University of Wollongong
School of Computing and Information Technology
CSCI446/946 Big Data Analytics
Spring 2024

Assignment 2
(20 marks)

Aim
This assignment aims to provide students with essential experience conducting big data analytics experiments with the Python programming language. In this assignment, you should

1. procedure big data analytics by following Big Data Analytics Lifecycle, appropriately choose, apply and evaluate core models/algorithms and analytics techniques to complete the analysis tasks,
2. understand and integrate the knowledge and skills learned in this subject, including big data analytics lifecycle, data preparation, clustering, classification, regression, association rules, data/model evaluation, data visualization and text processing.

Group work: You are to work on this assignment as a group. Each group is to work independently from other groups on this assignment. Groups and group memberships are to remain as they were for Assignment 1. All group members shall contribute to this assignment. Please plan before starting the assignment, then keep a detail digital work log and timesheet for each group member. A justification and/or explanations must accompany all your answers to this assignment. One submission per group only.

Penalties: If a group member fails to make a minimum contribution, the member will be awarded zero marks. Claims of less or no contribution should provide evidence like a work log. Plagiarism of any part in this assignment will result in zero marks being awarded to the whole group.

Preliminaries
Read through the lecture slides, lab instructions and the recommended readings in Weeks 1 – 9. Conduct relevant background studies. You should use Python for the tasks in this assignment. You can use any publicly accessible toolbox of library for Python. Your submission must include the source code file(s) which, when run, would re-create all your results.

About the Dataset and the Original Project
The dataset for Assignment 2 was originally used to train a CrowdFlower AI gender predictor. Contributors have been posting code and discussion on the web as well as here, here, here, here, among others. The resources are listed here to motivate your design of Assignment 2.
NOTE: Assignment 2 is different to any public project. Copy of any part of any public project will lead to zero mark for Assignment 2.

The dataset contains 20,050 rows, each with a username, a random tweet, account profile and image, location, and even link and sidebar colour. Detail dataset overview is available in "Twitter User Data.pdf" provided with the assignment instruction.

Essential Tasks of Assignment 2

Assignment 2 aims to find misinformation on social network, i.e., **identify profiles that are mistakenly recorded as human/non-human profiles**, focusing on the twitter_user_data dataset. Your essential tasks include the following Tasks 1-4.

Task 1: Design a big data analytics project by following Big Data Analytics Lifecycle. (3 marks)

Task 2: Process the data by considering the different data types and considering the various data properties. Correspondingly apply (mandatory) core models/algorithms. Consider, select, and justify the use of regression, association rules, clustering, classification, and text processing methods. Higher marks will be awarded when more models/algorithms are investigated and properly applied. (10 marks)

Task 3: Visualize the data and use visualization for the evaluation of results. (5 marks)

Task 4: Study factors in multiple views/modes (e.g., text, colour, tweet, etc.) and make suggestion to amend non-human and human profiles. (2 marks)

A report is required to summarize Tasks 1-4 in a well-organized way and cite referred articles and programming resources in your writing. Tasks 2-4 need Python programming to support your analysis.

Submission:

The submission link for Assignment 2 is on the subject's Moodle site. Only one submission per group. **The submission must be a zip file named "A2.zip", under 200 MB, and contains a report (mandatory), and code (mandatory).** The following file formats are acceptable: a report in **.pdf** format, and code files in **.py**.

Important:

1. The report must be in a single file and in .pdf format.
2. The title page must list the full name and student ID of all members in the group.
3. The title page must clearly show the contribution of each member (in percentage) and the task(s) that each member has done.
4. The report does not have a page limit.
5. Marks will be deducted for incomplete or vague descriptions.
6. Sufficient, suitable, and legible annotation shall be provided in your code to make it easy to understand. Marks will be deducted for untidy code, code difficult to read, code that does not run, or code that does not reproduce the results in your report.

Note: Failure of your code to run on a computer in the lab may attract zero marks. Marks may be deducted if we cannot reproduce your results by using your code. **Plagiarism of any part of your code, or any part in your report will attract zero marks for this assignment.** It is the responsibility of the group to ensure that your submission does not contain plagiarized material. You may be requested to demonstrate and explain your program or explain your answer in the report. Marks are deducted if you are unable to offer an explanation. Marks will be awarded for correct design, implementation, style, completeness, and justification.

---------- ***END***----------