Semester: 6th
Subject Name:- NLP & Code:- IT-3035
Branch (s): - CSE

SPRING MID SEMESTER EXAMINATION-2023

School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University
Subject Name: Natural Language Processing
[Subject Code: IT-3035]

Time: 1 1/2 Hours **Full Mark: 20**

Answer any four Questions including Q. No.1 which is Compulsory.
The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

1. Answer all the questions. [1 x 5 = 5 Marks]

a) If the first corpus has TTR1= 0.073 and the second corpus has TTR2 =0.67 , where TTR1 and TTR2 represent *type/token ratio* in the first and the second corpus respectively, then which of the following statements is/are false?

i. First corpus has more tendency to use different words.
ii. Second corpus has more tendency to use different words.
iii. TTR value sometime can be greater than 1.
iv. A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite.

b) Which of the following is correct about the Markov assumption?

i. The probability of a word depends only on the current word.
ii. The probability of a word depends only on the previous word.
iii. The probability of a word depends only on the next word.
iv. The probability of a word depends only on the current and the previous word.

c) Fill in the blank with the correct alternative: *Morphemes attached at the front and back of stem are called …………*

i. Prefixes
ii. Infixes
iii. Circumfixes
iv. Suffixes

d) Let the rank of two words, w_1 and w_2, in a corpus be 1600 and 400, respectively. Let m_1 and m_2 represent the number of meanings of w_1 and w_2 respectively. The ratio $m_1 : m_2$ would tentatively be:

i. 1:4
ii. 4:1
iii. 1:2
iv. 2:1

e) Consider a simplified language that has an alphabet consisting of only nine letters {a, e, i, b, c, d, f, g, h} having their probabilities of occurrence in a corpus as per the below mentioned table:

Letter	a	e	i	b	c	d	f	g	h
Prob of Occurrence in Corpus	1/16	1/8	1/4	1/16	1/16	1/8	1/16	1/8	1/8

Then which of the following is correct about the per letter entropy E of the language?

i. $E \leq 2.5$

ii. $2.5 < E < 3.5$

iii. $E \geq 3.5$

iv. None of the above

2. Consider the following corpus consisting of four sentences:

<s> three students rohan preeti and akhil are reading book </s>
<s> rohan is reading malgudi days </s>
<s> preeti is reading a detective book </s>
<s> akhil is reading a book by rk narayan </s>

Calculate the Probability of the sentence S: <s> rohan is reading a book </s>, assuming a bigram language model. [5 Marks]

3. (a) What is the difference between Bag Of Words (BOW) and TF-IDF model?

(b) Compute the TF-IDF score for the following corpus having three documents without the removal of stopwords:

D1: data science is one of the most important fields of science
D2: this is one of the best data science courses
D3: data scientists analyze data [5 Marks]

4. Describe classic NLP pipeline. Describe lexical ambiguity with suitable example. Write at least four challenges of Natural Language Processing. [5 Marks]

4. (a) Assume that we modify the costs incurred for operations in calculating Levenshtein distance as follows: (i) both the insertion and deletion operations incur a cost of 1 each, (ii) substitution incurs a cost of 2. Now, calculate the minimum edit distance between the strings "reading" and "writing" by drawing a suitable table. Also derive the number of insertions, deletions and substitutions required corresponding to the optimal alignment obtained between the strings "reading" and "writing" from the same table.

(b) *SpamAssassin* is an online spam filtering tool that works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. It has the following observations during its training: (i) the word "free" appears in 20% of the mails marked as spam; (ii) 0.1% of non-spam mails includes the word "free"; and (iii) 50% of all mails received by the user are spam mails. Based on the observations above, find the probability that a mail is spam if the word "free" appears in it. [5 Marks]

*** Best of Luck ***