COMP 336/533 Information Retrieval

MID-TERM EXAMINATION

October 17, 1996

I hereby declare that no illegal aids were used in completing this examination. And I understand that any act of cheating will result in severe penalty, including expulsion from the university.

Sign: ______________________ Date: ____________

Name: ______________________ Student ID: ____________ Circle : 336/533

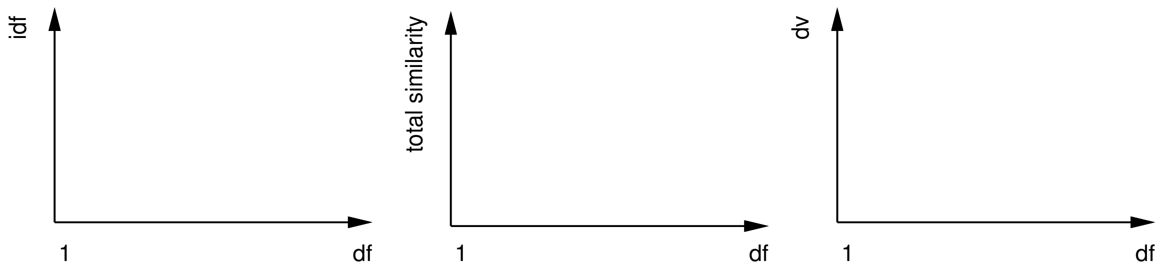
1. **[15]** Plot on the graphs below the typical behavior of the following parameters when a term appears in more and more documents (i.e., document frequency, df, increases):

   (a) the inverse document frequency (idf) of the term,

   (b) the total document similarity between the set of documents, and

   (c) the term discrimination value (dv) of the term.

   Notice that we are only concerned with the "average" or "typical" behaviour.



2. **[20]** Using the (improved) KMP method as described in the textbook (i.e., NOT the "improved" method), fill in the following shift array, # is the end-of-string character.

pattern char	no. of shifts
b	
a	
b	
d	
b	
d	
#	

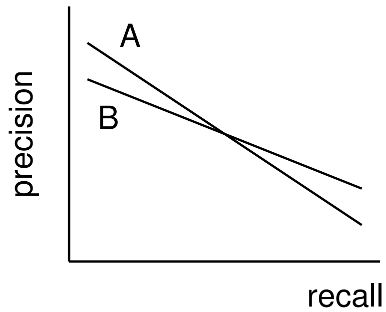
What are the values for the next[] array?

3. **[10]** Given the following precision/recall graph for two systems, A and B, which have *exactly* the same average precision:

   (a) under what situation will a user prefer A over B? why?

   (b) under what situation will a user prefer B over A? why?

   Notice that answers like "A is better when recall is low" and "B is better when recall is high" are NOT what I am looking for.



4. **[20]** Fill in the precision and recall values in the following table. 'x' means the document is relevant. There are a total of 3 relevant documents among a total of 100 documents.

	Recall-precision after retrieval of n documents				
	n	Doc ID	Recall	Precision	Fall Out
	1	a			
x	2	b			
	3	c			
x	4	d			
x	5	e			
	6	f			
	...	...	....	...	...
	...	...	....	...	...
	100	...			

5. **[15]** Given the following inverted file (or index):

x	$\longrightarrow$	1, 2	3, 1
y	$\longrightarrow$	2, 3	
z	$\longrightarrow$	2, 3	1, 1

A postings is of the form $\langle docid, tf \rangle$. Suppose there are only 3 documents,

(a) What are the *idf* values for the three terms, x, y, z? the index file?

x:
y:
z:

(b) Given the query "x y" with query term weights of 1 for both query terms, calculate the scores of the documents using tfxidf weights and the cosine similarity measures.

(c) If the query term weights are changed to 0.5, does it change the document scores? does it change the document ranking? Why? You don't need to recalculate the scores. A qualitative explanation is sufficient.