Information Retrieval (CS60092)
End-semester examination, Autumn 2013 – 2014

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

Q. 1> We wish to apply the naïve Bayes text classification with a multinomial model with the document class information below.

set	**doc-id**	**document text**	**class label**
train	1	*money invest in bank*	finance
	2	*river bank is green and green*	general
	3	*time invest in exam*	general
	4	*money invest bank for interest*	finance
	5	*paper is money is paper*	general
	6	*bank bank money money*	finance
test	7	*invest bank on river bank*	?

Use the stop list: *a, an, the, how, what, why, when, who, in, is, and, for, to, from, by, with, of*.

a. What are the prior probabilities needed to classify the test document?

b. What is the vocabulary size (without the stop words)?

c. What are the conditional probabilities required to classify the test document?

d. What is the class label assigned to the test document? **[2 + 1 + 6 + 1 = 10]**

Q. 2> We wish to use the following queries and corpus for document ranking using the MLE unigram model. Use the stop list: *a, an, the, how, what, why, when, who, in, is, and, for, to, from, by, with, of*.

doc-id	**document text**	**relevance (*q*1)**	**relevance (*q*2)**
1	*jaguar is a fast animal*	R	NR
2	*jaguar is a fast computer*	NR	R
3	*animal family jaguar leopard panther*	R	NR
4	*animal animation with jaguar computer*	NR	R

We have two queries – *q*1: *animal jaguar* and *q*2: *jaguar computer.*

a. Rank the documents with respect to *q*1 and *q*2 using individual document models only.

b. Now rank the documents for both queries mixing document and corpus models with the weight ratio 1:3.

c. Which of the rankings – part a or part b – has higher MAP? Is corpus modelling advantageous in this context? **[4 + 4 + 2 = 10]**

Q. 3> Consider the following points P1 through P13 on a 2-D plane: (2, 4), (2, 5), (3, 3), (4.5, 4.5), (3, 6), (5.5, 4.5), (5.5, 2.5), (8, 0), (10, 0), (12.5, 2.5), (12.5, 4.5), (10, 7), and (8, 7). Note that the order and nomenclature of the points are important, i.e. P1 is (2, 4), P2 is (2, 5), P3 is (3, 3), and so on up to P13. Points P1 through P5 belong to class 1 and P7 through P13 belong to class 2.

a. Classify P6 using Rocchio classification.

b. To which class would 3-NN classify P6?

c. Now assume that it is known that P6 belongs to class 2 while the class information for P7 is missing. Classify P7 using Rocchio. Is P7's class label the same as earlier? **[4 + 3 + 3 = 10]**

You may wish to plot the points on *x*- and *y*-axes to have a better picture of the scenario.

Q. 4> Consider the following points P1 through P9 on a 2-D plane: (1, 3), (2, 3), (1, 2), (2, 2), (1, 1), (2, 1), (5, 3), (5, 2) and (5, 1). Note that the order and nomenclature of the points are important, i.e. P1 is (1, 3), P2 is (2, 3), P3 is (1, 2), and so on up to P9. It is known that points P1 through P6 are red and points P7 through P9 are blue.

a. Break the set of points into two clusters using the *k*-means algorithm, starting at P2 and P6. Stop when there is no change in cluster membership between iterations. In case of tied distances from centroids, select the centroid with lower *y*. If the *y*-coordinates also match, select the centroid with lower *x*.

b. What is the NMI of this clustering output with respect to red and blue colouring? Show all steps of the computation. Use log base 2. **[5 + 5 = 10]**

You may wish to plot the points on *x*- and *y*-axes to have a better picture of the scenario.

P. T. O.

Q. 5> Consider the set of six documents below to be your corpus.

doc-id	**document text**
1	*java coffee is the best coffee for programming*
2	*java is a good island*
3	*java programming with java coffee*
4	*programming with java*
5	*best island in java island*
6	*java programming is the best programming*

Use the stop list: *a, an, the, how, what, why, when, who, in, is, and, for, to, from, by, with, of, good, better, best*.

We use a vector space model. Assume that the index of a term is determined by its order of encounter in the log (*java* gets position 1 in the vector, *coffee* gets position 2, *programming* position 3, and so on). Using **simple TF-IDF** (use raw TF, multiplied by IDF(t) = $\log_{10}(N/\text{DF}(t))$) for **term weighting of documents**. Do **not** apply **any** normalization on the document weights. Use **Euclidean distance** as the distance between two vectors.

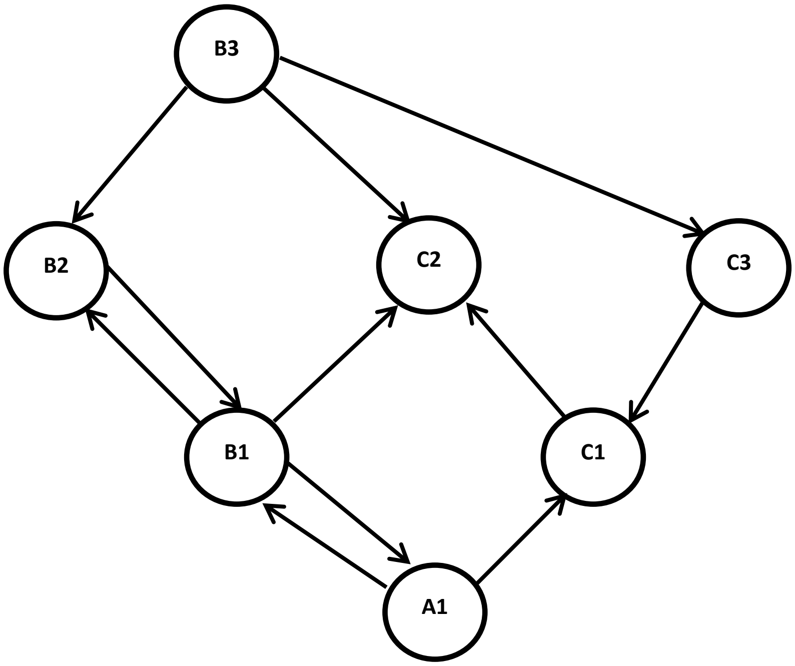
a. Build a doc-doc unnormalized Euclidean distance matrix and apply HAC using the *complete link* concept.

b. Draw the complete resultant dendrogram and mark the combination similarities at each join.

c. Cut the dendrogram where the gap between two successive combination similarities is the largest. How many clusters do you get? List the elements of these clusters. **[7 + 1 + 2 = 10]**

P. T. O.

Q. 6>



Consider the following toy Web graph, which shows webpages as labelled circles and their link structure using directed arrows. There are three domains in this graph: A, B and C. A has one webpage A1, B has three webpages B1, B2 and B3, and C has three webpages C1, C2 and C3. Suppose a crawler takes 100 milliseconds to fetch one webpage. *Politeness policy* states that a crawler cannot issue more than one page request to any domain in 50 milliseconds.

a. Crawler X must finish crawling all pages in a domain before switching to a different domain. What is the shortest time in which X can finish crawling all accessible pages if it observes the *Politeness Policy* and starts crawling on page A1?

b. What is the total crawl time if X starts from B3? What important challenge faced by a crawler is being highlighted in this scenario?

c. What is a common method for modeling the challenge in the previous problem in mathematical formulations of the Web graph?

d. Let *A* is the adjacency matrix of any Web graph (or its subset), where *A* is a square matrix with one row and one column for each page. The entry A_{ij} is one if there is a hyperlink from page *i* to page *j*, and zero otherwise. What would the cells of AA^{T} represent? What about $A^{\mathrm{T}}A$?

[3 + 4 + 1 + 2 = 10]

Q. 7> a. Find the eigenvalues and corresponding eigenvectors for the matrix $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

b. Since *S* is symmetric, what special property does its eigenvectors possess?

c. Compute the eigen decomposition for *S.*

d. Is the decomposition for the given *S* unique? Why or why not? **[3 + 1 + 5 + 1 = 10]**

Q. 8> Multiple choice questions on text classification. Provide your option with a brief reason (<= 3 sentences).

a. Which of the following are linear classifiers?
(A) Naïve Bayes and k-NN (B) Rocchio and k-NN (C) Naïve Bayes and Rocchio (D) All

b. What is the testing time complexity for multinomial Naïve Bayes text classifier (|C| is the number of priors and M_a is the number of types in the test document)?
(A) $\Theta(|C|^2 M_a)$ (B) $\Theta(|C| M_a)$ (C) $\Theta(|C|)$ (D) $\Theta(M_a)$

c. The number of linear separators for two classes is:
(A) Infinite (B) Zero (C) Infinite or zero (D) Depends on the data points

d. Which of the following has optimal running time complexity?
(A) Naïve Bayes and k-NN (B) Rocchio and k-NN (C) Naïve Bayes and Rocchio (D) None

e. Is the Naïve Bayes classifier optimal with respect to error rate on test data?
(A) Never (B) Always (C) Under specific assumptions (D) Depends on the data points

[2 x 5 = 10]

Q. 9> a. What are the two key parameters that determine the rank score of an ad?

b. What are the three steps involved in delivering an ad?

c. What are the three components of an ad?

d. Which location does the user go to upon clicking an ad?

e. Can a given ad have multiple bids by the same advertiser? **[2 + 3 + 3 + 1 + 1 = 10]**

Q. 10> While selecting ads beyond exact string matches what are ten different approaches used to expand/rewrite a query? Give an example for each approach. **[10]**
