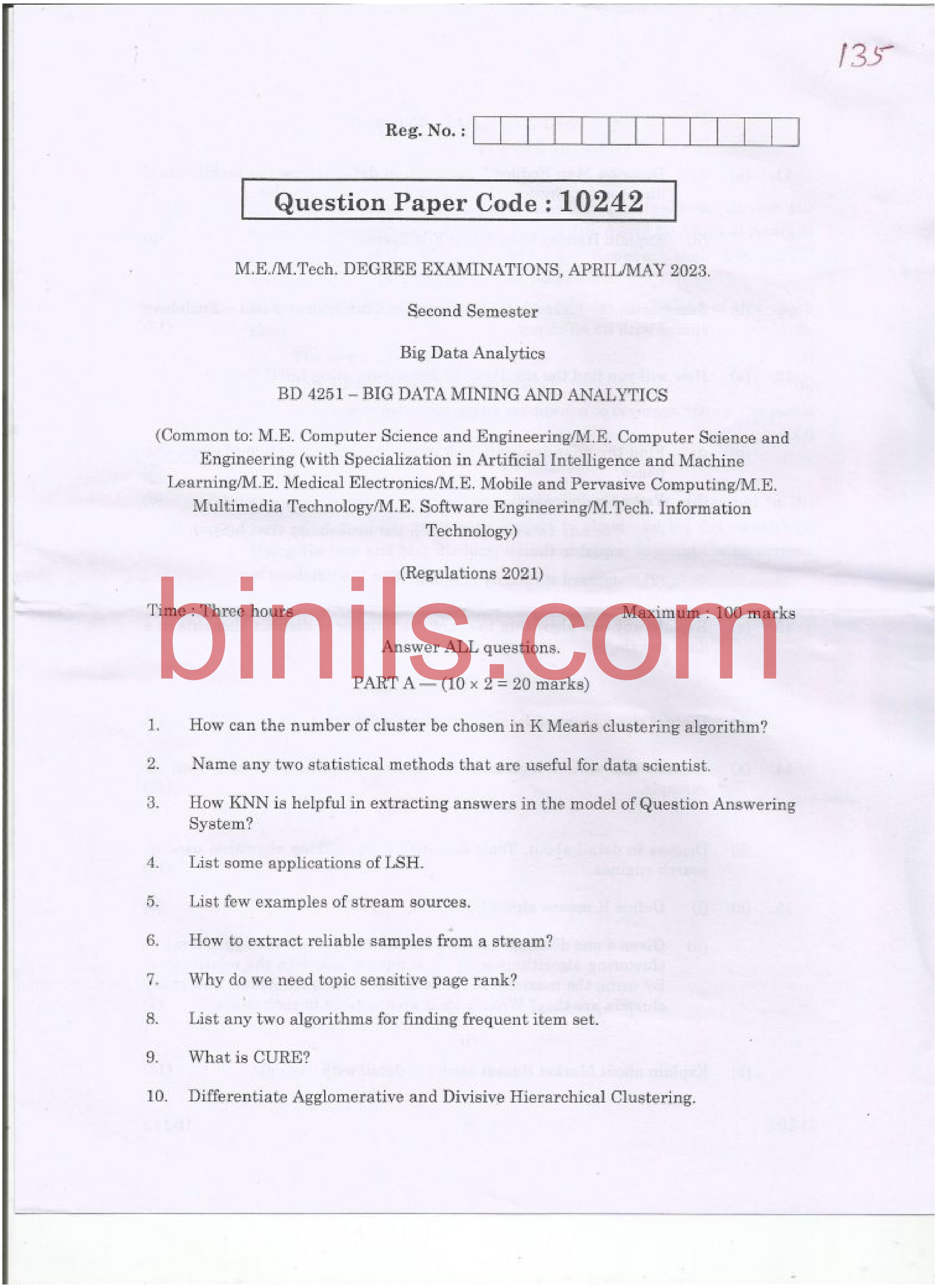


135

Reg. No. : □□□□□□□□□□□□

Question Paper Code : 10242

M.E./M.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2023.

Second Semester

Big Data Analytics

BD 4251 – BIG DATA MINING AND ANALYTICS

(Common to: M.E. Computer Science and Engineering/M.E. Computer Science and Engineering (with Specialization in Artificial Intelligence and Machine Learning/M.E. Medical Electronics/M.E. Mobile and Pervasive Computing/M.E. Multimedia Technology/M.E. Software Engineering/M.Tech. Information Technology)

(Regulations 2021)

Time : Three hours | Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. How can the number of cluster be chosen in K Means clustering algorithm?

2. Name any two statistical methods that are useful for data scientist.

3. How KNN is helpful in extracting answers in the model of Question Answering System?

4. List some applications of LSH.

5. List few examples of stream sources.

6. How to extract reliable samples from a stream?

7. Why do we need topic sensitive page rank?

8. List any two algorithms for finding frequent item set.

9. What is CURE?

10. Differentiate Agglomerative and Divisive Hierarchical Clustering.

135

PART B — (5 × 13 = 65 marks)

11. (a) (i) Describe Map Reduce framework in detail. Draw the architectural diagram for physical Organization of computer nodes. (7)

(ii) Explain Hadoop Distributed File System. (6)

Or

(b) Summarize the hierarchical clustering in Euclidean and non – Euclidean spaces with its efficiency. (13)

12. (a) How will you find the similarity of documents using LSH? Explain. (13)

Or

(b) (i) Find the near duplicate web pages from the collection of given web pages? (5)

(ii) Prove the following: (8)

(1) For any two sets Si and Sj, the probability that h(Sj)=h(Sj) is equal to the.

(2) Jaccard similarity of Si and Sj.

13. (a) Explain any one algorithm to count the number of distinct elements in a data stream. (13)

Or

(b) Discuss the concept of decaying window in detail. (13)

14. (a) Explain the apriori algorithm for mining frequent item sets with an example. (13)

Or

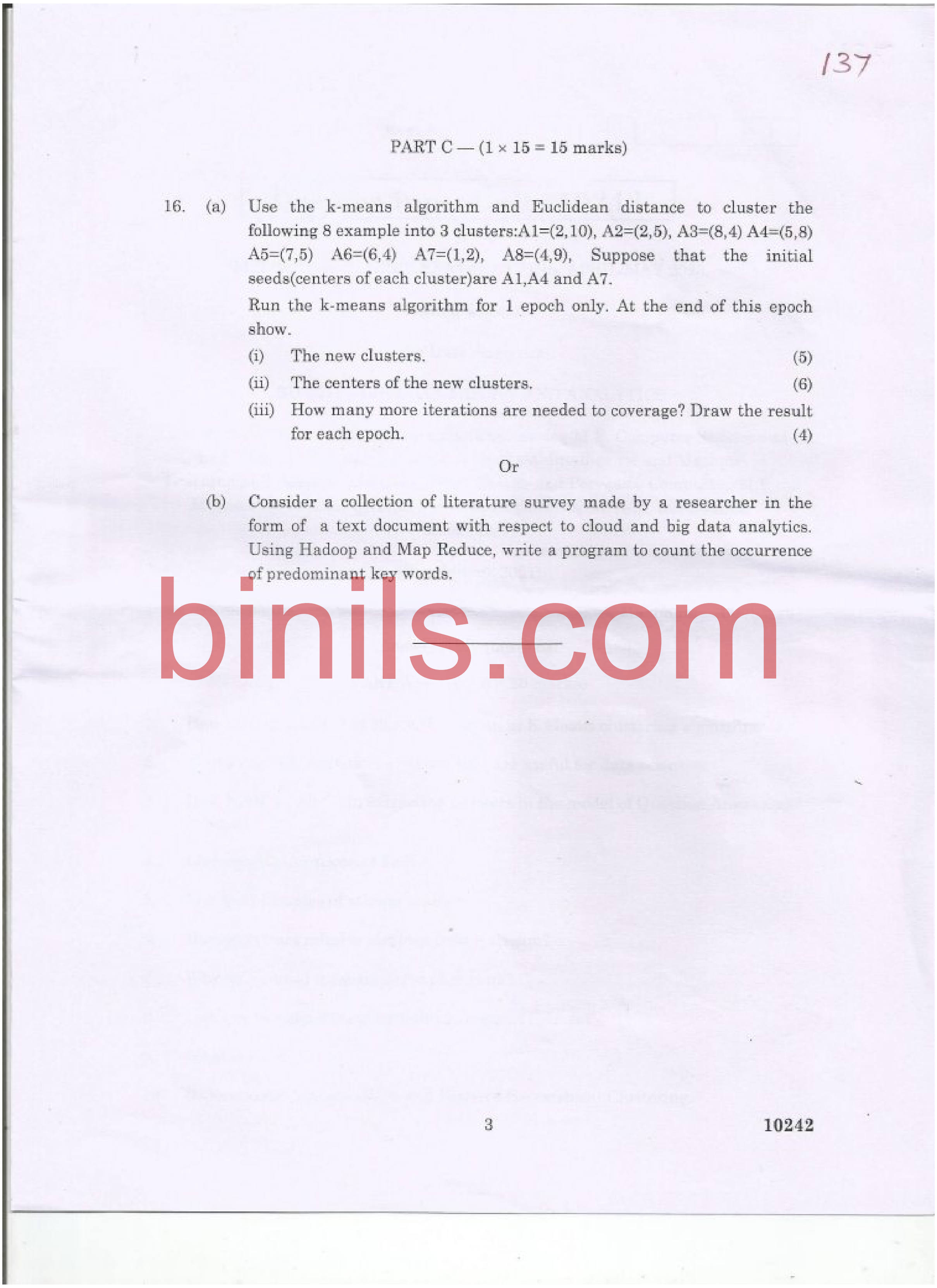
(b) Discuss in detail about. Topic Sensitive page ranking algorithm used by search engines. (13)

15. (a) (i) Define K means algorithm. (6)

(ii) Given a one dimensional data set {1,5,8,1,0,2} use the agglomerative clustering algorithms with the complete link with the relationship. By using the maximal life time as the cutting threshold, how many clusters are they? What is their membership in each cluster? (7)

Or

(b) Explain about Market Basket Model in detail with example. (13)



137

PART C — (1 × 15 = 15 marks)

16. (a) Use the k-means algorithm and Euclidean distance to cluster the following 8 example into 3 clusters:A1=(2,10), A2=(2,5), A3=(8,4) A4=(5,8) A5=(7,5) A6=(6,4) A7=(1,2), A8=(4,9), Suppose that the initial seeds(centers of each cluster)are A1,A4 and A7.

Run the k-means algorithm for 1 epoch only. At the end of this epoch show.

(i) The new clusters. (5)

(ii) The centers of the new clusters. (6)

(iii) How many more iterations are needed to coverage? Draw the result for each epoch. (4)

Or

(b) Consider a collection of literature survey made by a researcher in the form of a text document with respect to cloud and big data analytics. Using Hadoop and Map Reduce, write a program to count the occurrence of predominant key words.