Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
AIML 2021-22
Comprehensive Examination
(Regular)

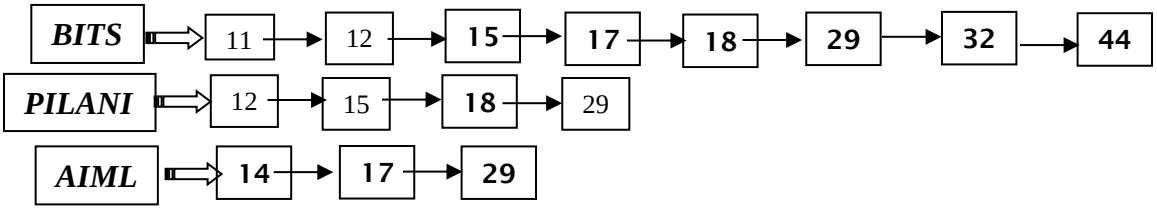
Course No. : PCAM ZC231
Course Title : Text Mining
Nature of Exam : Open Book
Weightage : 40
Duration : 2 Hours and 15 minutes

No. of Pages = 2
No. of Questions = 5

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. Document vectorization – M1 **[6+2=8 Marks]**

A. Following is the inverted index for three words.



(i) How query optimization is implemented for the query "*BITS* AND *PILANI* AND NOT *AIML*"? What Documents will be returned as output from the given set of documents for the above query? **[2 + 1 = 3 Marks]**

(ii) What is best order of query processing for the query "*BITS* AND *PILANI* AND *AIML*"? What Documents will be returned as output from the given set of documents for the query. **[2 + 1 = 3 Marks]**

B. How does the inverted index handle variant forms of the same word like USA, U.S.A, usa etc. [**2 Marks]**

Q2. [6+2=8 Marks]

A. What is the tagging of the following sentence? **[6 Marks]**

"Robots process programs automatically."

(Part of) lexicon:

Robots	N	0.123
process	N	0.1
process	V	0.2
programs	N	0.11
programs	V	0.15
automatically	Adv	0.789

(Part of) transitions:

P (N|V) =0.5 P (N|Adv) =0.12
P (V|Adv) =0.05 P (V|N) =0.4
P (Adv|N) =0.01 P (Adv|V) =0.13
P (N|N) =0.6 P (V|V) =0.05

B. Is POS tagging, a sequence classification problem? How is hidden markov model suitable for POS tagging problem. **[2 Marks]**

Q3. [4+4= 8 Marks]

A. Explain graphical representation of latent dirichlet allocation for three different topics T1, T2 and T3 with distribution of 4 words: w1, w2, w3 and w4 in each of the topics with some probability. **[4 Marks]**

B. What role does Dirichlet distribution play in Latent Dirichlet Allocation? **[4 Marks]**

Q4. [2+3+3= 8 Marks]

A. Suppose the mean rating of books is 2.4 stars. Alice, a faithful customer, has rated 350 books and her average rating is 0.7 stars higher than average users' ratings. Animals Farm, is a book title in the bookstore with 250,000 ratings whose average rating is 0.9 higher than global average. What would be a baseline estimate of Alice's rating for Animals Farms?[2marks]

B. How does Wordnet help in sentiment analysis? **[3 Marks]**

C. Calculate Polarity (battery life) given following 3 reviews using pointwise mutual information. **[3 marks]**

Reviewer1: "bad battery life and bad camera"
Reviewer2: "long battery life"
Reviewer3: "long battery life but bad camera"

Pointwise mutual information: How much more do events x and y co-occur than if they were independent?
If two words are statistically independent, PMI=0
If two words tend to not at all co-occur, PMI is negative
If two words tend to co-occur, PMI is positive
Does phrase appear more with "poor" or "excellent"?

Polarity(battery life**) = PMI(**battery life**, "long")− PMI(**battery life**, "bad")**

Q5. [5+3=8 Marks]

A. Explain with an example, Use of latent factor models in finding missing values in recommendation systems?. **[3 Marks]**

B. Find out which users have similar interests using cosine similarity for given book ratings in the table below. **[5 Marks]**

	Will	Aman	Anna
Book 1		5	4
Book 2	2		3
Book 3		2	1
Book 4	2	4	4