Network Tour of Data Science

Spectral analysis of 5000 movies network

Macko Vladimir, Novakovic Milica, Pavué Clément, Roussaky Mehdi
Lausanne 2017

Goal:
Analysis of the graph representing a network of movies and their relations using descriptions of 5000 selected movies.

Data description:
The studied *TMDB 5000 Movie Dataset* contains information about 5000 selected movies provided by users and reviewers from *The Movie Database (TMDb)*. Namely, each of the selected movies has the following attributes: *budget, genres, homepage, id, keywords, original language, original title, overview, popularity, production companies, production coutries, release data, revenue* and personnel aspects of *cast* and *crew* members listing their names, genders and role or contribution to the movie production and other details. The majority of the data is in the text format. Dataset is available at the Kaggle web page.

Project description: The aim is to create a graph in which nodes represent movies and edges represent similarity between the movies they are connecting. Construction of edges and induced impact on below-described results will be examined. Since movie attributes are not numerical, it is necessary to calculate *overlap* matrix for each feature separately using a feature-specific *overlap* calculating function which will be developed. Generated series of *overlap* matrices will be combined into one single weight matrix. However, the choice of combining method is arbitrary. Therefore, several examples of combining method will be considered. Once the graph is created, e.g. the weight matrix is established, it will be sparced using *k-nearest neighbors* and Laplacian matrix and its corresponding eigendecomposition will be calculated. The constructed graph will be visualized using PCA and using Laplacian eigendecomposition in order to observe *clusters of similarity* in terms of genres, popularity etc. Furthermore, studying connectivity of the graph and degrees of nodes it is possible to estimate how *mainstream* a movie is. Also, a study of choice of a *category representative* may be conducted choosing a movie which represents the cluster in the best way. For example, if *k-nearest neighbors* method is utilized with k equal to 1, we can obtain disconnected graph with clusters of movies, and the node to which the most of the edges oriented point, is appointed as a *category representative*. Using a *category representative* allows a viewer to get an impression on entire category based on watching only a single movie.