Data Management I (Data Profiling and Quality Dimensions)
Exam Guidelines and Requirements

Prof. Dr. Ajinkya Prabhune
Prof. Dr. Barbara Sprick

- The submission for this module has two parts; detailed are explained as **Goal-A** and **Goal-B** below.
- **Goal-A** →Each student will be given a dataset and it is expected that you profile and clean the dataset considering the different quality dimensions.
- The exam format for **Goal-A** will be a report submission – the report should not be more than 15-20 pages, following a single column format.
  - Each student has to submit his/her completed report until 15th March 14:00 in Moodle.
  - Source code, if any must be zipped and uploaded with the report in Moodle.
  - The report should not contain any subjective description (general information) but should be specific with objective statements on the steps taken in cleaning the data.
- **Goal-B** → Each student has to also prepare an unclean dataset**.**
- The exam submission for **Goal-B** will be a short document, recipe/source-code if any and the unclean dataset in either CSV, XL, or a MySQL database dump.

Goal - A

a) Data Profiling

- Using Talend Data Quality Management tool, you have to document the data profiling steps performed on the dataset
- For clear understanding of the profiling steps performed, you can paste the screenshot from the respective tool.

b) Data Cleaning

- Mention all the relevant data quality dimension necessary for cleaning the dataset.
- Document in detail the steps performed for handling a given quality dimension.
- OpenRefine, Talend Data Preparation, or Trifacta can be used for performing the data cleaning process
  - **Bonus point** for those who will be showing a comparison of data cleaning features between the 3 above mentioned tools (this task is not mandatory).
- The datasets are in CSV format, exporting them in a DB system for performing specific/advance queries is up to each individual. However, that information should be part of the report.
- Following questions have to be answered for the task
  - What insights did you gain from dataset analysis?
  - Which all data quality dimension does your analysis fits into?
    - Explain the dimension with a use case (the use case can be hypothetical one created by you)
    - For example, in the classroom tutorial on the Patent dataset, the use case was finding the different countries filing patents;

*If you wish to use a different dataset than the one provided by us, then please inform us as soon as you possible.

    - To answer this use case, the dimensions conformity + consistency + accuracy was considered for the column “PublicationNumber”.
- Why did you choose a given technique and the associated actions steps for a task?
    - Is there an alternative technique that could have been used?

GOAL – B

- Creating a Raw dataset refereeing to any online source like kaggle.com.
- Conditions on the Raw dataset
    - Minimum dimensions to be considered while making the dataset Raw are
        - Accuracy
        - Completeness
        - Conformity
        - Validity
        - Consistency
        - Uniqueness
        - Currency
    - You are free to include any other dimension from the Literature.
- The dataset must be made at least **25% Raw/Unclean** based on the above-stated dimensions.
- Every record/set of records that you make unclean should be documented highlighting the quality dimension that is considered.
- The recipe/source-code should be also submitted for verification
    - You can use any or a combination of OpenRefine, Data Preparation, or Trifacta.

*If you wish to use a different dataset than the one provided by us, then please inform us as soon as you possible.