COMP SCI 4094/4194/7094 - Distributed Databases and Data Mining
Term Project Guidelines

Project Objectives and Scope

The objectives of the term project is that you will have a good understanding of the given research topic, provide insight into its solution and a well defined strategy for its solution. You should treat the term project as if you were doing the initial background study for further in-depth research. In other words, the report should demonstrate an understanding of and an insight into the problem such that given enough time, you could carry it to its logical conclusion and complete the research.

Project Description

The project has two parts: an in-depth literature review and an implementation of a classification problem. For groups that with **1** (i.e., individual project) or **2** student(s), only literature review is required, see details in Deliverable section.

- **Literature review**. It describes the problem domain with proper problem definition, and a survey of existing work. The research topic of this term project is **Web Mining and Content Analysis**.

  The sub-topics include: **a.** Crawling and indexing Web content; **b.** Web recommender systems and algorithms; **c.** Summarization of Web data; **d.** Data, entity, event, and relationship extraction; **e.** Knowledge acquisition and automatic construction of knowledge bases; **f.** Large-scale graph analysis. **Please pick one of them.**

  You should be looking at the proceedings of conferences such as WSDM, WWW, SIGIR, ICDM, KDD, SIGMOD, VLDB, ICDE, ... and at journals such as IEEE TKDE, Data Mining and Knowledge Discovery Journal, Journal of Intelligent Information Systems, Intelligent Data Analysis, World Wide Web, VLDB Journal, Knowledge and Information Systems and many others. Most of these publications can be obtained through DBLP: `https://dblp.uni-trier.de/db/index.html`. This is not meant to be a complete list or may not even be the most important ones from your perspective. Please do your research and find the relevant papers to your chosen topic. Our hope is that by the time you complete the project, you'll have a good idea of what the area is about and what the most important publications are.

- **Classification Implementation**: You need to implement C4.5 Decision Tree algorithm in **c++** and perform classification on the following dataset: `https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009`. You are required to split the dataset into 80% training and 20% test.

Deliverable

The term project (70% of the overall mark) consists of a project outline, a project report and a project presentation:

- Project outline: 10%
- Project report: 40%
- Project presentation: 20%

The project could be undertaken individually, or in a group with a maximum of 5 students. You should submit the three submissions to MyUni by ONE group member (could be different ones for the three submissions). For project outline and project report, all members of the group will receive the same mark. You should prepare your presentation slides as a group. Each member should present her/his own contributions of this report. Hence we mark the presentation individually based on your contribution and presentation. It is incumbent upon you to make sure that all group members share the tasks in the work if you are in a group of more than 1 student.

Important

- For groups that with **1** (i.e., individual group) or **2** student(s), you only need to conduct literature review. So all the 70% marks go to literature review. The overall length must be less than 20 pages using ACM Computer Survey submission format (`https://dl.acm.org/journal/csur/author-guidelines`), without considering references.
- For groups that with **3** or **4** or **5** students, you **must** do both literature review and classification. To report the classification, you need to attach your code on the report and illustrate the results in proper visualizations, e.g, figures and/or tables. The length of literature review must be less than 20 pages, using ACM Computer Survey submission format (`https://dl.acm.org/journal/csur/author-guidelines`), without considering references.. The length for classification implementation is unlimited.
  - Project outline: 10% = 8% on literature review + 2% on classification implementation
  - Project report: 40% = 30% on literature review + 10% on classification implementation
  - Project presentation: 20% = 15% on literature review + 5% on classification implementation