Network Analysis with learned Document Embeddings

Student

Matthias König

Introduction: Today it is important to be well connected. This applies for person as well as for corporations, projects, organisations etc. To improve the connections it could be helpful to analyse the networking. There are several ways to achieve this. In this work, networking was analysed through content related similarities. The aim of this work was to develop a method to analyse the networking of different organisation based on their website content. The method was developed on the basis of a database of the Institute for Landscape and Open Space at OST. The organisations belong to the field of nature conservation. The database contains about 1900 organisations with their names and urls.

Approach: The content of the organisation's website was retrieved in an automated manner using the url of the organisations.
A vector representation for the content of each organisation was learned using a natural language model. The similarity between two contents was calculated using cosine similarity.
To visualise the data, a similarity network was created using the similarity matrix. Each organisation represents a node und the similarities to all the other organisation represents the edges.
Further on a clustering was done in order to extract common topics trough the content of the organisations. For the determination of a kind of positions in the found topics, a sub clustering was made only with the vector representation of a cluster.

Result: The different organisations could be compared and plotted using similarity networks as shown in figure 1. The color of the nodes represents the cluster assigned to the organisation. The edges between the organisations represents the content based similarity of the organisations. To make the visualisation more comprehensive, a threshold for similarity was introduced. All edges with a similarity below the threshold were removed from the network. In figure 2 only the cluster 4 of figure 1 is shown. The sub clusters in figure 2 represents a kind of position in the cluster.
This work was about the developing of the method. A structured analysis of the data is still pending.

Fig. 1 - Similarity Network
Own presentment

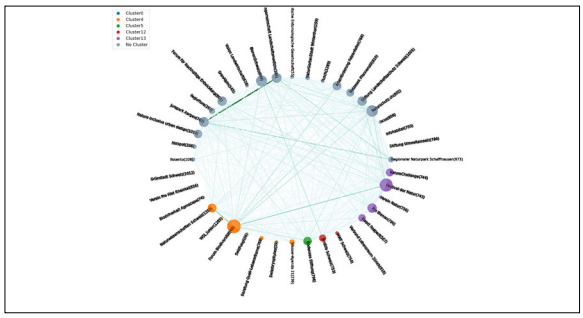
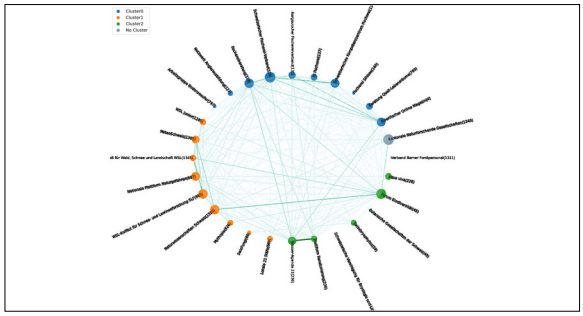


Fig. 2 - Similarity Network of Cluster 4 in Fig. 1 with sub cluster assignment.
Own presentment



Examiner
Prof. Dr. Guido Schuster

Subject Area
Data Science