National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Information Retrieval and Text Mining**	**Course Code:**	**CS567**
Program:	**MS(Computer Science)**	**Semester:**	**Fall 2018**
Duration:	**180 Minutes**	**Total Marks:**	**36**
Paper Date:	**18-Dec-18**	**Weight**	**45%**
Section:	**CS**	**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*

Q1) 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]

	apple	green	health	benefit	fruit	vitamin
Idf	1	1.24	1.48	1.24	1.48	1.48
Doc1	1.47	1.6*	1.48	1.77*	1.48	1.48
Doc2	0.6	1.24	1.6*	1.24	1.48	1.47*
Doc3	0.7	1.47*	1.48	1.6*	1.6*	1.48

	apple	green	health	benefit	fruit	vitamin
Idf	1	1.24	1.48	1.24	1.48	1.48
Doc1	1.47	1.98	2.2	2.2	2.2	2.2
Doc2	1.6	1.53	2.4	1.53	2.2	2.2
Doc3	1.7	1.82	2.2	1.98	2.4	2.2

	apple	green	health	benefit	fruit	vitamin
Idf	1	1.58	2.58	1.58	2.58	2.58
Doc1	1.58	3.16	0	5.16	0	0
Doc2	2	0	5.16	0	0	4.1
Doc3	2.3	2.5	0	3.16	5.16	0

(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

Q2) The Rocchio algorithm is a classic algorithm for implementing relevance feedback. Use Rocchio to compute the new query vector for "apple fruit" using doc3 for relevance feedback (i.e., doc3 has been marked as relevant).(alpha = 1, Beta = 0.8, Gamma = 0). [2 Marks]

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) Indexing New York Times newswire from 1991–1995 reveals that it contains about 400 million word tokens, and a lexicon (vocabulary) of size about 1 million (given certain fixed decisions on term

normalization, lowercasing, treatment of numbers etc.). What would be a good prediction of how many word tokens and what lexicon size one would get in indexing New York Times newswire from 1991–2000? (Hint: Use Heap's Law) [4 Marks]

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)

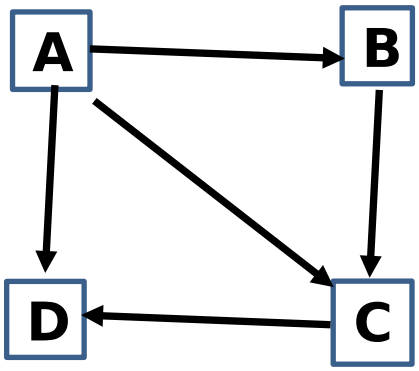
DocID	Document Text
1	exam apple fruit sugar computer exam exam exam
2	exam exam
3	metal here
4	Metal fruit exam here

Use language model (with Jelinek Mercer smoothing with lambda = 0.7) 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]**

Solution

	Doc 1	**Doc 2**	**Doc 3**	**Doc 4**
exam	0.7*(4/8) + 0.3*(7/16) =0.48	0.7*(2/2) + 0.3*(7/16) = 0.83	0.7*(0/2) + 0.3*(7/16) = 0.13	0.7*(1/4)+ 0.3*(7/16) = 0.31
fruit	0.7*(1/8) + 0.3*(2/16) = 0.124	0.7*(0/2) + 0.3*(2/16) =0.037	0.7*(0/2) + 0.3*(2/16) = 0.037	0.7*(1/4)+ 0.3*(2/16) = 0.212
exam fruit	0.059	0.031	=0.0048	0.066

Q8) Compute page rank of all nodes of following graph. Teleportation probability = 0.2. Perform only 3 iterations of page rank algorithm. [5 Marks]



Fill in the page rank values of each page in following table.

Solution

	A	B	C	D
Initial	0.25	0.25	0.25	0.25
Iteration 1	0.1	0.167	0.37	0.37
Iteration 2	0.12	0.15	0.28	0.45
Iteration 3	0.14	0.17	0.29	0.39

Q9) A crawler gathers documents and sends them to an indexer, which employs the following modules: [1 Mark]
(A) a stemmer;
(B) a language detector to detect the language of each document;
(C) a stop-word eliminator
(D) a filter that detects the format (pdf, Word, etc.) of the document.

Which of the following is correct sequence in which the indexer should apply these modules to a document:

i. D C B A
ii. C B A D
iii. B A C A
iv. D B C A

Q10) Mark these statements **true/false**: [2 Marks]

A. Stemming increases the size of the lexicon.
B. Stemming should be invoked at indexing time but not while doing a query
C. Stemming increases recall
D. Smoothing is necessary because otherwise the model would assign a zero probability to queries that contain terms not present in the original document (from which the model was built).

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:

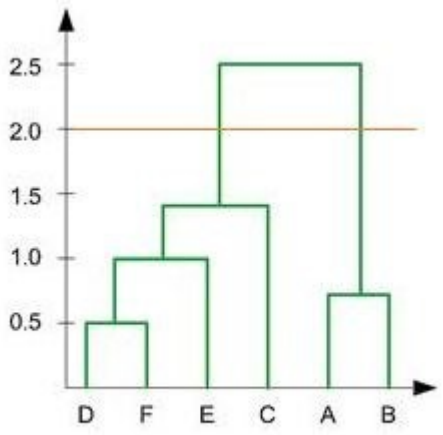
A. 1 and 2

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?

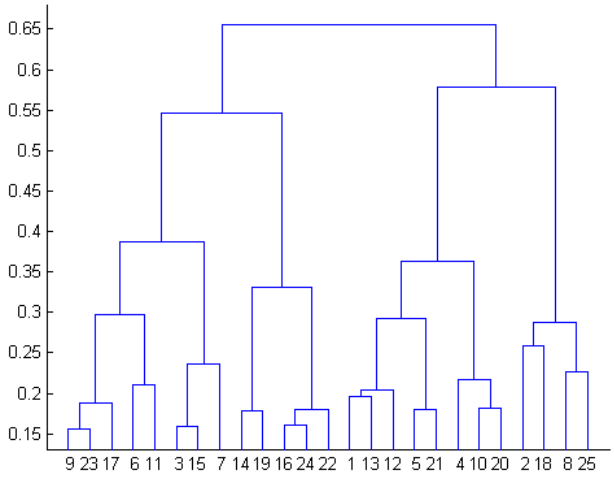
A. C1: (4,4), C2: (2,2), C3: (7,7)

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]



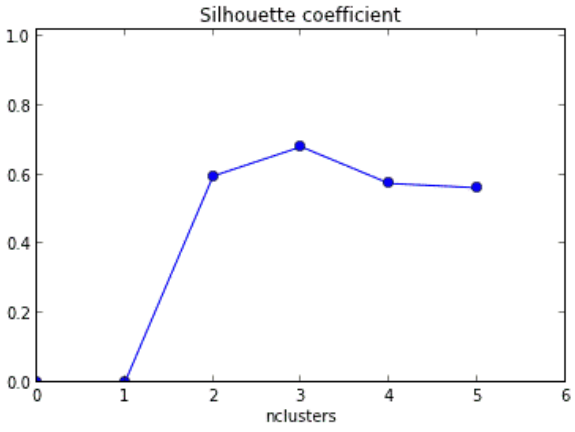
A. There were 28 data points in clustering analysis

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

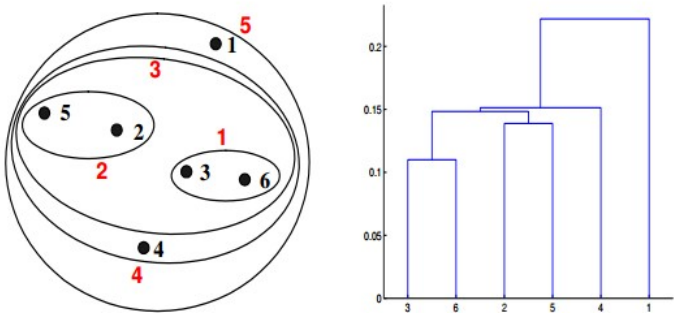
Q17) The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. What should be the best choice of no. of clusters based on the following results: [1 Mark]



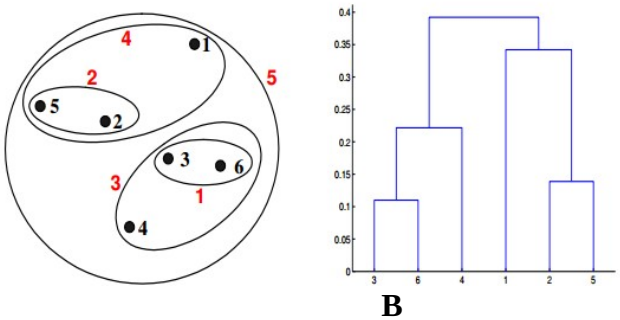
Q18) Given the following **distance matrix** between 6 data points, which of the following clustering representations and dendrogram depicts the use of **Single link** similarity function in hierarchical clustering: [3 Marks]

	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

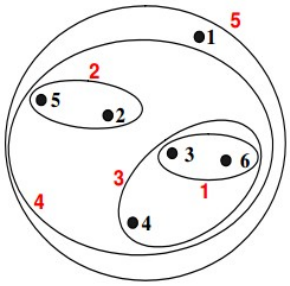
Table 1: Distances for six points

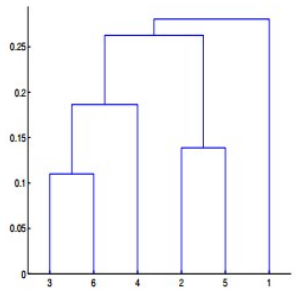


A

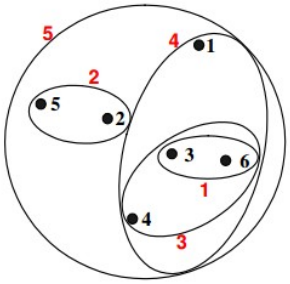


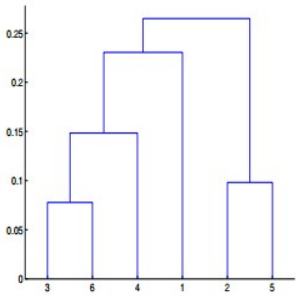
B


5
1
2
5
2
3
6
3
1
4
4


0.25
0.2
0.15
0.1
0.05
0
3
6
4
2
5
1

C


5
4
1
2
5
2
3
6
1
4
3


0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

D