011

University of Colombo, Sri Lanka

UCSC *University of Colombo School of Computing*

BACHELOR OF SCIENCE HONOURS IN COMPUTER SCIENCE
BACHELOR OF SCIENCE HONOURS IN SOFTWARE ENGINEERING
BACHELOR OF SCIENCE HONOURS IN INFORMATION SYSTEMS

Fourth Year Examination — Semester I— 2022

SCS4209/IS4108 — Natural Language Processing

(Two (2) Hours)

Answer all 4 questions

Number of Pages = 12 Number of Questions = 4

To be completed by the candidate

Index Number | | | | | | | | |

Important Instructions to candidates:

- The medium of instruction and questions is **English**.
- **Write your answers in English.**
- Note that questions appear on both sides of the paper. If a page is not printed, please inform the supervisor immediately.
- Answer all the **4 questions** in the **question paper**.
- Write your index number on each and every page of the **question paper**.
- The duration of the paper is **Two (2)** Hours.
- This paper has **4** questions on **12** pages.
- Any electronic device capable of storing and retrieving text including electronic dictionaries and mobile phones are **not allowed**.
- Non-programmable Calculators may be used.

To be completed by the examiners

1	
2	
3	
4	

Index Number | | | | | | | | |

1. In the following MCQs, **more than one choice** could be correct. Indicate **all correct** answers. Cross or color the correct choice(s) for each question in the following boxes.

[12 x 2 = 24 marks]

1.	(a)	(b)	(c)	(d)	(e)
2.	(a)	(b)	(c)	(d)	(e)
3.	(a)	(b)	(c)	(d)	(e)
4.	(a)	(b)	(c)	(d)	(e)
5.	(a)	(b)	(c)	(d)	(e)
6.	(a)	(b)	(c)	(d)	(e)

7.	(a)	(b)	(c)	(d)	(e)
8.	(a)	(b)	(c)	(d)	(e)
9.	(a)	(b)	(c)	(d)	(e)
10.	(a)	(b)	(c)	(d)	(e)
11.	(a)	(b)	(c)	(d)	(e)
12.	(a)	(b)	(c)	(d)	(e)

1. Which of the following is/are primary ways of collecting data for data-driven NLP?
   (a). Questionnaire
   (b). Querying an API provided by a platform data provider
   (c). Web scraping
   (d). Excel or CSV file
   (e). DBMS

2. Which of the following tools provide functionality for accessing textual data?
   (a). The requests package
   (b). Jupyter Notebook
   (c). The beautifulsoup package
   (d). The scikit-learn package
   (e). The lxml parser

3. Which of the following is/are python packages which can be used for natural language processing tasks such as POS tagging?
   (a). Pandas
   (b). Scrapy
   (c). SciPy
   (d). spaCy
   (e). NLTK

Index Number | | | | | | | | |

4. Which of the following is/are text preprocessing tasks in a classification or clustering task?
   (a). Tokenization
   (b). Case conversion
   (c). Feature extraction
   (d). Stop word removal
   (e). Removing markup

5. Which of the following is/are common feature extraction methods for supervised and unsupervised learning for NLP tasks?
   (a). Bag of words binary vectors
   (b). Byte pair encoding
   (c). Bag of words frequency vectors
   (d). TFIDF vectors
   (e). TFIDF n-gram vectors

6. Which of the following is/are dense vector methods for feature extraction?
   (a). Bag of n-gram vectors
   (b). Word2Vec
   (c). BERT
   (d). GloVe vectors
   (e). Jaccard

7. If the number of unique words in a text collection is very large, which of the following can be used to extract features in order to not run out of memory?
   (a). Use word embeddings instead of term-document vectors.
   (b). Ignore very high frequency and very low frequency terms when vectorizing.
   (c). Use dimensionality reduction techniques.
   (d). Do POS tagging.
   (e). Use a dependency parser.

8. Which of the following is/are true about word embeddings?
   (a). They produce sparse vectors.
   (b). They capture semantic meaning of words.
   (c). They give each sense of a word a different representation.
   (d). They can be learned from raw text data.
   (e). Pretrained word embeddings are publicly available for Sinhala and Tamil.

Index Number | | | | | | | | |

9. Which of the following is a/are algorithm(s) for text classification?
   (a). Naïve Bayes
   (b). XGBoost
   (c). Affinity Propagation
   (d). Linear Regression
   (e). Logistic Regression

10. Which of the following is a/are algorithm(s) for text clustering?
   (a). MeanShift
   (b). NNMF
   (c). PCA
   (d). DBSCAN
   (e). K-Nearest Neighbor

11. Which of the following is/are **TRUE** for NLP model building?
   (a). Repeatedly improving a model by changing parameters and checking performance on test data is the best practice for model building.
   (b). Grid search is a way of tuning parameters in a systematic way in order to arrive at the best model.
   (c). Explainability of the built models is important to convince management.
   (d). In deep learning, the model is built by fitting it repeatedly on the training data to minimize the loss.
   (e). Overfitting can be avoided by regularization in general and using dropout nodes in deep learning models.

12. Which of the following is/are **TRUE** about deep learning models for NLP?
   (a). Dense layers (in Keras) refer to fully connected feedforward neural networks.
   (b). CNNs are more suitable for sequential input than RNNs.
   (c). LSTMs and GRUs are types of RNN.
   (d). NLP problems can be solved by combining RNNs with CNNs.
   (e). Typical RNN models for NLP tasks would have an embedding layer before the RNN and a fully connected layer after it.

Index Number | | | | | | | | |

2. (a). What are speech sounds? Why does the number of speech sounds differ from language to language? Describe them briefly by giving suitable examples.

[5 marks]

(b). Briefly describe the concepts: *Phone*, *Phoneme* and *Allophone* in Phonology with suitable examples.

[6 marks]

(c). Briefly explain the difference between acoustic phonetics and auditory phonetics.

[5 marks]

Index Number | | | | | | | | |

(d). Identify five (5) minimal pairs of words from the following word list.
{lock, glow, bin, bloom, bleed, bloom, block, fool, grow, full, rock, bag, luck, clock, bed, beg, file, bun, room, look }

[5 marks]

1.
2.
3.
4.
5.

(e). By considering the following IPA chart, convert the sentences written in IPA to the English Language.

[5 marks]

Phonemic Chart

		monophthongs				diphthongs			
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	ʧ cheese	ʤ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

(1)	/ ʃiː ɪz maɪ sɪstə /	
(2)	/ duː jʊ laɪk fʊtbɔːl /	
(3)	/ maɪ feɪvərɪt kʌlə ɪz bluː /	
(4)	/ aɪ ɔːlwəz get ʌp ɜːlɪ /	
(5)	/ðə ˈbeɪbɪ wəz stɪl ˈsliːpɪŋ/	

Index Number | | | | | | | | |

3. (a). Briefly describe why text preprocessing is an essential step and important in the process of building models in NLP.

[5 marks]

(b). Identify the *stems*, *derivational morphemes* and *inflectional morphemes* separately of all the following words.

[6 marks]

Word	Stem	Derivational Morphemes (if any)	Inflectional Morphemes
establishment			
repaired			
unfortunately			
wolves			
mismanagement			

Index Number | | | | | | | | |

(c). Consider the following paragraph.

To Sherlock Holmes she is always THE woman. &&All emotions&&, and that one particularly, were abhorrent to his cold##, precise but admirably balanced mind. He was, I take it, draw like, the most perfect reasoning and observing machine that the WORLD has seen, but as a lover he love and would have placed himself in a false position. They were admirable things for the observer–excellent for drawing the veil from men's motives and actions.

What are the text preprocessing steps that need to use to extract root words from the above text? Justify your answer.

[6 marks]

(d). Consider the following corpus of five sentences, with tags added to represent the beginning and the end of the sentence.

[8 marks]

$< S >$ *he is saman* $< /S >$
$< S >$ *saman is not tall* $< /S >$
$< S >$ *he does not like mary* $< /S >$
$< S >$ *saman does not do it* $< /S >$
$< S >$ *we like him* $< /S >$

Calculate the Bigram probability of the following sentence using the above corpus. Show all the steps in your calculations

Sentence: $< S >$ *saman does not like him* $< /S >$

Index Number | | | | | | | | |

Index Number | | | | | | | | |

4. (a). What is Named Entity? Explain why Named Entity Recognition is important in language processing.

[5 marks]

(b). Briefly describe two (2) unique features of each of Top-down Parsing and Bottom-up Parsing.

[4 marks]

Top-down Parshing

1.

2.

Bottom-up Parsing

1.

2.

(c). Consider the following fragment of English grammar.

S → NP VP

NP → Det N | Det N PP | Adj N

VP → V NP | V PP | V S

PP → P NP

Det → a | an | the

N → chair | fox | dress | forest | boy | hat | man | dog

V → ran | walked | barked | looks | run | eat | saw

Adj → angry | nice | smaller | hungry

P → at | on | under | with | into | in

Index Number | | | | | | | | |

i. Write down **(a)** three (3) structurally different and grammatical sentences and **(b)** one (1) grammatical but senseless sentence generated by this grammar.

[4 marks]

(a)

(b)

ii. What additional rule(s) / lexical items would you include to the above grammar to accommodate the following sentences?

[6 marks]

1. The dress looks nice

2. John saw a boy with a dog

3. Boy ran into the forest with a hat

(d). What kind of ambiguity is there in the following sentences? Justify your answers by drawing parse trees or using any other method.

[6 marks]

Index Number | | | | | | | | |

1. I invited the person with the microphone

2. I went to the bank

3. They fed her rat poison.

_____ *****_____