UNIVERSITY
OF LONDON

CM2015

BSc EXAMINATION

COMPUTER SCIENCE

Programming with Data

Release date: Monday 18 September 2023 at 12:00 midday British Summer Time

Close date: Tuesday 19 September 2023 by 12:00 midday British Summer 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.

Do not write your name anywhere in your answers.

PART A

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

PART B

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

Question 1

a. Define exploratory data analysis (EDA) and explain its significance in the data analysis process. Provide three commonly used EDA techniques and describe how each technique helps in gaining insights into the dataset.

[6 marks]

b. Describe the process of EDA in your own coursework assignment.

[6 marks]

c. Identify and explain five contemporary challenges faced in data visualisation using Python. Discuss how these challenges impact the effective representation and communication of complex data.

[10 marks]

d. When working with datasets downloaded from Kaggle, there are several limitations that users may encounter. Describe five considerations that you might take into account when reviewing your dataset.

[5 marks]

e. Identify three ways that you can improve upon the quality of datasets acquired from places like Kaggle.

[3 marks]

Question 2

a. Define web scraping and explain its purpose in the context of data extraction. Describe the fundamental steps involved in the process of web scraping.

[6 marks]

b. Discuss the ethical considerations and potential challenges associated with web scraping. Explain at least two ethical considerations and two challenges that individuals or organisations may face when implementing web scraping techniques.

[4 marks]

c. Define test-driven development (TDD) and describe its purpose and benefits in the software development lifecycle. Explain the fundamental steps involved in the TDD process.

[3 marks]

d. Discuss advanced techniques or best practices that experienced developers can employ to enhance the effectiveness and efficiency of the test-driven development process. Provide specific examples or strategies where possible.

[4 marks]

e. Describe SQLite and its role as a database management system. Explain the advantages and use cases of SQLite in comparison to other database systems.

[6 marks]

f. Explain the process of implementing SQLite in Python. Describe the necessary steps to create a database, execute SQL queries, and retrieve data using Python's SQLite library.

[2 marks]

g. Discuss the benefits of using version control systems (VCS) in the context of data science projects. Explain at least three key advantages that VCS provides to data scientists and their projects.

[3 marks]

h. Identify and explain two potential pitfalls or challenges that data scientists may encounter when utilising version control systems. Discuss how these pitfalls can impact project management, collaboration, or data integrity.

[2 marks]

Question 3

a. Discuss three difficulties in natural language processing, for instance when working with technical documentation such as developer guides or domain-specific texts such as medical documents.

[6 marks]

b. Provide four examples of how you would create stopwords for a technical piece of text.

[8 marks]

c. Share your opinion on the preferred approach for grouping words or sentences. Explain how you would group fragments of text together to preserve semantic meaning effectively.

[4 marks]

d. Provide two reasons why regular expressions might not be appropriate for solving problems in Python.

[4 marks]

e. Describe four contexts where catching errors is desirable and provide examples of scenarios where a system could benefit from robust error handling.

[8 marks]

END OF PAPER