4 = 1

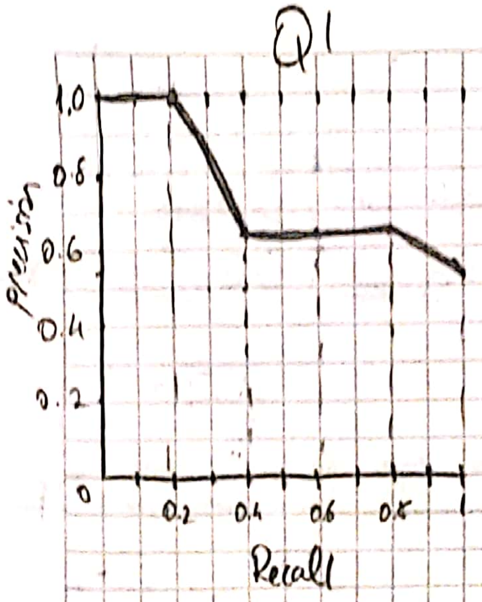
c)

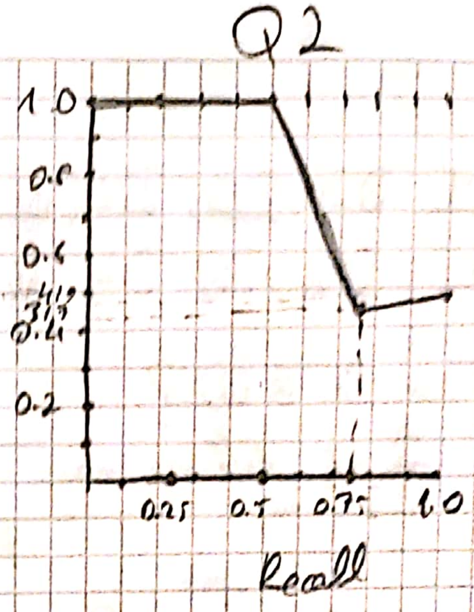
Q1 Precision-Recall Table

# of Docs	1	2	3	4	5	6	7	8	9	10
Relevance	1	0	1	0	1	1	0	0	1	0
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	5/9	5/10
Recall	1/5	1/5	2/5	2/5	3/5	4/5	4/5	4/5	5/5	5/5

Q2 Precision-Recall Table

# of Docs	1	2	3	4	5	6	7	8	9	10
Relevance	1	1	0	0	0	0	1	0	1	0
Precision	1/1	2/2	2/3	2/4	2/5	2/6	3/7	3/8	4/9	4/10
Recall	1/4	2/4	2/4	2/4	2/4	2/4	3/4	3/4	4/4	4/4





Precision - Recall Curves for Q1, Q2 Respectively.

Purpose of the TREC Interpolated Approach

- To create a consistent evaluation standard by using fixed recall levels. This makes results more
- To smooth Precision - Recall curves. By interpolation, we are sure that curves remain non-increasing, which makes them more interpretable.

According to TREC 6 Appendix A, the purpose of interpolation is to facilitate computing average performance over a set of topics.

Q2 In evaluating the effectiveness of an interactive IRS, I would use <u>R-precision.</u> Because R-precision is calculated based on the total number of relevant docs in the collection. However, P@10 considers only 10 documents (relevant or irrelevant.) Since it is an interactive system, users explore results beyond a single fixed rank; R-precision is more suitable in this case.

Q3

$$D = \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} & \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} & \begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \end{array}$$

(a) Straightforward approach (Brute-force)

\# of calculation: $\frac{n.(n-1)}{2} = \frac{6.5}{2} = 15$

(b) Term inverted indexes:

$t_1 \to d_3, d_4$

$t_2 \to d_1, d_2$

$t_3 \to d_3, d_5$

$t_4 \to d_2, d_3, d_4$

$t_5 \to d_1, d_2, d_5, d_6$

$t_6 \to d_5, d_6$

Calculated Similarities

$d_1 : \{d_1, d_2, d_5, d_6\} \Rightarrow S_{12}, S_{15}, S_{16}$

$d_2 : \{d_1, d_3, d_4, d_5, d_6\} \Rightarrow S_{23}, S_{24}, S_{25}, S_{26}$

$d_3 : \{d_2, d_3, d_4, d_5\} \Rightarrow S_{34}, S_{35}$

$d_4 : \{d_2, d_3, d_4\} \Rightarrow$ No new similarity

$d_5 : \{d_1, d_2, d_3, d_5, d_6\} \Rightarrow S_{56}$

$d_6 : \{d_1, d_2, d_5, d_6\} \Rightarrow$ No new similarity

10 calculation: $S_{12}, S_{15}, S_{16}, S_{23}, S_{24}, S_{25}, S_{26}, S_{34}, S_{35}, S_{56}$

(c) Using inverted index approach:

$$S = \begin{array}{c} \begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \end{array} & \\ \begin{bmatrix} 1 & S_{12} & - & - & S_{15} & S_{16} \\ X & 1 & S_{23} & S_{24} & S_{25} & S_{26} \\ X & X & 1 & S_{34} & S_{35} & - \\ X & X & X & 1 & - & - \\ X & X & X & X & 1 & S_{56} \\ X & X & X & X & X & 1 \end{bmatrix} & \begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \end{array}$$

Dice Coefficient: $\frac{2 \sum X_i Y_i}{\sum X_i^2 + \sum Y_i^2}$

$S_{12} = \frac{2(1+1)}{2+3} = \frac{4}{5}$ $S_{15} = \frac{2.1}{2+3} = \frac{2}{5}$ $S_{16} = \frac{2.1}{2+2} = \frac{1}{2}$

$S_{23} = \frac{2.1}{3+3} = \frac{1}{3}$ $S_{24} = \frac{2.1}{3+2} = \frac{2}{5}$ $S_{25} = \frac{2.1}{3+3} = \frac{1}{3}$ $S_{26} = \frac{2.1}{3+2} = \frac{2}{5}$

$S_{34} = \frac{2(1+1)}{3+2} = \frac{4}{5}$ $S_{35} = \frac{2.1}{3+3} = \frac{1}{3}$ $S_{56} = \frac{2.(1+1)}{3+2} = \frac{4}{5}$

$$S = \begin{bmatrix} 1 & 0.8 & 0 & 0 & 0.4 & 0.5 \\ X & 1 & 0.33 & 0.4 & 0.33 & 0.4 \\ X & X & 1 & 0.8 & 0.33 & 0 \\ X & X & X & 1 & 0 & 0 \\ X & X & X & X & 1 & 0.8 \\ X & X & X & X & X & 1 \end{bmatrix}$$

Q4 (a) I will prove by example.

Let's say: $D = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 100 & 200 & 0 & 100 & 0 \end{bmatrix}$

$$\underset{d_s}{S} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/4 & 1/2 & 0 & 1/4 & 0 \end{bmatrix}_{2\times 5} \qquad \underset{k_s}{S'} = \begin{bmatrix} 1/101 & 1/201 & 0 & 0 & 0 \\ 100/101 & 200/201 & 0 & 1 & 0 \end{bmatrix}$$

$C = S S'^T$

$$S = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/4 & 1/2 & 0 & 1/4 & 0 \end{bmatrix} \qquad S'^T = \begin{bmatrix} 1/101 & 100/101 \\ 1/201 & 200/201 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}_{5\times 2}$$

$$C = \begin{bmatrix} \frac{1}{202} + \frac{1}{402} & \frac{100}{202} + \frac{200}{402} \\ \frac{1}{404} + \frac{1}{402} & \frac{100}{404} + \frac{200}{402} \end{bmatrix}$$

As it can be seen from left, $c_{11} < c_{12}$

So, c_{ii} may be less than c_{ij} by the example above.

(b) If $\forall c_i = 1/m$, this means $\forall d_i$ in D are identical. So, $d_i = d_j$. This implies that $c_{ii} = c_{ij}$ for $\forall i,j \leq m$. Verbally, every document in the collection is identical. So, their coverage are the same for all, there is only one cluster, every document are in that cluster. Their coverage is $1/m$ for $\forall$ doc in D.

(c) $n_c = n_c'$ (# term clusters are equal to # doc clusters)

$$\sum_{i=1}^{m} \delta_i = \sum_{i=1}^{n} \delta_i'$$

Multiply by $\frac{1}{m}$

$$\delta = \frac{1}{m} \sum_{i=1}^{n} \delta_i'$$

multiply by $\frac{1}{n}$

$$\frac{\delta}{n} = \frac{1}{m} \delta'$$

$$\delta = \frac{n}{m} \delta'$$

MOS

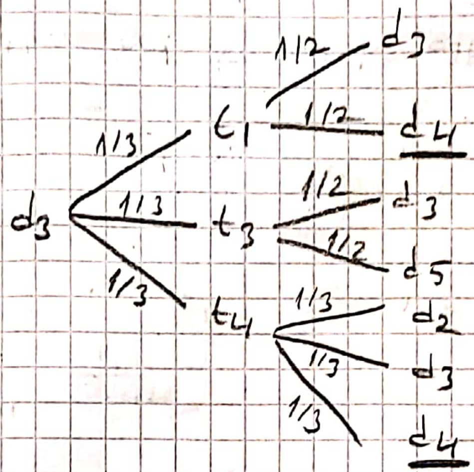
Q5

$$D = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$\alpha_1 = 1/2$, $\alpha_2 = 1/3$, $\alpha_3 = 1/3$, $\alpha_4 = 1/2$, $\alpha_5 = 1/3$, $\alpha_6 = 1/2$

$\beta_1 = 1/2$, $\beta_2 = 1/2$, $\beta_3 = 1/2$, $\beta_4 = 1/3$, $\beta_5 = 1/4$, $\beta_6 = 1/2$

a



$$c_{34} = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} = 5/18 \approx 0.28$$

using double stage experiment

$c_{34} = 5/18 \approx 0.28.$

b

$$S = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix}_{6\times 6}$$

$$S' = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/4 & 0 \\ 0 & 1/2 & 0 & 1/3 & 1/4 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/4 & 1/2 \\ 0 & 0 & 0 & 0 & 1/4 & 1/2 \end{bmatrix}_{6\times 6}$$

$$\begin{array}{cc} & \begin{array}{cccccc} & & \text{seed} & & \text{seed} & \end{array} \\ C = SS'^T = & \begin{bmatrix} 3/8 & 3/8 & 0 & 0 & 1/8 & 1/8 \\ 1/4 & 13/36 & 1/9 & 1/9 & 1/12 & 1/12 \\ 0 & 1/9 & 8/18 & 5/18 & 1/6 & 0 \\ 0 & 1/6 & 5/12 & 5/12 & 0 & 0 \\ 1/12 & 1/12 & 1/6 & 0 & 5/12 & 1/4 \\ 1/8 & 1/8 & 0 & 0 & 3/8 & 3/8 \end{bmatrix} \end{array}$$

$i =$
$j = 6$

$$n_c = \sum c_{ii} = \frac{3}{8} + \frac{13}{36} + \frac{8}{18} + \frac{5}{12} + \frac{5}{12} + \frac{3}{8} = \frac{6}{8} + \frac{44}{36} = \frac{11}{9} + \frac{6}{8} = \frac{142}{72}$$
$$= 1.97$$

Seed power of $d_i \Rightarrow P_i = \delta_i \ \psi_i \cdot$ (# terms in d_i)

δ_i : Decoupling Coefficient : C_{ii}

ψ_i : Coupling Coefficient $(1 - \delta_i)$

$P_1 = (3/8) \cdot (5/8) \cdot 2 = 0.47$
$P_2 = (13/36) \cdot (23/36) \cdot 3 = 0.69$
$P_3 = (8/18) \cdot (10/18) \cdot 3 = 0.74 \rightarrow$ seed
$P_4 = (5/12) \cdot (7/12) \cdot 2 = 0.49$
$P_5 = (5/12) \cdot (7/12) \cdot 3 = 0.73 \rightarrow$ seed
$P_6 = (3/8) \cdot (5/8) \cdot 2 = 0.47$

$\boxed{n_c \approx 2 \text{. So, we have 2 clusters}}$

Clusters $\rightarrow \{d_1, \underset{}{\overset{\text{(seed)}}{d_5}}, d_6\}$
$\{d_2, \underset{\text{(seed)}}{d_3}, d_4\}$

To assign non-seed to seeds, $(m-n_c) \times n_c = (6-2) \times 2 = 8$ calculations on entries of C matrix. In addition, 6 C_{ii} values also needed to be calculated for seed power of documents. In total, 14 entries of C matrix should be calculated.

Inverted Index for Seed Documents:

$t_1 \rightarrow \langle 3,1 \rangle$

$t_2 \rightarrow \langle \quad \rangle$

$t_3 \rightarrow \langle 3,1 \rangle, \langle 5,1 \rangle$

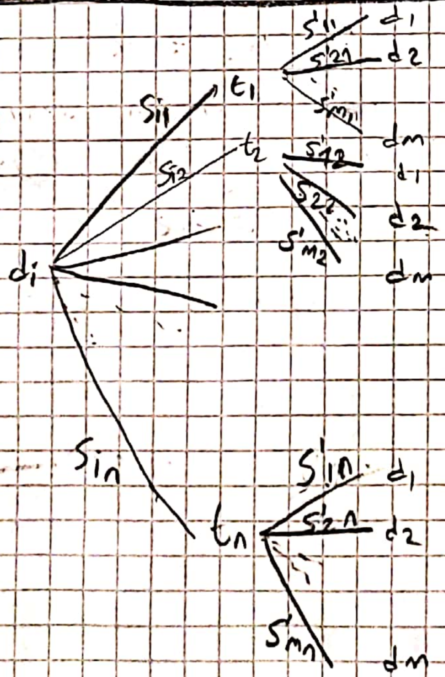
$t_4 \rightarrow \langle 3,1 \rangle$

$t_5 \rightarrow \langle 5,1 \rangle$

$t_6 \rightarrow \langle 5,1 \rangle$

Q6

a



Maximum # of active branches for a C matrix entry is reached when each document contains every term. For a document, n number of active branches for n terms and mn number of active branches for m documents. For document i $n+mn$ active branches are obtained.

For all documents, $m(n+mn)$ active branches in total, as there are m docu

(b) $d_i \overset{s_{ij}}{\text{———}} t_j \overset{s'_{ij}}{\text{———}} d_i$

Minimum number of active branches is obtained when all docs contain any of the terms. Since each document contains at least 1 term, there should be 1 branch for term and 1 branch for document. Thus, there must be 2 active branches for a document. In total, minimum of 2m active branches exist as we have m docs.

Q7 $$n_c = \frac{mn}{t} = \frac{m}{t_g} = \frac{n}{x_d}$$

t_g: term generality

x_d: depth of indexing

Indexing clustering relationship implied by the cover coefficient concept is intuitively justifiable. If the # of nonzero entries in D matrix is increased, the similarity among docs is increased; hence, a smaller n_c results. The reverse, more zero entries, would imply higher n_c. Thus, the relationship above can be used to check the clustering tendency of a document collection.

Q8

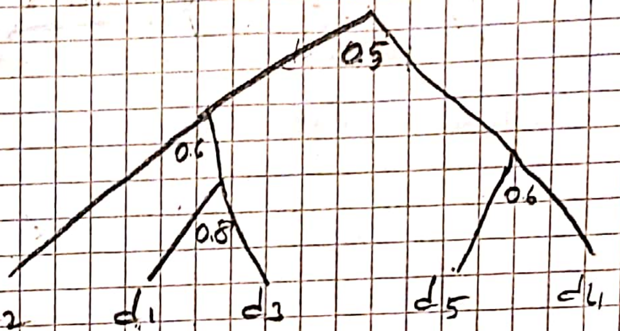
$$\begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & \\ \end{array}$$
$$\begin{bmatrix} 1 & 0.4 & 0.8 & 0.5 & 0.3 \\ X & 1 & 0.6 & 0.3 & 0.4 \\ X & X & 1 & 0.5 & 0.2 \\ X & X & X & 1 & 0.6 \\ X & X & X & X & 1 \end{bmatrix} \begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array}$$

Single link

Step	Pair	Sim value
1	$d_1 - d_3$	0.8
2	$d_2 - d_3$	0.6
3	$d_4 - d_5$	0.6
4	$d_1 - d_4$	0.5
5	$d_3 - d_4$	0.5
6	$d_1 - d_2$	0.4
7	$d_2 - d_5$	0.4
8	$d_1 - d_5$	0.3
9	$d_2 - d_4$	0.3

a) The List C will be

[$d_1 - d_3$, $d_2 - d_3$, $d_4 - d_5$, $d_3 - d_4$]

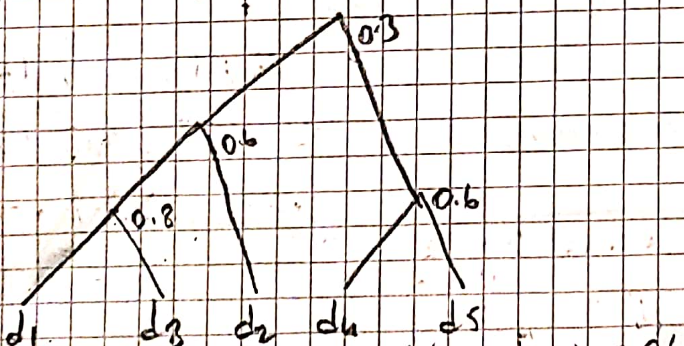


Ultimate clustering structure by single-link

$$S' = \begin{bmatrix} 1.0 & 0.6 & 0.8 & 0.5 & 0.5 \\ X & 1.0 & 0.6 & 0.5 & 0.5 \\ X & X & 1.0 & 0.5 & 0.5 \\ X & X & X & 1.0 & 0.6 \\ X & X & X & X & 1.0 \end{bmatrix}$$

Similarity matrix implied by the dendrogram.

b) complete link



$d_2 - d_3 = 0.6$ $d_4 - d_5 = 0.6$

Ultimate clustering structure by complete link

$$S' = \begin{bmatrix} 1.0 & 0.6 & 0.8 & 0.3 & 0.3 \\ X & 1.0 & 0.6 & 0.3 & 0.3 \\ X & X & 1.0 & 0.3 & 0.3 \\ X & X & X & 1.0 & 0.6 \\ X & X & X & X & 1.0 \end{bmatrix}$$

Similarity Matrix implied by the dendrogram

Q9

$m = 360$

$n_c = 24$

$k = 3$

Cluster size: $p = \frac{m}{n_c} = \frac{360}{24} = 15$

By the Yao's formula from the given paper

$$E(X) = n_c \left[1 - \prod_{i=1}^{k} \frac{md - i + 1}{m - i + 1} \right]$$

where $d = 1 - \frac{1}{n_c}$

$d = 1 - \frac{1}{24}$

$d = \frac{23}{24}$

$$= 24 \left[1 - \left(\frac{23}{24} \cdot \frac{344}{359} \cdot \frac{343}{358} \right) \right]$$

≈ 2.88

The expected # of clusters to be accessed to retrieve all relevant docs is approximately 2.88.
Therefore, we can expect 3 clusters to retrieve all relevant documents.

Q10

$P_1 = \{a, b, c\} \; \{d, e\}$

$P_2 = \{a\} \; \{b, c, d\} \; \{e\}$

$$\text{Rand Index} = \frac{TP + TN}{TP + TN + FP + FN}$$

a) P_2 is ground truth

Pair	ab	ac	ad	ae	bc	bd	be	cd	ce	de
Class	FP	FP	TN	TN	TP	FN	TN	FN	TN	FP

$$\text{Rand Index} = \frac{(4 + 1)}{10} = 0.5$$

b) P_1 is ground truth

Pair	ab	ac	ad	ae	bc	bd	be	cd	ce	de
Class	FN	FN	TN	TN	TP	FP	TN	FP	TN	FN

$$\text{Rand Index} = \frac{(4 + 1)}{10} = 0.5$$

Q11 a: <1,2>, <3,2>, <9,2> <10,3> <12,4>, <18,4>, <20,3> <23,3>, <25,4>, <33,4>, <37,4>, <40,5>, <43,4>, <55,3>, <64,2>, <68,4>, <72,3>, <75,1>, <88,2>

b: <15,7>, <66,3>, <75,1> <90,2>

a) Compare term b's <15,7> with <1,2> <3,2> <9,2> <10,3>, <12,4> and <18,4> => 6 comparisons.

Compare term b's <66,3> with <20,3> <23,3>, <25,4> <33,4>, <37,4>, <40,5> <43,4>, <55,3>, <64,2>, <68,4>
=> 10 comparisons

Compare term b's <75,1> with <72,3>, <75,1>
=> 2 comparisons

Compare term b's <90,2> with <88,2> => 1 comparison.

So, we can find intersection of term a & term b lists without using skipping 19 comparisons

b) Skip with chunk = 3 each chunk has lowest doc from next

term a

chunk 1: <1,2> <3,2>, <9,2>
chunk 2: <9,2> <10,3> <12,4>
chunk 3: <12,4> <18,4>, <20,3>
chunk 4: <20,3> <23,3>, <25,4>
chunk 5: <25,4>, <33,4>, <37,4>
chunk 6: <37,4>, <40,5>, <43,4>
chunk 7: <43,4>, <55,3>, <64,2>
chunk 8: <64,2>, <68,4>, <72,3>
chunk 9: <72,3>, <75,1>, <88,2>

For <75,1> chunk 8: False
chunk 9: True
| inside 2

For <75,1> => 4 comparisons

For <15,7> of term b
is it in chunk 1: False
chunk 2: False
chunk 3: True
inside chunk 2
For <15,7> => 5 comparisons

For <66,3> chunk 3: false
chunk 4: false
chunk 5: false
" 6: false
" 7: false
chunk 8: True
inside chunk 8: 2
For <66,3> => 8 comparisons.

For <90, 2> chunk 9 : True
inside 3 comparisons

<90, 2> ⟹ 4 comparisons.

In total 5 + 8 + 4 + 4 = 21 comparisons needed.

(c)

	Small Skips	Large Skips
Advantages	More chunks can be skipped Less comparisons within the chunk as chunk is small.	Total # chunks decreases Total # of comparisons with chunk descriptors decreases
Disadvantages	The # of comparisons with the chunk descriptors increases	The # of comparisons within the chunk increases.

(d) I believe it is not possible to take advantage of a skipping structure for disjunctive queries.
In the case of a conjunctive (AND) queries, because it takes intersection of posting queries in order to find the common terms, it is useful. However, disjunctive (OR) queries take the union of posting lists. So, it is useless to find common terms as union of posting queries is the set of docs either in one query or in the other or both. It is not useful for finding the common terms of posting lists.

Q12

term 1 = <16,4>, <34,3>, <47,4>, <109,7>

term 2 = <15,3>, <22,3>, <33,2> <34,6> <86,4>, <108,7>

(a) term 1 = <109,7>, <16,4>, <47,4> <34,3>

term 2 = <108,7>, <34,6> <86,4>, <15,3>, <22,3> <33,2>

(b) For doc 16 : $f_{dt} = 4$

For doc 34 : $f_{dt_1} = 3$ $f_{dt_2} = 6$ total => 9

For doc 47 : $f_{dt} = 4$

For doc 109 : $f_{dt} = 7$

For doc 15 : $f_{dt} = 3$

For doc 22 : $f_{dt} = 3$

For doc 33 : $f_{dt} = 2$

For doc 86 $f_{dt} = 4$

For doc 108 $f_{dt} = 7$

i) Ranked list

34, 108, 109, 16, 47, 86, 15, 22, 33

ii) Top 5 results

34, 108, 109, 16, 47

iii) Round 1 : Process <109,7> from 1 <108,7> from 2

Round 2 : Proces <16,4> from 1 <34,6> from 2

Round 3 : Process <47,4> from 1 <86,4> from 2

Top 5 after interleaving

34, 109, 108, 16, 47