UNIVERSITY OF LONDON

CM3060

BSc EXAMINATION

COMPUTER SCIENCE

Natural Language Processing

Release date: Thursday 13 March 2025 at 12:00 midday Greenwich Mean Time

Close date: Friday 14 March 2025 by 12:00 midday Greenwich Mean Time

Time allowed: 4 hours to submit

INSTRUCTIONS TO CANDIDATES:

Part A of this assessment consists of a set of **TEN** Multiple Choice Questions (MCQs). You should attempt to answer **ALL** the questions in **Part A.** The maximum mark for Part A is **40**.

Candidates must answer **TWO** out of the **THREE** questions in **Part B**. The maximum mark for Part B is **60**.

Part A and Part B will be completed online together on the Inspera exam platform. You may choose to access either part first upon entering the test area but must complete both parts within **4 hours** of doing so.

Calculators are **NOT** permitted in this examination. Credit will only be given if all workings are shown.

You may use **ONE** A4 page of previously prepared notes in this examination. Please hold up your notes to the camera at the start of the examination.

File upload is **NOT** permitted in this examination.

Do not write your name anywhere in your answers.

PART A

Question 1

Candidates should answer the **TEN** Multiple Choice Questions (MCQs) in Part A.

PART B

Candidates should answer any **TWO** questions from Part B.

Question 2

(a) List down and define the required preprocessing step to clean only the text below. Write the output for each preprocessing step with its output text.

> *“. Dr. Matthew Yee-King is Goldsmiths, University of London’s Programme Director,*
>
> *@Computer Science BSc online. He graduated from the People’s University of Peckham and + goes train spotting with Lauren S.... on @Sundays..."*

[8 marks]

(b) Define Name Entity Recognition and identify any Name Entity Recognition and its types in the example given above.

[6 marks]

(c) Write output in the form of tokens. How many types and tokens are there in the following sentence “As a Machine Learning Scientist, you can design LLM models with the help of https://www.nltk.org/”?

[8 marks]

(d) Real-world scenario: a website uploads and shares different types of news items. You are tasked with designing a high-level solution to automatically categorize this news into true and fake news and help users understand the authenticity of the news.

i. Propose a high-level solution design for processing and classifying these news.

[4 marks]

ii. Justify your choice of fake news classification techniques for this specific application.

[4 marks]

Question 3

(a) Briefly explain how you might solve a text classification problem using a Naïve Bayes classifier for Fake News binary classification.

[8 marks]

(b) Given the following training data, calculate the likelihood probabilities for each word given in each part of speech (POS).

- house/NNS moving/VBG is/VBZ stressful/JJ
- moving/JJ houses/NNS are/VBZ dangerous/JJ
- I/PRP saw/VBZ Essi/NNP moving/VBG houses/NNS
- She/PRP houses/VBZ refugees/NNS

[6 marks]

(c) You are working on a Fake News dataset with a binary classification problem. Your dataset has 12,000 true news and 4,000 fake news samples, for a total of 16,000 samples. What is the problem with the dataset, and how can you resolve this problem?

[8 marks]

(d) Extracting unigram with bigram combined features after removing the stop words, write down the output for the below text.

Sentence = ['As a Machine Learning Scientist you can design LLM models with the help of https://www.nltk.org/']

[8 marks]

Question 4

(a) Explain how WordNet is structured. What kind of tasks is it useful for? Give examples.

[8 marks]

(b) Suppose that you are building a model to identify topics in a corpus of 10 million movie reviews. What techniques could you use to reduce the dimensionality of the data? How might you use TF.IDF to represent the documents?

[6 marks]

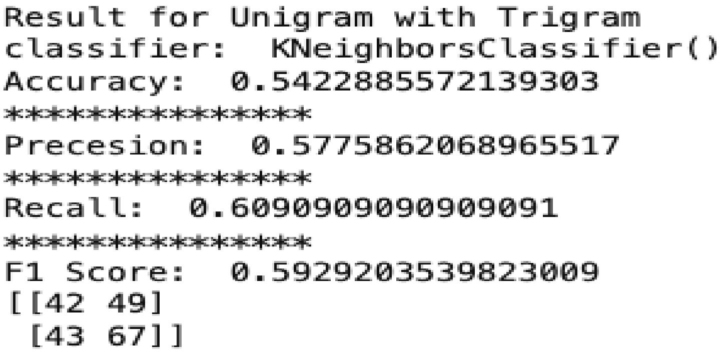
(c) What are the values for F1, Precision and Recall in the following confusion matrix?

	Predicted		
		Fake	True
Actual	Fake	**10** (True Positive)	**10** (False Negative)
	True	**20** (False Positive)	**60** (True Negative)

[8 marks]

(d) Consider the screenshot below, with the confusion matrix presented at the bottom of the screenshot. The parameters of the confusion matrix (Actual, Predicted, etc.) can be assumed to be the same as the confusion matrix in the above question; however, the values you should consider are the ones in the screenshot. Report on the outputs and explain how the accuracy results for the fake news binary classification problem could be improved.

Do you think the dataset is unbalanced? How can you resolve this problem?



```
Result for Unigram with Trigram
classifier:  KNeighborsClassifier()
Accuracy:  0.5422885572139303
***************
Precesion:  0.5775862068965517
***************
Recall:  0.6090909090909091
***************
F1 Score:  0.5929203539823009
[[42 49]
 [43 67]]
```

[8 marks]

END OF PAPER