OBE

REFNO038739

MCA SEMESTER – IV EXAMINATION 2020-21

COMPUTER APPLICATIONS

CS - 320T : Text Analytics

Time : 4.30 hours **Max. Marks : 70**

Instructions

1. **The Question Paper contains 08 questions out of which you are required to answer any 04 questions. The question paper is of 70 marks with each question carrying 17.5 marks.**

   प्रश्नपत्र में आठ प्रश्न पूँछे गये हैं जिनमें से 4 प्रश्नों का उत्तर देना है। प्रश्नपत्र 70 अंकों का है, जिसमें प्रत्येक प्रश्न 17.5 अंक का है।

2. The total duration of the examination will be **4.30 hours** (Four hours and thirty minutes), which includes the time for downloading the question paper from the Portal, writing the answers by hand and uploading the hand-written answer sheets on the portal.

   परीक्षा का कुल समय 4.30 घंटे का है जिसमें प्रश्नपत्र को पोर्टल से डाउनलोड करके पुनः हस्तलिखित प्रश्नों का उत्तर पोर्टल पर अपलोड करना है।

3. For the students with benchmark disability as per Persons with Disability Act, the total duration of examination shall be **6 hours** (six hours) to complete the examination process, which includes the time for downloading the question paper from the Portal, writing the answers by hand and uploading the hand-written answer sheets on the portal. .

   दिब्यांग छात्रों के लिये परीक्षा का समय 6 घंटे निर्धारित हैं जिसमें प्रश्नपत्र को पोर्टल से डाउनलोड करना एवं हस्तलिखित उत्तर को पोर्टल पर अपलोड करना है।

4. Answers should be hand-written on a plain white A4 size paper using black or blue pen. Each question can be answered in upto 350 words on 3 (Three) plain A4 size paper (only one side is to be used).

   हस्तलिखित प्रश्नों का उत्तर सादे सफेद A4 साइज के पन्ने पर काले अथवा नीले कलम से लिखा होना चाहिये। प्रत्येक प्रश्न का उत्तर 350 शब्दों तक तीन सादे पृष्ठ A4 साइज में होना चाहिये। प्रश्नों के उत्तर के लिए केवल एक तरफ के पृष्ठ का ही उपयोग किया जाना चाहिए।

5. Answers to each question should start from a fresh page. All pages are required to be numbered. You should write your Course Name, Semester, Examination Roll Number, Paper Code, Paper title, Date and Time of Examination on the first sheet used for answers.

   प्रत्येक प्रश्न का उत्तर नये पृष्ठ से शुरू करना है। सभी पृष्ठों को पृष्ठांकित करना है। छात्र को प्रथम पृष्ठ पर प्रश्नपत्र का विषय, सेमेस्टर, परीक्षा अनुक्रमांक, प्रश्नपत्र कोड, प्रश्नपत्र का शीर्षक, दिनांक एवं समय लिखना है।

Questions

P.T.O.

Questions

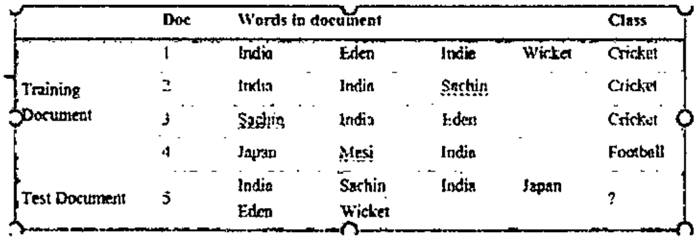
1. What is Term Document Matrix (TDM)? What is the rationale behind length normalization of document vectors and how is it done? Construct the TDM for the following three document collection:

   D1: "Shipment of Gold damaged in a fire"
   D2: "Delivery of silver arrived in a silver truck"
   D3: "Shipment of Gold arrived in silver truck"

2. (a) Distinguish between Boolean and Vector Space representation of documents. What are advantages of Vector Space representation over Boolean representation?

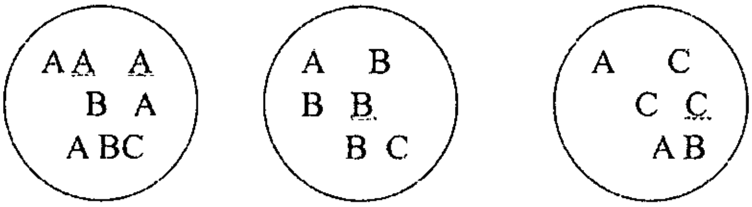
   (b)Explain the notion of Document Similarity. Distinguish between Syntactic and Semantic Similarity. How is Cosine Similarity computed? What features make it a popularly used document similarity measure?

3. Explain the operation of Naïve Bayes classifier by computing the probability of the test document to belongto both target classes.



	Doc	Words in document				Class
Training Document	1	India	Eden	Indie	Wicket	Cricket
	2	India	India	Sachin		Cricket
	3	Sachin	India	Eden		Cricket
	4	Japan	Mesi	India		Football
Test Document	5	India Eden	Sachin Wicket	India	Japan	?

4. What is the main difference between Document Classification and Clustering? What internal measure is used to measuring Clustering quality? Show with a case that it is not the best metric for measuring clustering quality. Compute Purity of the following clustering result:



Cluster I Cluster II Cluster III

5. What do you understand by Sentiment Analysis? Explain the difference between sentiment and opinion. What are different levels of sentiment analysis? How can sentiment analysis be done using Machine Learning Classifiers?

P.T.O.

6. Explain the following in brief:
   a. Stop Words
   b. Stemming and Lemmatization
   c. Smoothing in Sequence Labelling Problem

7. (a) Explain the process of generation of extractive summaries? What is the role of context in automatic summary generation? Explain.

   (b) Explain the Trigram Hidden Markov Model for Named Entity Recognition. How are parameters estimated for this model using training data?

8. Write short notes on the following:
   a. Structured vs. Unstructured Data
   b. K-Means clustering algorithm
   c. POS Tagging

*****