Computitional social science - Assignment 4

JProf. Dr. Claudia Wagner
clwagner@uni-koblenz.de

Nada Beili
nbeili@uni-koblenz.de

June 24, 2020

Submission until: July 13, 2020 - 11:00 a.m.

Instructions:

1. Send your solution before the deadline to nbeili@uni-koblenz.de
2. Use as subject of the email "CSS Assignment 4 " + your full name and immatriculation number.
3. For the programming tasks, please do not add your code to the PDF. You need to submit only the .ipynb file. **The file name has to be the same as the email subject otherwise will not be accepted.**
4. Do not work in groups or copy from other students. The submissions of all students that are involved in a plagiarism case will not counted (independent of who copied from whom)
5. You are not allowed to use any library to do the most of the tasks
6. You are allowed to use Pandas, Networkx or Graph-tool, Matplotlib and/or Seaborn.

In this assignment, you are provided with a file 'netflix_titles.csv'. This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset contains the following features:

- show_id: A unique identifier for each movie.
- type: Identifier - A Movie or TV Show
- title: Title of the Movie / Tv Show
- director: Director of the Movie
- cast: Actors involved in the movie / show
- country : Country where the movie / show was produced

- date_added: Date it was added on Netflix
- realease_year: Actual Release year of the move / show
- rating: TV Rating of the movie / show
- duration: Total Duration - in minutes or number of seasons
- listed_in: Genre
- description:The summary description

Tasks:

1. **Data Preparation**:

   1.1. Check duplicates values in the dataset.

   1.2. Find the missing rows in each column.

   1.3. Remove the NaN values from the dataset.

   1.4. change the 'date_added' values to the only year. And rename the column: 'added_year'.

   1.5. Make a copy of the dataset.

   1.6. In the copied dataset,remove all the columns, except for the 'show-id', 'type' and 'cast'.

   1.7. In the copied dataset, you make sure that each artist will be in a row with the showid that they have participated in and the type (see the table below).

actor_name	show_id	type
Alan Marriott	81145628	Movie
Andrew Toth	81145628	Movie

2. **Data Discovering**:

   2.1. Plot the number of movies vs TV shows.

   2.2. Plot the number of releases per year. Give an interpretation.

   2.3. What is the highest year in adding new shows?

   2.4. What are the types of ratings for the movies? Plot for each rating the number of movies.

3. **Social Network analysis**:

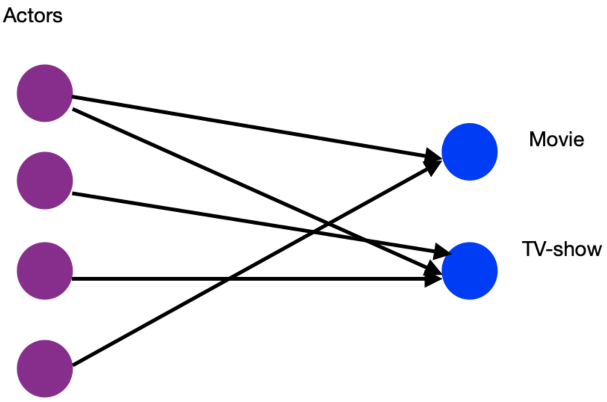
   3.1. Create a co-acting network, where the nodes are actors and the edges are the links between actors if they have participated in the same movie or TV-show at least once.

3.2. Compute the betweenness centrality, degree centrality , and closeness centrality for each node.

3.3. For each centrality measure, find the actor with the highest value.

3.4. Compute the average length of the shortest path and the average clustering coefficient.

3.5. Shuffle your data and take the first 100 rows and plot two modes network (see the figure below)


Actors
Movie
TV-show