National University of Computer and Emerging Sciences, Lahore Campus

Course: Information Retrieval Course Code: CS317 Program: BS(Computer Science) Semester: Fall 2018 Duration: 180 Minutes Total Marks: 38
Paper Date: 19-Dec-18 Weight 50%
Section: A,B Page(s): 10
Exam: Final Roll No:

Instruction/Notes: ***Attempt the examination on this question paper****. You can use extra sheets for rough work but do not attach extra sheets with this paper. Do not fill the table titled Question/marks*

Question	1-4	4-6	7	8-9	10	11-17	18	Total
Marks	/ 9	/ 6	/4	/ 5	/4	/8	/2	**/38**

Q1) a) Suppose we have the following documents: [1 Mark]

Document	Words
D1	a b b a b b c
D2	a a b a b a
D3	b b b b b b c c

How will a positional inverted index for this corpus look like?

i. a => D1 -> D2 -> D3;
b => D1 -> D2 -> D3;
c => D1 -> D2 -> D3
ii. a => D1 -> D1 -> D2 -> D2 -> D2 -> D2;
b => D1 -> D1 -> D1 -> D1 -> D2 -> D2 -> D3 -> D3 -> D3 -> D3 -> D3 -> D3; c => D1 -> D3 -> D3
iii. a => D1 -> D2;
b => D1 -> D2 -> D3;
c => D1 -> D3
iv. a => D1:1,4 -> D2:1,2,4,6;
b => D1: 2,3,5,6 -> D2:3,5 -> D3: 1,2,3,4,5,6;
c => D1: 7 -> D3: 7,8.

b) Let V = Vocabulory size, n= Total number of words, d = total documents, AveD = Average Document Length, |q| = query length, |posting| = length of posting list of a word. [2 Marks]

What is time complexity of building inverted index

i. using HashMaps

ii. without using Hashmaps

Q2) Given the query “apple fruit” and the following term frequencies for the three documents doc1, doc2 and doc3 :

	apple	green	health	benefit	fruit	vitamin
Doc1	3	4	0	6	0	0
Doc2	4	0	4	0	0	3
Doc3	5	3	0	4	4	0

(a) Represent each document in vector space model using tf*idf weights. [3 Marks]

(b) Which of the following will be correct order of documents if we rank them according to tf*idf weights for given query [2 Marks]

i. Doc3, Doc1, Doc2
ii. Doc2, Doc3, Doc1
iii. Doc3, Doc2, Doc1
iv. Doc1, Doc2, Doc3

Q3) If document vectors are length normalized, the ranking according to Euclidean distance and ranking according to cosine of the angle between the document vectors gives the same ranking or different ranking? Justify your answer. [1 Mark]

Q4) In the class, we discussed several ways to evaluate ranking systems, for e.g., precision, recall, NDCG etc. However, for all these metrics, we need the relevance values for the results. Two possible

methods to collect relevance values are: a) click feedback from users, b) expert judgements. Write one advantage and one disadvantage of each of these methods. [2 Marks]

a) Click feedback:

b) Expert judgments:

Q5) Instead of using recall/precision for evaluating IR systems, we could use accuracy of classification. Consider a classifier (non-ranking IR system) that classifies documents as being either relevant or non relevant. The accuracy of a classifier that makes *c* correct decisions and *i* incorrect decisions is defined as: $c/(c+i)$. Why do the recall and precision measures reflect the utility (i.e., quality or usefulness) of an IR system better than accuracy does? Give example. [2 Marks]

Q6) Suppose that we have a collection of 10 documents, and two different Boolean retrieval systems A and B. Give an example of two result sets, A*q* and B*q*, assumed to have been returned by the system in response to a query *q*, constructed such that A*q* has clearly higher utility and a better score for precision than B*q*, but such that A*q* and B*q* have the same scores on accuracy. [2 Marks]

Q7)

DocID	**Document Text**
1	exam apple fruit sugar computer exam exam exam
2	exam exam
3	metal here
4	Metal fruit exam here

Use language model (with Dirichlet smoothing with mu = 3) to calculate the probabilities of the queries "exam", "fruit", and hence "exam fruit" according to each document, and use those probabilities to rank the documents returned by each query. Fill in these probabilities in the below table: **[4 Marks]**

	Doc 1	**Doc 2**	**Doc 3**	**Doc 4**
exam				
fruit				
exam fruit				

Q8) Let D be a document in a text collection. Suppose we add a copy of D to the collection. How would this affect the IDF values of all the words in the collection? Why? [2 Marks]

Q9) The goal of a retrieval model is to score and rank documents for a query. Different retrieval models make different assumptions about what makes a document more (or less) relevant than another. Suppose you issue the query "lemur" to a search engine. And, suppose that documents D101 and D123 both contain the term "lemur" twice. Answer the following questions. [3 Marks]

a) Would the ranked Boolean retrieval model necessarily give both documents the same score? If not, what information would determine which document is scored higher?

b) Would the cosine similarity necessarily give both documents the same score? If not, what would determine which document is scored higher?

c) Would the query-likelihood model (without linear interpolation) necessarily give both documents the same score? If not, what would determine which document is scored higher?

Q10) a) Compute page rank of all nodes of following graph. Teleportation probability = 0.2. Perform only 1 iterations of page rank algorithm. [3 Marks]

b) Why do we need teleportation in page rank algorithm? [1Mark]

Q11) Suppose a search returns documents D1, D2, and D3 in this order. The correct results in the system would have been D2, D1, D4, and D5 in this order. Which are the precision and recall for the engine in this case? [2 Marks]

a) P = 0.67; R = 0.5
b) P = 0.5; R = 0.67
c) P = 0.67; R = 0.4
d) P = 0.4; R = 0.67

Q12) Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means [1 Mark]

A. Yes

B. No

C. Can't say

D. None of these

Q13) Which of the following metrics, do we have for finding similarity between two clusters in hierarchical clustering? [1 Mark]

1. Single-link
2. Complete-link
3. Average-link

Options:

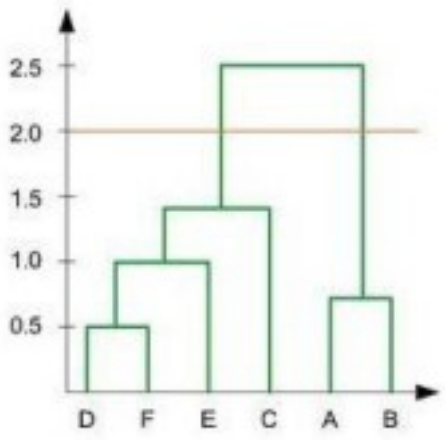
A. 1 and 2

B. 1 and 3

C. 2 and 3

D. 1, 2 and 3

Q14) In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed? [1 Mark]



A. 1

B. 2

C. 3

D. 4

Q15) Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: [1 Mark]

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

What will be the cluster centroids if you want to proceed for second iteration?

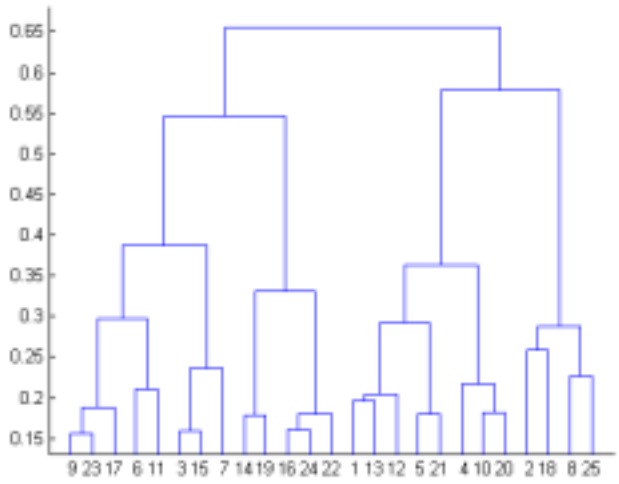
A. C1: (4,4), C2: (2,2), C3: (7,7)

B. C1: (6,6), C2: (4,4), C3: (9,9)

C. C1: (2,2), C2: (0,0), C3: (5,5)

D. None of these

Q16) After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram? [1 Mark]



A. There were 28 data points in clustering analysis

B. The best no. of clusters for the analyzed data points is 4

C. The proximity function used is Average-link clustering

D. The above dendrogram interpretation is not possible for K-Means clustering analysis

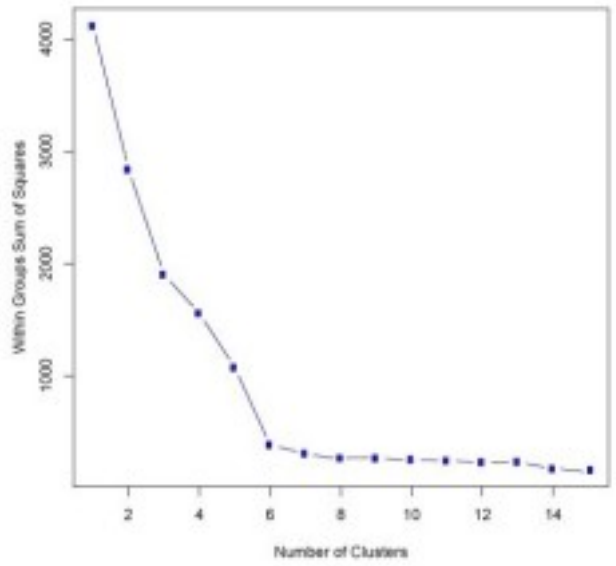
Q17) What should be the best choice for number of clusters based on the following results: The value on Y-axis is RSS. [1 Mark]

A. 5

B. 6

C. 14

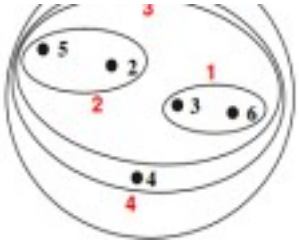
D. Greater than 14

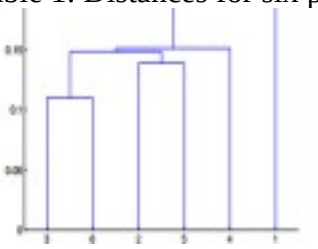


Q18) Given the following **distance matrix** between 6 data points, which of the following clustering representations and dendrogram depicts the use of **Complete link** similarity function in hierarchical clustering: [2 Marks]

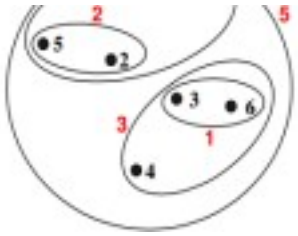
	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1	0					
P2	0.235	0				
P3	0.22	0.148	0			
P4	0.36	0.204	0.153	0		
P5	0.34	0.138	0.284	0.293	0	
P6	0.234	0.254	0.11	0.22	0.39	0

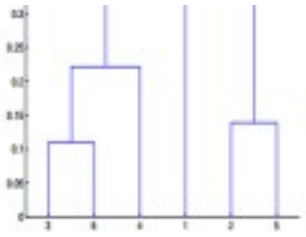
Table 1: Distances for six points



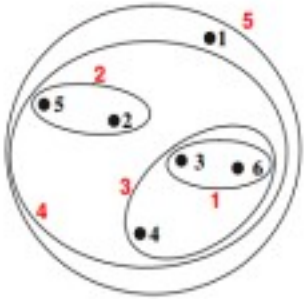


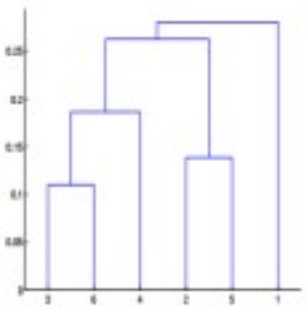
A



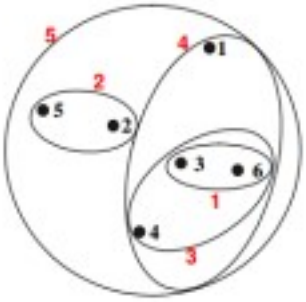


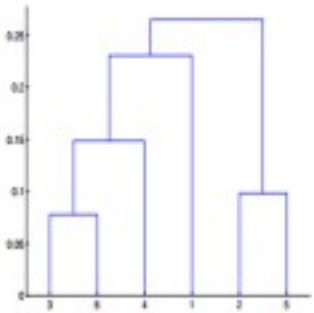
B





C





D