1. **[30]** Using the KMP algorithm (including the improvement on slide 25 and cascade mismatch), determine the number of character positions to shift in the shift array for the following pattern.

Basic (before applying cascade mismatch):

pattern	**a**	**b**	**b**	**a**	**b**	**b**	**a**	**a**
shift	1	1	2	3	3	3	3	3

After applying cascade mismatch:

pattern	**a**	**b**	**b**	**a**	**b**	**b**	**a**	**a**
shift	1	1	2	4	4	5	7	3

Based on the shift table obtained, show the steps in matching the pattern against the text string in the following table. In each step, circle the pattern character that causes the mismatch.

Text ->	**a**	**b**	**c**	**a**	**b**	**b**	**a**	**a**	**b**	**b**	**a**	**b**	**b**	**a**	**a**
Pattern ->	a	b	**b**	a	b	b	a	a							
Step 1			**a**	b	b	a	b	b	a	a					
Step 2				a	b	b	a	**b**	b	a	a				
Step 3								a	**b**	b	a	b	b	a	a
Step 4															

FOR REFERENCE ONY: Using the basic (before optimizing for cascade mismatch):

Text ->	**a**	**b**	**c**	**a**	**b**	**b**	**a**	**a**	**b**	**b**	**a**	**b**	**b**	**a**	**a**
Pattern ->	a	b	**b**	a	b	b	a	a							
Step 1			**a**	b	b	a	b	b	a	a					
Step 2				a	b	b	a	**b**	b	a	a				
Step 3							a	**b**	b	a	b	b	a	a	
Step 4								a	b	b	a	b	b	a	a

2. **[25]** Fill in the precision, recall and fallout values in the following table. There are a total of 100 documents, and all of the relevant documents are shown in the table; they are marked with a √ in the first column. Draw the precision/recall graph and fallout/recall graph as in slide 11 of the lecture notes (i.e., no need to interpolate or smooth the graph).

	Rank	doc ID	Recall	Precision	Fallout
	1	1001	0	0	1/96=0.01
	2	2873	0	0	2/96=0.021
	3	3916	0	0	3/96=0.031
	4	0983	0	0	4/96=0.042
	5	8310	0	0	5/96=0.052
√	6	7892	1/4=0.25	1/6=0.1667	5/96=0.052
	7	4562	1/4=0.25	1/7=0.143	6/96=0.063
√	8	4921	2/4=0.5	2/8=0.25	6/96=0.063
√	9	7934	3/4=0.75	3/9=0.333	6/96=0.063
	10	9248	3/4=0.75	3/10=0.3	7/96=0.073
...	...	...	...	...	...
	98	1688	3/4=0.75	3/98=0.031	95/96=0.99
√	99	0926	4/4=1	4/99=0.04	95/96=0.99
	100	3861	4/4=1	4/100=0.04	96/96=1

Given that the top 10 documents are retrieved. Compute:

(i) the average precision Ap(q) and explain what is "q" in the question, i.e., "q is the query which …"

(ii) $NDCG_{10}$ assuming that relevant documents get scores of 1 and non-relevant documents get scores of 0.

ANS:

(1) Ap(q)= (1/6+2/8+3/9)/3 =0.25
q is the query which

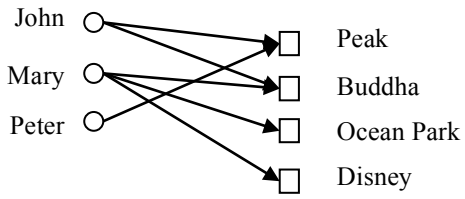
(2) $DCG_{10} = 0 + 1/\log_2 6 + 1/\log_2 8 + 1/\log_2 9 = 1.036$

$IDCG_{10} = 1 + 1/\log_2 2 + 1/\log_2 3 = 2.631$

$NDCG_{10} = \frac{DCG_{10}}{IDCG_{10}} = 0.394$

Note: The point of the graph is that precision can start from 0, goes up and eventually come down to nearly zero.

3. [25] The bipartite graph represents the places visited by a traveler.



A link indicates that the person has visited a place. Define the **interestingness** of a place as the summation of the **experience** of travelers who visited the place, and the experience of a traveler as the summation of the interestingness of places he/she visited.

(a) Compute the interestingness of each place and experience of each traveler using the hub/authority metaphor. Assuming at the beginning all hub and authority values are one. Perform 2 iterations.
(b) Given that Peak and Buddha both have two travelers explain why they get the interestingness values obtained in (a).
(c) In plain English, how would you describe an interesting place and an experienced traveler, each in one sentence? I.e., "A place is interesting if it … A traveler is experienced if he/she …"

H(John) = A(Peak) + A(Buddha)
H(Mary) = A(Buddha) + A(Ocean Park) + A(Disney)
H(Peter) = A(Peak)

A(Peak) = H(John) + H(Peter)

A(Buddha) = H(John) + H(Mary)
A(Ocean Park) = H(Mary)
A(Disney) = H(Mary)

Solution: [The graph in question is changed.]

(a) In the bipartite graph, view sites have authority weight and travelers have hub weight. After two rounds we can get:

Authority\round	0	1	2
Peak	1	2	3
Buddha	1	2	5
Ocean Park	1	1	3
Disney	1	1	3

Hub\round	0	1	2
John	1	2	4
Mary	1	3	4
Peter	1	1	2

(b) According to (a), Buddha's interesting value is larger than Peak's. Peak and Buddha both have two travelers. Peak has John and Peter, Buddha has John and Mary. As a traveler, Mary went to 3 places and Peter only went to one place. Also from (a), we can find that Mary has larger experienced value than John. In turn, Buddha's interesting value is larger than Peak's.

(c) A place is interesting if it is visited by many experienced travelers.
A traveler is experienced if he/she visited many interesting places

4. [20] The following sentences are extracted from Wikipedia and considered as two documents.

Document 1: Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Document 2: Automated information retrieval systems are used to reduce what has been called "information overload".

After removing all punctuation marks and words with three characters or less:
(i) Show the bigrams of these two documents. No stemming is needed. Suppose an index is built on the bigrams (no need to show the index).

Document 1: information retrieval activity obtaining information resources relevant information need from collection information resources

information retrieval, retrieval activity, activity obtaining, obtaining information, information resources, resources relevant, relevant information, information need, need from, from collection, collection information, information resources

Document 2: Automated information retrieval systems used reduce what been called information overload

automated information, information retrieval, retrieval systems, systems used, used reduce, reduce what, what been, been called, called information, information overload

For the query **information retrieval systems** (note: no double quote applied):
(ii) Explain how you would transform the query before the index is searched.

The query has to be broken into bigrams: information retrieval, retrieval systems

(iii) Assuming that TF is used as term weight (i.e., IDF is not used) what is the similarity of the query to the two documents using inner product similarity?

D1: 1
D2: 2

(iv) Using this example, explain the advantage of bigram compared to unigram (i.e., single term) indexing in terms of search quality (this part is an open discussion; a few sentences are enough).

Bigrams can rank D2 higher D1, which is correct. If unigram is used, the similarity scores will be:
D1: 4 + 1 = 5
D2: 2 + 1 + 1 = 4
D1 will rank higher than D2 even without containing the word “system” because 4 instances of “information” is counted in weighting, of which only one concerns “information retrieval”

(correctly) while the other three are about “information resources” and “information need”. Bigram would not match these latter three instances.

Disadvantage (not required): There are a lot more bigrams than unigrams in a typical document.