B.C.S.E. 4th Year 1st Semester Examination – 2025

Natural Language Processing

Time: 3 Hours

Full Marks: 100

CO1: Mathematical Models

Answer any 3 questions 3*10 = 30

1. Find out the edit distance and alignment between the two strings *"processing"* and *"impression"*, considering the following costs for the edit operations. 10

insertion cost = deletion cost = 1

substitution cost = (your roll number % 2) + 1

2. Given the following term-context matrix, compute which of the following word pairs - [digital, information] and [fish, potato], is more similar according to distributional similarity using add-*n* smoothing, where *n* = (your roll number % 2)+1. 10

term \ context	computer	boil	data	result	fry
fish	0	2	0	0	2
potato	0	2	0	0	1
digital	2	0	3	2	0
information	1	0	6	4	0

3. Consider the following Movie Review document: 10

It's hokey. There are virtually no surprises, and the writing is second-rate. So why was it so enjoyable? For one thing, the cast is great. Another nice touch is the music. I was overcome with the urge to get off the couch and start dancing. It sucked me in, and it'll do the same to you.

Consider the following 6 features:

count(positive Lexicon) ε doc,

count(negative Lexicon) ε doc,

1 if "no" ε doc 0 otherwise,

count(1st and 2nd pronouns) ε doc,

1 if "!" ε doc 0 otherwise,

log (word count of doc)

The feature weights are: 2.5, -5.0, -1.2, 0.5, 2.0, 0.7 and bias = 0.1
Note: sigmoid(0.833) = 0.70 and ln(66) = 4.19
Compute whether the document belongs to positive or negative sentiment.

4. a. Given the following training documents, compute which class the test document belongs to. 5

	Doc_ID	Words	Class
Training	1	wicket wicket run pitch	C
	2	wicket run run bat ball	C
	3	score boundary ground	C
	4	score goal goal penalty	F
Test	5	score ball goal penalty	?

b. If P(0)=0.19 and P(any other digit)=0.09, compute the perplexity of your 12 digit class roll number according to the unigram language model. 3

c. What is the role of the sigmoid function in logistic regression, and why it is important for classification tasks? 2

CO2: Algorithms

Answer any 3 questions 3*10 = 30

1. a. Derive the trigram language model using maximum likelihood estimation, chain rule, Markov assumption and add-1 smoothing. 7
   b. Discuss about the performance issues of the Naïve Bayes Classifier. 3

2. a. Describe and draw the architectural diagram of a simplified feed forward neural language model with N = 3. 8
   b. Why the Neural LMs work better than N-gram LMs? 2

3. a. Compute the cross-entropy loss for a single observation x in Logistic Regression. 4
   b. In the context of logistic regression for text classification, what is overfitting, and what techniques can be applied to reduce overfitting? 2+4

4. a. What is contextual embedding? 2
   b. Define the query, key and value roles of each vector in the transformer. 3
   c. Derive the final equation for single attention head. 5

CO3: Linguistics

1. Answer any 5 questions: 2*5 = 10
   a. Why are stop-words not considered in Information Retrieval?
   b. Mention some cases where case folding is useful and where it is not.
   c. Differentiate between inflectional morphology and derivational morphology.
   d. Differentiate between stemming and lemmatization.
   e. Why the maximum matching word segmentation algorithm works very well for languages like Chinese, but not for English?
   f. Mention some features that you could use in a machine Learning framework for deciding whether a '.' (dot) in English marks the end of a sentence or not.
   g. Differentiate between word similarity and word relatedness.

CO4: Generalization and Analysis

Answer any 3 questions 3*10 = 30

1. a. Given the following confusion matrices for two classes, compute class-wise precision and recall. Also compute macroaveraged precision and microaveraged precision. 6

Class 1

System \ Actual	yes	No
Yes	10	10
No	10	970

Class 2

System \ Actual	yes	no
yes	80	20
no	10	890

b. Give 2 use cases in NLP where (i) precision s preferred over recall and (ii) recall is preferred over precision. Explain your answer. 2+2

2. a. What are the three architectures for large language models? 6
   b. What is temperature sampling? 4

3. a. Let us imagine that you are doing a search on the following three documents with the free text query. Calculate cosine similarity scores of the query with three documents using tf-idf based technique. 8
   Query: **Covid19 pandemic**
   Document1: *Pandemic of Covid19 becomes important for its new variant Omicron*
   Document2: *Pandemic made worth tolerating life style due to Covid19 variants*
   Document3: *Never stop pandemic issue*
   b. Compare multivalue classification and multinomial classification. 2

4. a. Why do you require smoothing in N-gram Language model? 2
   b. Describe linear interpolation and backoff in N-gram Language model. 5
   c. Discuss some features that you can use for developing a spam filter. 3