Programming for Big Data / Data Application Development
Project Description

You are to carry out a series of analyses of two datasets utilizing appropriate programming languages and programming environments. For each of the chosen datasets you are required to compile a report of the analysis. For the report, you should provide the following:

- a formal description of the underlying datasets **[2 x 5%]**
- a description of the objective of the analysis **[2 x 5%]**
- the data processing activities carried out **[2 x 5%]**
- the presentation of the final analysis results along with any associated annotations and supplementary information **[2 x 15%]**

Additionally, you should also identify a particular dataset and analysis that can utilize the MapReduce programming paradigm for processing. You should then:

- implement and present the algorithms to process the dataset in a relevant environment **[15%]**
- present configuration details on how you set up the environment **[5%]**
- present and discuss your results in a meaningful manner **[20%]**

Note: A minimum implementation is R or Python, with MapReduce. However, at least one of the following technologies: Pig, Hive, H2O can also be used.

All deliverables should be compiled into an accompanying paper, which should be submitted along with any programming code elements.

Your project report should discuss the challenges that you encountered whilst handling your chosen datasets and the means and mechanisms you implemented to overcome these challenges. It should be structured as follows:

- **Abstract**: a roughly 200-word executive summary of the project and the key results
- **Introduction**: set the scene of the project, i.e., the objectives of the project (for example what are you trying to find out)
- **Data**: present the data sets chosen, and why. Conceivable here, would also be to discuss how other people have used the data sets you have chosen.
- **Methodology**: essentially, how have you applied KDD, CRISP-DM, SEMMA or a similar methodology to your project?
- **Implementation and Architecture**: how have you built your application workflow, what components and/or forms of analytics have you used and why?
- **Results**: what did you find out about your data sets? E.g.: what was surprising? what was expected? what did you find out with respect to your motivational question that is presented in the introduction? Finally discuss any interesting aspects of your results or key challenges you solved in achieving your results.
- **Conclusions and future work**: what (in general) did you learn and find out? If you were to do the project again, what would you do differently? If you had more time (e.g. in your final project) what would you do next to extend your work?

The paper should be formatted in the IEEE double column format and ideally be 6 pages long including all figures and any references to existing work. Please refer to this link for the formatting requirements: https://www.ieee.org/conferences_events/conferences/publishing/templates.html

Note*:* the project contributes towards a maximum of 50% of the marks for the module.