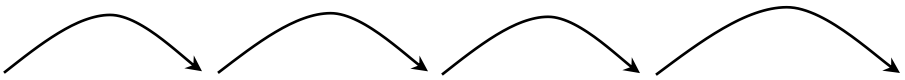
Information Retrieval (CS60092)
Computer Science and Engineering, Indian Institute of Technology Kharagpur

Supplementary End-Semester Examination for Session 2012 - 2013, July 2013

Answer as many questions as you can. *Time: 3 hours*
Use of calculator is allowed. *Maximum Marks: 100*
State any assumptions made clearly.

Q. 1> Consider the following postings lists p1 and p2. p1 has skip pointers. p2 does not have any skip pointer.



p1: 3 5 9 15 24 39 60 68 75 81 84 89 92 96 97 100 115

p2: 3 5 89 95 97 99 100 101

(a) Intersect the postings lists WITHOUT USING the skip pointers. Write down the comparisons (x,y) made while doing the intersection, where x is a docID from p1 and y is a docID from p2. How many comparisons are required?

(b) Intersect the postings lists USING the skip pointers. Write down the comparisons (x,y) made while doing the intersection, where x is a docID from p1 and y is a docID from p2. How many comparisons are required?

(c) Do skip pointers help in processing AND queries? Justify your answer.

(d) Do skip pointers help in processing OR queries? Justify your answer. **[4 + 4 + 1 + 1 = 10]**

Q.2> Suppose that a document collection consists of following two documents:

d1: *free eBooks free software eBooks*

d2: *hundred free pdfs*

User's initial query is **q:** *free eBooks free pdfs free computer eBooks*

The user judges **d1** relevant and **d2** non-relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize

vectors. Using Rocchio relevance feedback, what would the revised query vector be after relevance feedback?

Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$. **[6]**

Q.3> Draw the DOM tree for the following XML document:

```
<db>
 <customer>
   <name>
    <firstname>John</firstname> <lastname>Doe</lastname>
   </name>
   <phone>
    <areacode>512</areacode> <number>471-9558</number>
   </phone>
   <purchases>
    <item>
     <camera>
       <type>Canon digital</type> <price>200</price>
     </camera>
    </item>
   </purchases>
 </customer>
</db>
```

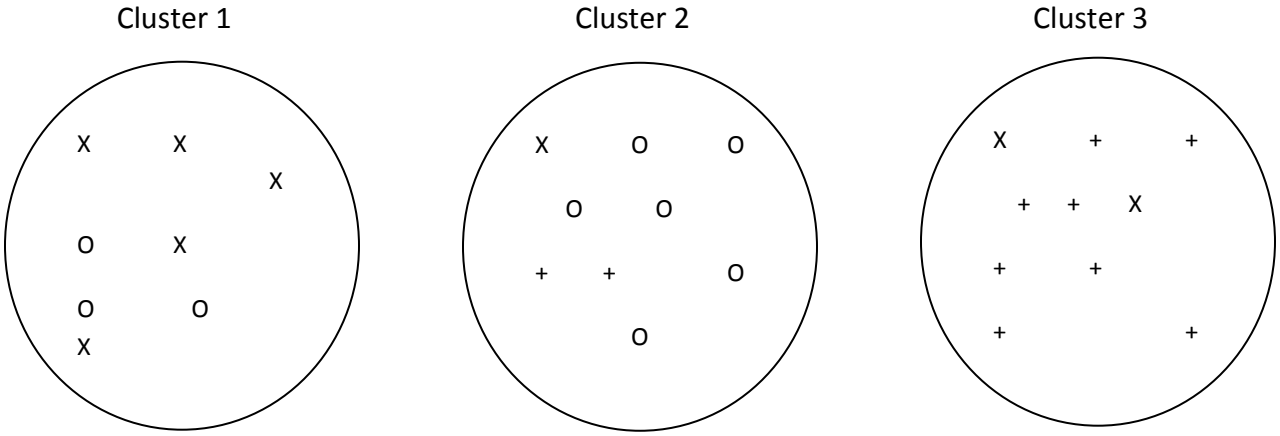
[4]

Q. 4> Consider the following matrix representing **distance** between six documents:

Document	A	B	C	D	E	F
A	0	662	877	255	412	996
B	662	0	295	468	268	400
C	877	295	0	754	564	138
D	255	468	754	0	219	869
E	412	268	564	219	0	669
F	996	400	138	869	669	0

Compute hierarchical single-linkage clustering of these six documents. Clearly show the matrices at each step of building the dendrogram. (No marks will be given for showing only the final dendrogram.) **[10]**

Q. 5 > Consider the following figure for clusters found after performing flat clustering (*k*-Means) on a set of documents. The gold standard for each document is produced by human judges. Each document belongs to one of the three gold standard classes (x, o and +).



Calculate the following quality measures for the above clustering
(a) Purity
(b) NMI
(c) Rand Index
(d) F-Measure **[2 + 4 + 2 + 2 = 10]**

Q. 6> Consider the problem of learning to classify a name as being Food or Beverage. Assume the following training set:

Document	**Class**
Cherry Pie Chocolate	Food
Chicken Wings Crispy	Food
Cream Soda Water	Beverage
Orange Soda	Beverage

(a) Train a Multinomial Naive Bayes Classifier on the above data. Calculate the multinomial parameters (Priors and Conditional Probabilities). Use *Laplace Smoothing* for calculation of conditional probabilities.

(b) What does this classifier predict about the class of the following test document: ***chocolate cream soda***? Assume *positional independence* of terms. **[7 + 3 = 10]**

Q. 7> Consider a document collection that contains the following documents:
d_1: *tick goes the clock goes tick tick tick*
d_2: *tick tock big time*
d_3: *clock tower*
d_4: *big tower of clock*
Let a query be *"clock tick".* Compute the tf-idf scores of each document with respect to this query and provide the resultant document ranking. **[10]**

Q. 8> Consider two queries for which there are 4 and 6 relevant documents in the collection respectively. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1, Query 1: R N R R R N N N N N

System 1, Query 2: R R R R N N N R N R

System 2, Query 1: N N N N R R R N N R

System 2, Query 2: N N N N R R R R R R

(a) What is the MAP of each system? Which system has a higher MAP?

(b) What does the result say about what is important in getting a good MAP score?

(c) How is R-precision of a system defined? What is the R-precision of each system here? Does it rank the systems in the same order as MAP? **[6 + 1 + 3 = 10]**

Q. 9> Consider the following documents:

D1: *english channel atlantic*

D2: *national geography channel english*

D3: *doordarshan national english news*

Using unigram language model, rank the above documents for the query ***national news channel english***. To compute the model probabilities, combine MLE estimates from documents and the collection giving equal importance to both. **[10]**

Q. 10> Consider the query *obama health plan*. The document collection consists of six documents only, which are marked as relevant (R) or non-relevant (NR):

d_1: *president rejects rumors about his own bad health* (NR)
d_2: *the plan is to visit obama* (NR)
d_3: *obama raises concerns with us medical reforms* (R)
d_4: *president states a health vision* (R)
d_5: *romney states a health issue* (NR)
d_6: *obama states a health plan* (R)

Assume a binary independence model (BIM) of retrieval. Rank the documents in descending order of their retrieval status value (RSV). Use contingency tables to show intermediate steps. Do not use any smoothing. The RSV for a BIM model is given by

$$RSV_d = \sum_{t:x_t=\, q_t=1} \log_{10} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

where, for each term *t*, the probabilities of occurrence p_t and u_t can be represented in the form of the following contingency table:

	Document	R	NR
Term present	$x_t = 1$	p_t	u_t
Term absent	$x_t = 0$	$1 - p_t$	$1 - u_t$

[10]

Q. 11> Consider the following term document matrix *C*.

Terms	*D1*	*D2*	*D3*	*D4*	*D5*	*D6*
Ship	1	0	1	0	0	0
Boat	0	1	0	0	0	0
Ocean	1	1	0	0	0	0
Voyage	1	0	0	1	1	0
Trip	0	0	0	1	0	1

(a) Suppose vector space model is used to represent the documents. Vector dimensions are filled with raw frequency counts of the corresponding terms. According to this representation, what is the similarity between the documents D2 and D3?

(b) C is decomposed as $C = U\sum V^T$. The matrices U, ∑ and V are given below.

U =

	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0	0.73
ocean	−0.48	−0.51	−0.37	0	−0.61
voyage	−0.70	0.35	0.15	−0.58	0.16
trip	−0.26	0.65	−0.41	0.58	−0.09

∑ =

2.16	0	0	0	0
0	1.59	0	0	0
0	0	1.28	0	0
0	0	0	1	0
0	0	0	0	0.39

V^T =

	d1	d2	d3	d4	d5	d6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0	0	0.58	0	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

<i> Suppose a low rank approximation of C is obtained as C_2 by keeping the *most* important two terms. According to C_2, what is the cosine similarity between documents D2 and D3?

C_2 =

[0.8511 0.5189 0.2807 0.1272 0.2087 -0.0815
0.3628 0.3567 0.1559 -0.2042 -0.0228 -0.1814
1.0128 0.7201 0.3614 -0.0443 0.1637 -0.2081
0.9726 0.1284 0.1967 1.0310 0.6214 0.4096
0.1215 -0.3905 -0.0840 0.9038 0.4127 0.4911]

<ii> Suppose another low rank approximation of C is obtained as C'_2 by keeping the *least* important two terms. According to C'_2, what is the cosine similarity between documents D2 and D3?

C'_2 =

[0.8511 0.5189 0.2807 0.1272 0.2087 -0.0815

0.3628 0.3567 0.1559 -0.2042 -0.0228 -0.1814
1.0128 0.7201 0.3614 -0.0443 0.1637 -0.2081
0.9726 0.1284 0.1967 1.0310 0.6214 0.4096
0.1215 -0.3905 -0.0840 0.9038 0.4127 0.4911]

(c) Find out the Eigen Values of the matrix CC^T.

[2 + 4 + 4 = 10]
