UNIVERSITY OF LONDON

CM3060

BSc EXAMINATION

COMPUTER SCIENCE

Natural Language Processing

Release date: Thursday 10 March 2022 at 12:00 midday Greenwich Mean Time

Submission date: Friday 11 March 2022 by 12:00 midday Greenwich Mean Time

Time allowed: 24 hours to submit

INSTRUCTIONS TO CANDIDATES:

Section A of this assessment paper consists of a set of **TEN** Multiple Choice Questions (MCQs) which you will take separately from this paper. You should attempt to answer **ALL** the questions in Section A. The maximum mark for Section A is **40**.

Section A will be completed online on the VLE. You may choose to access the MCQs at any time following the release of the paper, but once you have accessed the MCQs you must submit your answers before the deadline or within **4 hours** of starting whichever occurs first.

Section B of this assessment paper is an online assessment to be completed within the same 24-hour window as Section A. We anticipate that approximately **1 hour** is sufficient for you to answer Section B. Candidates must answer **TWO** out of the THREE questions in Section B. The maximum mark for Section B is **60**.

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

You should complete **Section B** of this paper and submit your answers as **one document**, if possible, in Microsoft Word or a PDF to the appropriate area on the VLE. Each file uploaded must be accompanied by a coversheet containing your **candidate number**. In addition, your answers must have your candidate number written clearly at the top of the page before you upload your work. Do not write your name anywhere in your answers.

UL22/0109

SECTION A

Candidates should answer the **TEN** Multiple Choice Questions (MCQs) quiz, **Question 1** in Section A on the VLE.

SECTION B

Candidates should answer any TWO questions from Section B.

Question 2

(a) Ambiguity is considered one of the major problems in Natural Language Processing (NLP). Briefly discuss three different types of ambiguity that make NLP difficult. [6]

(b) Briefly define what is meant by the semantics of a natural language utterance, and how this differs from the pragmatics.

[4]

(c) Define types and tokens. How many types and tokens are there in the sentence "*The sands of time were eroded by the river of constant change*"?

[6]

(d) Describe the process of sentence tokenisation. What challenges might you encounter with a sentence like "*Ms. Smith likes boohoo.com (but not a lot :)*"

[7]

(e) Define word tokenization. What challenges might you encounter with a sentence like "*Prof. Russell-Rose can't believe that's the students' long-term prospects*."?

[7]

Question 3

(a) What are the main shortcomings of Context Free Grammars? Discuss how the use of data-driven methods may be used to address these shortcomings.

[5]

(b) Describe or draw the correct parse tree for the most natural interpretation of the following English sentence: "*She took the advice on board*".

[5]

(c) Give an example of a sentence where knowledge of the syntactic structure (rather than just lexical knowledge) is needed in order to determine the meaning. Explain briefly how syntactic information helps in this case.

[7]

(d) What mechanisms other than word order might be used to convey syntactic roles?

[3]

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

[6]

(f) What is the difference between derivation and inflection? Give examples of each.

[4]

Question 4

(a) What is the motivation for smoothing in statistical language processing?

[2]

(b) Compare and contrast a number of different approaches to smoothing in the context of language modelling.

[6]

(c) Briefly explain how you might solve a text classification problem using a Naïve Bayes classifier.

[6]

(d) What assumptions might you typically make, and how might you approach the task of feature selection?

[4]

(e) What are the values for accuracy, precision and recall in the following confusion matrix?

[6]

	Predicted positive	**Predicted negative**
True positives	10	10
True negatives	20	60

(f) Why is it important to measure more than just accuracy when dealing with an imbalanced data set?

[2]

(g) Some NLP models take a “bag of words” approach and base their decisions on the words contained in a document irrespective of word order. However, two sentences such as This film was exciting and not at all boring and This film was boring and not at all exciting may contain the same words but convey opposite sentiments. Briefly explain how this problem has been addressed in sentiment analysis systems.

[4]

END OF PAPER