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

Assignment 2

(20 marks)

Due: 23:59 Beijing Time, 15th November 2024

Aim

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

- 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,
- 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 as specified on Moodle. You can form groups of your own accord. Each student can only join in one group. Each group should contain no more than 5 members and no less than 3 members. All group members are expected to contribute to this assignment. 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 either R or Python for the tasks in this assignment. You can use any publicly accessible toolbox of library for R and Python. Your submission must include the source code file(s) which, when run, would re-create all your results.

About Dataset and 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 here: https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification
The resources can motivate your design of Assignment 2.
NOTE: Assignment 2 is different to any public project. Copy from any public project will lead to zero mark for Assignment 2. The dataset contains 20,000 rows, each with a user name, a random tweet, account profile and image, location, and even link and sidebar color. Detail dataset overview is available in "Twitter User Data.pdf".

Essential Tasks of Assignment 2

Assignment 2 aims to find some misinformation on the social network, i.e., identify the mistakely recorded 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 data in different types and having different properties, and correspondingly apply (mandatory) core models/algorithms. They are regression, association rules, clustering, classification, and text processing. (10 marks)
Task 3: Visualize the data and visualization for evaluations. (5 marks)
Task 4: study factors in multiple views (e.g., text, color, tweet, etc.) and make suggestion to amend switching between 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 R/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 two files. One is the report (mandatory) named in "`A2.pdf`"; another is a zip file named "`A2.zip`", under 200 MB, and contains code (mandatory) and video presentation (optional).** Either following way is acceptable:

1. a report in `.pdf` format, and code files in `.R` or `.py`; or in `.ipynb`

A video presentation is optional in `.mp4` format or a shared link saved in a `.txt` file.

Important:

- The report must be in a single file and in `.pdf` format. The title page must list the full name and student ID of all members in the group. Clearly indicate members' contributions.
- The report does not have a page limit.
- The report will be checked by Turnitin system in Moodle site for Plagiarism test.
- Marks will be deducted for incomplete or vague descriptions.
- 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 that is 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 may attract zero marks. Plagiarism of any part in 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***----------