Project 2

Suppose that you are a big fun of movies, and you try to do some analysis about movies. You find a dataset that includes information about 5000 movies, and you want to do some network analysis about it.

In this project, your specific tasks are as follows:

1. Given the IMDB 5000 Movie dataset, **create a network**. You should think about the metric to build a network. For example, a co-play network has actors as nodes and connections between two actors are determined based on that the two actors played the same movie(s).
2. **Find two subnetworks**. You are free to find any two subnetworks from the network that you have created, and each subnetwork should meet the following two requirements.
    - Including at least 20 nodes
    - All nodes in a subnetwork are **NOT** isolated, so each node should have at least one connection with other node(s).
3. **Determine similarity metrics**. You need to define a way of computing similarity between the two subnetworks by using the information given in this dataset. For example, based on the overall budget and movie genres, parts of the two subnetworks are similar, and their computed similarity score is 0.78.
4. **Interactively show the similarity**. You should design an interactive way of displaying the computed similarity between two subnetworks. For example, as a user selects a few nodes in subnetwork1, similar component(s) in subnetwork2 gets highlight, and a table pops up with necessary information to explain how the similarity is determined.

Your project should run in a web browser. You can use any web-based library, and make sure you give reference in the written document. You are encouraged to work on the project in a group. The movie dataset can be found in the attachment, and more detailed information about it can be found: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset.

Write a document with your discussion regarding to the above tasks. Specifically, your document should include discussions about the following important questions.

1. How do you build the network from this given dataset?
2. How do you find two subnetworks?
3. What is/are your metric(s) to compute similarity between the two subnetworks?
4. What is your design to show the similarity?
5. How to run your program?

The document should be in MS Word (Times New Roman font, size 12, single line spacing, no page limit).

Your submission should include both code and the document.