NTDS Project Proposal

El Amrani Ayyoub - Micheli Vincent - Myotte Frédéric - Sinnathamby Karthigan
Group №2

1 Data product

We would like to build a query based search engine for Wikipedia using Graph Machine Learning. In other words, given a list of keywords entered by a user, it would recommend Wikipedia articles related to the topics that are embedded within the query.

2 Dataset

We will use Wikipedia articles obtained from the following tool: *Seealsology*. Thanks to Seealsology crawls, we will be able to build a graph by using articles as nodes and the "See also" section as a set of edges to link articles. We will first pick a few topics and then expand the scope of the dataset according to the computational resources available.

3 Models and Methods

After a general study of the graph, we will need to build embeddings for the nodes. Such embeddings can be constructed thanks to Spectral Clustering or *the DeepWalk algorithm* for instance.
Node embeddings will be used to compute similarities between articles. The idea is to recommend the closest articles. For instance given the query: "Computer Science, Statistics and Business", the articles "Computer Science", "Statistics" and "Business" would be extracted. After computation of the mean of their embeddings, we hope that "Data Science" would be the first article proposed.
Embeddings can also be used to cluster the data and assign labels to pages in an unsupervised manner. A user could then refine the search space by specifying a label before the query.

4 Evaluation and Visualization

First, we will need to evaluate our model. We could perform a qualitative analysis of our search engine by comparing its suggestions to Google search's results.
Then we would like to build an interactive visualization as follows: for a given query, color the graph based on the predictions and give the possibility to interact with those nodes (consultation of the Wikipedia articles themselves). We could also visualize the nodes in the embedding space by applying dimensionality reduction techniques.

5 Motivation

This data product would enable users to quickly access new and relevant articles based on their queries. Moreover, it would be a Wikipedia-centered query engine leveraging the specificities of the graph spanned by the encyclopedia.