

**Computing, Digital Forensics and Cybersecurity**

**2018/19**

**MCOMD2CAL- Computational Algorithms**

**String matching**

**Date: OCT 2018**

**General guidelines for submission**

* This is an individual submission and must be presented as your own work.
* The required date of submission is 19th Nov 2018 14:00 via Turnitin on Blackboard.
* This assignment has been set by Vijay Sahota; Email: Vijay.sahota@canterbury.ac.uk
* This assignment has been moderated by Richard Henson
* This assignment is worth 50% of the course marks

**General advice**

* You are advised to back up your work regularly onto your N: drive and on removable storage devices. Always check the date-stamp on your files before submission.
* You must submit your work using the software versions we currently have on the University’s network.
* Please note that a high standard of presentation is expected. The report should have an appropriate Title Page and a Contents Page, and each page is to be numbered.
* The title page should have the following information:
  + Course Code
  + Course Title
  + Assignment Title
  + Lecturer’s Name
  + Student’s Name
  + Academic Year
  + Hand in Date
* You are required to use the originality checking feature available in Turnitin to help improve your academic report writing skill

**Learning Outcomes of this module:**

* Demonstrate an understanding of computational complexity.
* Demonstrate knowledge of existing algorithms.
* Critically evaluate existing algorithms.
* Implement and develop an effective algorithm to current issues in computing

**Scenario:**

You have been tasked to compile a survey paper outlining, comparing and suggesting the best string matching algorithm specific to a given task.

Your chosen language to implement these algorithms is C# and it should be clear due to the statistical nature of the problem that you are to execute tests multiple times.

**TASK 1: Algorithm:**

For this section you must research the list of algorithms below, and simply state their Big O performance (state others if they exist) and what this means in terms of runtime, computational complexity (impact on CPU type) and memory consumption. Furthermore for each algorithm provide a lay but complete explanation of their operation (provide your own worked example).

* Naïve implementation (Brute force)
* Horspool algorithm
* Reverse Colussi algorithm

[30 Marks]

**TASK 2: Experimentation:**

You are to now required to implement the algorithms researched for Task 1, using their reference implementations (found online - http://www-igm.univ-mlv.fr/~lecroq/string/) and re-compile (source on BB now) into C#.

Once achieved perform tests using relevant data with respect to your assigned scenario (as outlined in the table below) along with relevant search patterns.

As already stated you are to perform multiple tests (to obtain accurate averages) as well as performing test against different search patterns (+ lengths) and different sizes of source text.

Once you have ascertained all your experimental results, comment on each algorithm’s performance in this simulated real world scenario and attempt to explain and then justify your analysis of the results.

Note: use a single core CPU (use only 1 thread) in your testing

|  |  |
| --- | --- |
| Application (Source) | Assigned Student |
| Code parsing (C# code as text) | Sime, Kervanli, |
| Gene searching (Bio-informatics FASTA) | Gravesande, Truman, |
| English text (Public domain novels) | Herring, Booker, |
| MD5 Hashes (Auto generate) | San, White, |
| Telephone numbers (Auto generate) | Shadbolt, Prior, |

[40 Marks]

**TASK 3: Improvements:**

Given your understanding from Task1 and your results from Task 2, suggest suitable modifications and areas to investigate that would either improve the existing algorithms or provide significant basis for a new algorithm.

Along with justifications as to why your suggestion would increase and/or better performance, attempt to provide a rough (does not have to be Big O) idea or guesstimate the performance improvement.

[20 Marks]

**Quality** of report presentation, assumptions, proper referencing, etc.

[10 Marks]

**Deliverables:**

You are expected to produce a report of a high professional standard adhering to the guidelines given at:

P:\COURSEWORK\IT\\_Departmental Standards for Students

The length of report should be 2000 words.

**Assessment:**

* The report must show your understanding of the subject matter and be Harvard referenced throughout (P:\COURSEWORK\IT\\_Departmental Standards for Students\References and Bibliographies.doc).
* Note: Online Harvard Referencing style resource <http://www.canterbury.ac.uk/library/help-and-advice/citing-references/harvard08.pdf>

**Mark Distribution:**

|  |  |  |
| --- | --- | --- |
| **Task** | **Detail** | **Marks** |
| **1** |  | **30** |
|  | Performance equations and their discussion. | **0-15** |
|  | Explanation and worked examples | **0-15** |
|  |  |  |
| **2** |  | **40** |
|  | A complete set of relevant tests carried out | **0-20** |
|  | Comments on their performance and explanation/ justification | **0-20** |
|  |  |  |
| **3** |  | **20** |
|  | Suggested factors to consider/ investigate | **0-10** |
|  | Justification for plausible impact, if suggestion were implemented | **0-10** |
|  |  |  |
|  |  |  |
|  | *Quality of report: presentation, assumptions, proper referencing* | **10** |
|  |  |  |
|  | **MAX TOTAL** | **100** |