Network Tour of Data Science: Project Proposal

Team composition : Armand Boschin, Bojana Ranković, Quentin Rebjock

Graph: Wikipedia hyperlink network

Problem: Does the structure of the graph bears info on the content of the nodes ? We would like to find out if it is possible to detect communities of pages just by looking at the hyperlink connections and match these communities with real-world data such as categories of the pages.

Steps of the project:

1) Scraping the Wikipedia hyperlink network. Start from one node and get the pages as far as 2 or 3 hops depending on the number of nodes we get.
2) Try to apply spectral clustering in order to detect clusters of pages.
3) Visualize the clusters to match them with real-world categories (using some of the tools from the last guest lecture).
4) Model the network by a random graph/scale-free network/something else in order to try to retrieve some of its characteristics.
5) Improve the community detection using for example the Louvain algorithm.