AUTUMN MID SEMESTER EXAMINATION-2022

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]

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.

Answer Suggestion 1(a): {i} and {iii}
Evaluation Scheme 1(a): 1 mark for both the correct alternatives, 0.5 for one of the correct alternatives, 0 for wrong alternative(s).

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.

Answer Suggestion 1(b): {ii}
Evaluation Scheme 1(b): 1 mark for the correct alternative, 0 for wrong alternative(s).

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

Answer Suggestion 1(c): {iii}
Evaluation Scheme 1(c): 1 mark for the correct alternative, 0 for wrong alternative(s).

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

Answer Suggestion 1(d): {iii}
Evaluation Scheme 1(d): 1 mark for the correct alternative, 0 for wrong alternative(s).

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

Answer Suggestion 1(e): {ii}
Answer Justification: E = - [4/16 log 1/16 + 4/8 log 1/8 + 1/4 log 1/4] = 1/4 log 16 + 1/2 log 8 + 1/4 log 4 = 1/4 * 4 + 1/2 * 3 + 1/4 * 2 = 1 + 3/2 + 1/2 = 3
Evaluation Scheme 1(e): 1 mark for the correct alternative with justification, 0.5 for the correct answer without justification, 0 for wrong alternative(s).

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]

Answer Suggestion 2: 0.0208
Answer Justification 2:

Prob(rohan | <s>) = count(<s>, rohan) / count(<s>) = 1/4
Prob(is | rohan) = count(rohan, is) / count(rohan) = 1/2
Prob(reading | is) = count(is, reading) / count(is) = 3/3 = 1
Prob(a | reading) = count(reading, a) / count(reading) = 2/4 = 1/2
Prob(book | a) = count(a, book) / count(a) = 1/2
Prob(</s> | book) = count(book, </s>) / count(book) = 2/3
*Prob(Sentence) = 1/4 * 1/2 * 1 * 1/2 * 1/2 * 2/3 = 0.0208333333 ~ 0.0208*

Evaluation Scheme 2: Full marks (5) for complete sentence probability estimation using a bigram language model. Zero marks for wrong answer. Partial marks may be awarded for minor mistakes.

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]

Answer Suggestion 3(a): Technically BOW includes all the methods where words are considered as a set, i.e. without taking order into account. Thus TFIDF belongs to BOW methods: TFIDF is a weighting scheme applied to words considered as a set. There can be many other options for weighting the words in a set. Compared to regular TF-weighted BOW, the TFIDF weighting scheme gives more weight to words which appear in fewer documents and less weight to words which appear in many documents. The rationale is that a word which appears in many documents is unlikely to be relevant since it doesn't help selecting the most similar document. Typically the most frequent words are grammatical words (also called stop words, e.g. determiners, pronouns, etc.).

Answer Suggestion 3(b): A sample scheme for TF-IDF vectorization is illustrated below. However, other internationally-recognized models for TF-IDF vectorization are also acceptable.

$BagOfWords = \{data, science, is, one, of, the, most, important, fields, this, best, courses, scientists, analyze\}$

$TF = \{4, 3, 2, 2, 3, 2, 1, 1, 1, 1, 1, 1, 1, 1\}$

$$TF\text{-}Score = \begin{cases} 1 + log_{10} TF, & if\ TF > 0 \\ 0, & if\ TF = 0 \end{cases}$$

$TF\ Vector = \{1.6, 1.5, 1.3, 1.3, 1.5, 1.3, 1, 1, 1, 1, 1, 1, 1, 1\}$

$DF = \{3, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1\}$

$N/DF = \{1, 1.5, 1.5, 1.5, 1.5, 1.5, 3, 3, 3, 3, 3, 3, 3, 3\}$

$IDF\text{-}Score = log_{10}(N/DF)$

$IDF\ Vector = \{0, 0.2, 0.2, 0.2, 0.2, 0.2, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5\}$

$TF\text{-}IDF\ Vector = TF\text{-}Score * IDF\text{-}Score$

$= \{0, 0.3, 0.26, 0.26, 0.3, 0.26, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5\}$

Evaluation Scheme 3: One (1) mark for 3(a), Four (4) Marks for 3(b). In 3(b), Full marks (4) for complete evaulation of TF-IDF vector using an acceptable TF-IDF vectorization model. Zero marks for wrong answer. Partial marks may be awarded for minor mistakes.

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

Evaluation Scheme 4: One (1) mark allotted for NLP pipeline, Two (2) Marks allotted for lexical ambiguity, and Two (2) Marks allotted for challenges of Natural Language Processing. In lexical ambiguity part, one mark for description and one mark for meaningful example. In NLP Challenges part, 0.5 mark may be awarded for each challenge description. Mentioned challenges should be properly described and/or illustrated with suitable example.

5. (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.

Answer Suggestion 5(a): A suggestive answer is given below:

		W	R	I	T	I	N	G
	0	1	2	3	4	5	6	7
R	1	2	1	2	3	4	5	6
E	2	3	2	3	4	5	6	7
A	3	4	3	4	5	6	7	8
D	4	5	4	5	6	7	8	9
I	5	6	5	4	5	6	7	8
N	6	7	6	5	6	7	6	7
G	7	8	7	6	7	8	7	6

Steps Involved: (1) Insert (W): Cost: 1
(2) No Change (R): Cost: 0
(3) Delete (E): Cost: 1
(4) Substitute (A, I): Cost: 2
(5) Substitute (D, T): Cost: 2
(6) No Change (I): Cost: 0
(7) No Change (N): Cost: 0
(8) No Change (G): Cost: 0
Total Incurred Cost: 6

(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]

Answer Suggestion 5(b): A suggestive answer is given below:

Given Prob(Free | Spam) = 0.20, Prob(Free | Non Spam) = 0.001;
Also given, Prob(Spam) = 0.50, thus yielding Prob(Non Spam) = 0.50;
To find: Prob(Spam | Free) = ?
Using Bayes theorem, Prob(Spam | Free) = Prob(Spam) * Prob(Free | Spam) / Prob(Free)
= 0.50 * 0.20 / (0.50 * 0.20 + 0.50 * 0.001) = 0.995.

Evaluation Scheme 5: Three (3) marks allotted for 5(a) and Two (2) marks allotted for 5(b). In either cases, Full marks for complete evaluation and Zero marks for wrong answer. Partial marks may be awarded for minor mistakes.