

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATIONS- 2026

M.Tech-II Semester (DS)

COURSE CODE (CREDITS): 22M1WCI236 (3)

MAX MARKS: 35

COURSE NAME: Natural Language Processing

COURSE INSTRUCTOR: HRI

MAX. TIME: 2 Hours

Note: (a) All questions are compulsory.

(b) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

(c) Use of calculator is allowed

Q.No	Question	CO	Marks
Q1	Given that the 10th most frequent word in a corpus (which closely follows Zipf's Law) has a probability of occurrence of 0.012, what is the frequency of the most frequent word if the total size of the corpus is 10,000 words?	[CO1]	4
Q2	Draw a complete Edit Distance (Levenshtein Distance) matrix between the strings S1= Thin and S2= Thick, assuming a cost of 1 for insertion, deletion, and substitution. What are the minimum cost and the operation?	[CO1]	4
Q3	Describe the following for finding the similarity on the basis of Wordnet (a) Resnik Similarity (b) Lin Similarity	[CO3]	2+2
Q4	Suppose you are using Gibbs sampling to estimate the distributions, θ and β for topic models. The underlying corpus has 3 documents and 5 words, {machine, learning, language, nature, vision} and the number of topics is 2. At certain point, the structure of the documents looks like the following $\beta_i^{(j)} = \frac{c_{ij}^{WT} + \eta}{\sum_{k=1}^T c_{kj}^{WT} + W\eta}$ $\theta_j^{(d)} = \frac{c_{dj}^{DT} + \alpha}{\sum_{k=1}^T c_{dk}^{DT} + T\alpha}$ Doc1: nature(1) language(1) vision(1) language(1) nature(1) nature(1) language(1) vision(1) Doc2: nature(1) language(1) language(2) machine(2) vision(1) learning(2) language(1) nature(1) Doc3: machine(2) language(2) learning(2) language(2) machine(2) machine(2) learning(2) language(2) (number) –number inside the brackets denote the topic no. 1 and 2 denote whether the word is currently assigned to topics t1 and t2 respectively. $\eta = 0.3$ and $\alpha = 0.3$ (i) Compute "$\beta(2)\text{nature}$" (ii) Estimated value of $\theta_{t1}\text{doc2}$	[CO3]	3+3

Q5	"Python" appears in 1200 Wikipedia articles. Anchor breakdown: <table><tr><th>Candidate</th><th>Anchor count</th></tr><tr><td>Python (programming)</td><td>420</td></tr><tr><td>Pythonidae (snake)</td><td>190</td></tr><tr><td>Monty Python</td><td>105</td></tr><tr><td>Python (mythology)</td><td>35</td></tr><tr><td>No link (plain text)</td><td>450</td></tr></table> (i) What is the keyphraseness of "Python"? (ii) What is the commonness of (Python, Monty Python)?	Candidate	Anchor count	Python (programming)	420	Pythonidae (snake)	190	Monty Python	105	Python (mythology)	35	No link (plain text)	450	[CO4]	6
Candidate	Anchor count														
Python (programming)	420														
Pythonidae (snake)	190														
Monty Python	105														
Python (mythology)	35														
No link (plain text)	450														
Q6	Consider the system-generated summary (S) and the reference summary (R) as follows: S: Renewable energy sources like solar and wind power are rapidly replacing fossil fuels across many countries. R: Solar and wind energy are increasingly adopted worldwide as sustainable alternatives to traditional fossil fuels. (i) What is the ROUGE-1 recall for the given summary with respect to the reference? (ii) What is the ROUGE-2 recall for the given summary with respect to the reference?	[CO4]	6												
Q7	Which of the following words best represent high arousal states in affective lexicons? Justify. 1. Tired, sluggish 2. Peaceful, calm 3. Excited, alert 4. Sad, disappointed	[CO4]	5												