22

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

Fourth Year Examination – Semester I – 2021

SCS4209 – Natural Language Processing – (Part B)

TWO (2) HOURS (for both Part A and Part B)

To be completed by the candidate

Examination Index No: ……………………………………………………

Important Instructions to candidates:

1. The medium of instruction and question is **English**.
2. If a page or a part of this question paper is not printed, please inform the supervisor immediately.
3. Note that questions appear on both sides of the paper. If a page is not printed, please inform the supervisor immediately.
4. Write your index number on each and every page of the Question paper.
5. **This paper consists of two parts, Part A (Question No 1 and Question No 2) and Part B (Question No 3 and Question No 4) and need to be submitted separately.**
6. This Part (B) has **02** questions and **8** pages.
7. Answer **ALL** questions.
8. Any electronic device capable of storing and retrieving text including electronic dictionaries and mobile phones are **not allowed**.
9. **Non-Programmable** calculators are **allowed**.

For Examiner's use only	
Question No	**Marks**
[redacted]	[redacted]
[redacted]	[redacted]
3	
4	
Total	

Index No: ....

Part B

Question 3

Mark the **correct answer/s** by **circling the appropriate options in the Answer Box** given below. A question can have more than one correct option. You should **circle all correct options to get full marks**. Circling of incorrect options would score minus marks. However, such minus marks are not carried forward to calculate the total marks for this question. Therefore, the **lowest mark per sub-question is zero** (0).

Answer Box: Include all answers in the following box by circling the correct options.

Question No.	Options					Question No.	Options				
(a)	i	ii	iii	iv	v	(f)	i	ii	iii	iv	v
(b)	i	ii	iii	iv	v	(g)	i	ii	iii	iv	v
(c)	i	ii	iii	iv	v	(h)	i	ii	iii	iv	v
(d)	i	ii	iii	iv	v	(i)	i	ii	iii	iv	v
(e)	i	ii	iii	iv	v	(j)	i	ii	iii	iv	v

[2x10 Marks]

(a) Pragmatics refers to
- (i). turn taking in two-way conversation
- (ii). meanings of a sentences independent of context
- (iii). meanings of sentences in the real world
- (iv). meanings of a sentences in the context of the paragraph they are in
- (v). the functions a sentence plays in language

(b) Information extraction refers to
- (i). extracting entities and relations from text
- (ii). finding documents that are relevant to a search query
- (iii). extracting syntactic information from a sentence
- (iv). building a logical form from an input sentence
- (v). finding events of interest which relate entities in the text

(c) The following is a/are toolkit(s) available for natural language processing
- (i). NLTK
- (ii). Pandas
- (iii). TextBlob
- (iv). Scipy
- (v). Spacy

(d) The following is a/are machine learning library/libraries which can be used to build data-driven NLP models
- (i). Pattern
- (ii). Tensorflow
- (iii). scikit-learn
- (iv). Keras
- (v). Pandas

(e) The following is a/are library/libraries that can be used to collect online text data
- (i). Numpy
- (ii). BeautifulSoup
- (iii). Gensim
- (iv). Scrapy
- (v). Scipy

(f) The following is a/are ensemble learning algorithm(s) that can be used for text classification tasks
- (i). Gradient Boosted Trees
- (ii). SVM
- (iii). Neural Network
- (iv). Random Fo~~r~~rest
- (v). Affinity Propagation

(g) The following is a/are algorithm(s) suitable for the document clustering task
- (i). K-nearest neighbor
- (ii). AdaBoost
- (iii). Agglomerative Clustering
- (iv). SVM
- (v). K-means

(h) The following is a/are suitable performance measures for clustering tasks
- (i). Sensitivity
- (ii). K-means
- (iii). Silhouette Coefficient
- (iv). F1 measure
- (v). Sum of Squared Error elbow

(i) The following is a/are ways of determining the importance of features in supervised learning.
- (i). The magnitude of the coefficients of a Logistic Regression model
- (ii). The value of the hyperparameter C in an SVM
- (iii). The features used at the decision points further up in a decision tree model
- (iv). The weight matrix in a Neural Network
- (v). The features which correlate best with the target variable

(j) Which of the following is/are true with respect to deep learning
- (i). Neural networks with more than a single hidden layer are deep neural networks
- (ii). Deep neural networks are the only algorithms for deep learning
- (iii). Convolutional neural networks are good at modeling sequence data
- (iv). GRU and LSTM are two types of RNN
- (v). A major problem in deep neural networks was the vanishing gradient problem

Question 4

A Sri Lankan startup company wants to setup a sports information service for sports enthusiasts who want to keep up to date with their favourite sports and teams. In order to support their plan, they want to know how they can access every single relevant piece of sports news as soon as possible.

(a) How would you they be able to access as much of the world of sports as they possibly can?

(3 marks)

(b) Assume that you were able to access one million individual sports news items for the past 12 months. What kind of initial descriptive statistics would you be interested in obtaining to get an idea about the dataset?

(2 marks)

(c) What steps would you take to clean the data you accessed this way?

(3 marks)

(d) Since the dataset consists of all kinds of sports, how would you seek to identify the different sports apart from each other assuming that the individual news items have no label to help you?

(4 marks)

(e) How would you validate the categorization you achieve by applying the method you specified in your answer to (d) above?

(2 marks)

(f) What kind of representations of the news text would you explore to find out the best possible way to extract its features?

(3 marks)

(g) Assume that the categorization you arrive at is close enough to the 'ground truth' in order to be treated so and results in 12 different sports. Explain the different non-deep learning algorithms that you would explore in order to build a classifier for future use, giving reasons for the choice of each such algorithm.

(4 marks)

(h) You also would like to check if deep learning can give a better classifier model for sports data classification. Suggest the kind of deep learning architecture that you would explore for this dataset and which parameters in it you would try to improve your model.

(5 marks)

(i) How would you detect if your model has overfitted the training data and what you could do about it in your deep learning model.

(2 marks)

(j) Assume that the 12 classes used for training had 81k, 16k, 79k, 82k, 76k, 69k, 84k, 78k, 88k, 69k, 73k, 15k data points respectively. What could be done to improve the performance of the classifiers further? Explain how that could be done.

(2 marks)

*****