**Project Rules, PRI 2022/23**

*Information Processing and Retrieval (PRI)  
Master in Informatics Engineering and Computation (M.EIC)  
FEUP, U.Porto*

**I. Introduction**

The course project runs for the whole semester and has three milestones, each with a report and a presentation as deliverables. The end result of the project is an information search system, including work on data collection and preparation, information querying and retrieval, and retrieval evaluation.

The project is developed in groups of 3 students and starts with the selection of a topic and the relevant data sources. By the end of the second practical class, groups and topics have to be established.

Topics are free but cannot be repeated in the same class. Selected datasets must be mostly of an unstructured nature and rich in textual data (i.e. long text fields, not simply textual labels).

**II. Deliveries and oral presentations**

Each project delivery (milestone) has a corresponding presentation and discussion. In the weeks assigned to project presentations, the practical class will be organized in a workshop format, with project presentations and discussions according to an established schedule.

The grade for each project delivery considers the written report, and the presentation and discussion. The grade for deliveries has an individual component, that is positive if the student contributes to the workshop with questions or comments, and negative if the student is not present or shows unprofessional behavior.

The final project evaluation corresponds to a weighted average of the milestones evaluations. The final classification of the project may vary from element to element of the same group, in more or less 3 grade points, based on the opinion of the professors and on the self- and hetero-assessment to be carried out internally in each group.

Reports are written as short scientific papers, using a two-column format (4 pages maximum in each delivery; a final report with up to 12 pages is expected). The deadline for the submission is posted in the course plan. Each report is a self-contained work-in-progress and is based on the previous deliveries, i.e. each delivery corresponds to a new part and extends on the previous delivery. Students can use the standard two-column template from ACM ([Word and LaTeX templates available at ACM](https://www.google.com/url?q=https://www.acm.org/publications/authors/submissions&sa=D&source=editors&ust=1664463476440781&usg=AOvVaw0eaT8cyRT0ukLMYihf7lGH) and also on [Overleaf](https://www.google.com/url?q=https://www.overleaf.com/latex/templates/acm-conference-proceedings-new-master-template/pnrfvrrdbfwt%23.W5k05mRKgWo&sa=D&source=editors&ust=1664463476441076&usg=AOvVaw1vuiP6j3sD9x_D4KFoGtfE)) or other two-column based templates.

Electronic submissions of the project deliverables are accepted up to 18:00 on the day before the presentation day. Moodle is used for submissions, which include:

* PDF version of presentation slides (demo-gg.pdf);
* PDF version of the report (report-gg.pdf);
* Code, depending on the milestone (code-gg.zip)

**III. Project Milestones**

The project runs for the whole semester, starting with the group setup and ending with the final presentation. The project is reported and evaluated at three milestones, namely: data preparation, information retrieval, and final search system. The expected results depend to some extent on the selected topic, but some details are provided below.

**Milestone #1: Data Preparation**

The first milestone is achieved with the preparation and characterisation of the datasets selected for the project. The datasets are the foundation for the project and the goal of the first task is to prepare and explore them. This task is heavily dependent on the datasets, which may require some extraction actions such as crawling or scraping.

Work on these tasks depends on the nature, volume, organization and accessibility of the selected datasets. As a result of this milestone, a well-documented and reproducible pipeline of data processing is expected. Groups must include the Makefile in the milestone submission.

The following list has a sample of the actions which are required:

* search repositories for datasets;
* select convenient data subsets;
* assess the authority of the data source and data quality;
* perform exploratory data analysis;
* prepare and document a data processing pipeline;
* characterize the datasets, identifying and describing some of their properties;
* identify the conceptual model for the data domain;
* identify follow-up information needs in the data domain.

**Milestone #2: Information Retrieval**

The second milestone is achieved with the implementation and use of an information retrieval tool on the project datasets and its exploration with free-text queries.

This task makes use of state-of-the-art retrieval tools and involves the view of the datasets as collections of documents, the identification of a document model for indexing, and the design of queries to be executed on the indexed information.

Also included in this milestone is a brief description of the ideas to explore in the development of the final search system, i.e. Milestone #3.

The following list has a sample of the actions which are required:

* choose the information retrieval tool (Solr, Elasticsearch, Lucene, Terrier, …);
* analyze the documents and identify their indexable components;
* use the selected tool to build the indexes;
* use the selected tool to configure and execute the queries;
* demonstrate the indexing and retrieval processes;
* manually evaluate the returned results;
* evaluate the results obtained for the defined information needs.

**Milestone #3: Search System**

The third milestone is achieved with the development of the final version of the search system. This version is an improvement over the previous milestone, making use of features and techniques with the goal of improving the quality of the search results.

For this milestone, each group is expected to explore innovative approaches and ideas, and will heavily depend on the context and data of each group. Additionally, an extended evaluation of the results and a comparison with the previous version of the search system is also expected.

As examples of topics to explore, groups may: incorporate new information retrieval algorithms; expand the information available for each document by adding and linking new datasets; work on user interfaces by developing a frontend for the search system; etc.