Question Paper

Exam Date & Time: 30-Apr-2024 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
Second Semester Master of Engineering - ME (Artificial Intelligence and Machine Learning) Degree Examination - April / May 2024

Natural Language Processing Principles and Applications (Elective -II) [AML 5233]

Marks: 100 **Duration: 180 mins.**

Tuesday, April 30, 2024

Answer all the questions.

1)	Define the term Linguistics in Natural Language Processing and list the areas of study under linguistics. (CO1)(BL4)	(10)
2)	What is Language Semantics in Natural Language Processing? Distinguish the following in NLP applications with relevant examples: 1. Homonyms 2. Homographs 3. Homophones 4. Heteronyms 5. Heterographs. (CO2)(BL4)	(10)
3)	Write regular expressions for the following languages. "Word", here mean an alphabetic string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth. (2+2+2+4 Marks) a. the set of all alphabetic strings; b. the set of all lower-case alphabetic strings ending in a b; c. the set of all strings from the alphabet a;b such that each 'a' is immediately preceded by and immediately followed by 'a b'; d. to find cases of the English article 'an' without missing it anywhere in the sentence, with a word boundary on both sides (so not to pick ant or panther), also before the article we require either the beginning-of-line or a non-alphabetic character, and the same at the end of the line: (CO2)(BL3)	(10)
4)	List any five main components that go into logical representations and syntax for First Order Logic. (5 Marks) (CO3)(BL3) Write the FOL representation for the Natural Language Statement: (5 Marks) "All black widow spiders are poisonous". Write the Natural Language Statement for the FOL representation: (∃ x player(x) ∧ score(x, goal) ∧ (∀ y score(y, goal) → x=y))	(10)
5)	Show the minimum edit distance between two strings with illustrative steps and compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 1) of "algorithm" to "altruistic". Show your work (using the	(10)

edit distance grid). (7 Marks)

Identify whether 'drive' is closer to *brief* or to *divers* and what is the edit distance to each. (Insertion cost 1, deletion cost 1, substitution cost 1). (3 Marks) (CO3)(BL3)

6) What is n-gram? Describe the various applications of using n-grams in NLP with an example. (5 Marks) (CO3)(BL3) (10)

Identify all Bi-grams. Estimate the Bi-gram probability. What is the most probable next word predicted by the model for the following word sequences? (5 Marks)

<s> Sam ?

<s>do I like ?

Consider the following training corpus for your calculation.

<s>I am Sam</s>

<s>Sam I am</s>

<s>Sam I like</s>

<s>Sam I do like</s>

<s>do I like Sam</s>

7) Classify the different relations the words can have in Lexical Semantics in the context of NLP with suitable examples. (CO4)(BL4) (10)

8) (a) In a corpus of 10000 documents you randomly pick a document, say D, which has a total of 250 words and the word 'data' occurs 20 times. Also, the word 'data' occurs in 2500 (out of 10000) documents. (10)
What will be the tf-idf entry for the term 'data' in a bag of words vector representation for D. (5 Marks)

(b) You have the following three documents - D1, D2, D3:
D1: A simplified model of the human neuron as a kind of computing element that could be described in terms of propositional logic.
D2: A modern neural network is a network of small computing units, each of which takes a vector of input values.
D3: The use of modern neural nets is often called deep learning because modern networks are often deep.
Answer the following with respect to the above set of 3 documents. What are the number of bi-grams and tri-grams in D1? (5 Marks) (CO4)(BL4)

9) Consider that our document collection S has the following documents: D1, ..., D5: (10)

document	words
D1:	Data Base System Concepts
D2:	Introduction to Algorithms
D3:	Computational Geometry: Algorithms and Applications

document	words
D4:	Data Structures and Algorithm Analysis on Massive Data Sets
D5:	Computer Organization and Architecture

Our dictionary DICT consists of 8 words: {w1 = data, w2 = system, w3 = algorithm, w4 = computer, w5 = geometry, w6 = structure, w7 = analysis, w8 = organization}. We consider that, by stemming, "computer" and "computational" are regarded as the same word, and so are "algorithms" and "algorithm".

Problem 1. Let tf (w, D) denote the term frequency of term w in a document D. Evaluate the value of tf (wi, Dj) for all $1 \leq i \leq 8$ and $1 \leq j \leq 5$. (5 Marks)

Problem 2. Let idf (w) denote the inverse document frequency of term w. Evaluate the value of idf (wi) for all $1 \leq i \leq 8$. (5 Marks) (CO4)(BL4)

10) What is a neural network? Explain the usage of Perceptron Classifier in NLP Application (CO5)(BL5) (10)

-----End-----