Practical assignment

The practical assignment involves the construction of a datawarehouse and the elaboration of data analysis over it, with the construction of graphical reports supported by the Python module matplotlib.

The theme of the datawarehouse will be the national swimming competitions at the master level. More information about the organization of this level of swimmers is available at the following link:

https://en.wikipedia.org/wiki/Masters_swimming

To extract the base information, files in Lenex format (LXF) will be used: https://wiki.swimrankings.net/images/6/62/Lenex_3.0_Technical_Documentation.pdf

To facilitate the interpretation of the XML format of Lenex files, a Python script is available on the Moodle page of the course that reads LXF files and generates SQL to create and populate a database with the extracted information. Note that the database that is created by this script is not a database designed as a datawarehouse, as that component is part of this current assignment.

The python script assumes a directory "files" where the lxf files should be placed. Running the script will then generate a sql file for which of the lxf files.

There are several LXF files associated with the Master level on the websites of either the Portuguese Swimming Federation (www.fpnatacao.pt) or the Northern Portugal Swimming Association (www.annp.pt), among others. Some have been placed in Moodle, but you are free to enrich your datawarehouse based on other LXF files.

The first objective of the work involves modeling, creating and loading the datawarehouse with the extracted information. You can define dimensions and measures with some freedom, but clearly dimensions like Athlete, Club, and Swim Style (which is defined by a stroke type and a distance), should be part of the datawarehouse.

The second objective of the work involves the multi-dimensional analysis of the information stored in the datawarehouse, with specific operators such as GROUP BY CUBE, GROUP BY ROLLUP and several aggregation functions. The objective will be to build highly informative reports that analytically summarize the information stored in the datawarehouse. These reports must also be presented in graphical form, when applicable, using SQL embedded in Python scripts that generate the graphs with the matplotlib module.

In addition to the report, you should prepare a slide-based presentation of your work. These presentations, which count individually to the final grade, will be done on June 9th, in a schedule of presentation slots that will be defined. The work should be submitted in Moodle until the 00:00:00 of June 9th, and comprise the report, the code files (python/sql) and the presentation slides.