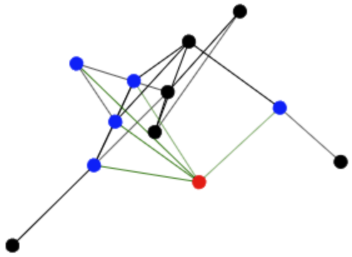
Massive Graph Data Extraction

CentralSupelec

8 novembre 2019



The aim of this project is to establish a listing of Gutenberg dataset books classified according to there unavoidability using the graph made with the jaccard similarity between the books.

You can estimate there unavoidability of a book using on of the three indexes (*closeness*, *betweeness* or *pagerank*).

In order to obtain a more precise estimate you can use a secondary and tertiary indexes. Also you can use different values of the threshold of the graph to expand the amount of information about the book and its position in the graph.

It is possible to find up to four or five books with the same rating.

The work must be submitted in the form of a documented project by explaining the options chosen when it is carried out.

In addition the work submitted should clarify the following points:

- The description of the dataset used
- A presentation of the algorithms used
- A concise argument for the choice of algorithms and options not selected
- Figures on the results of algorithms with precise explanation
- benchmarks

No more than 10 pages (between 5 and 10). It is possible to form groups of two or three students to work on the project.