Homework 5

Due: 27 May, 2019, 11:59 PM (before midnight)

From IMDB's top rated 250 movies[1] choose one movie (maybe the one you like). Scrape the title of the movie, the summary text, and the urls for 12 similar movies from the "More Like This" section. For each of the 12 movies do the same (scrape the title, the summary text and the urls of their 12 similar movies). Do this for 3 layers. Layer 0: the movie you chose, Layer 1: the "More Like This" movies from layer 0 movies, Layer 2: "More Like This" movies from layer 1 movies, Layer 3: "More Like This" movies from layer 2 movies. This is what we did in class for scraping Wikipedia. The technique is called snowball sampling[2].

Part A

Construct a directed network using the data, where each node is a movie and each edge represents connection between nodes (if the movie was in the "More Like This" section). Create a Dash app where you will have the description of the project, 3 plotly plots and comments/conclusion. Each plot should be the visualization of the network, with appropriate labels and coloring (make sure that titles are visible and are not overlapping too much). The name of each node should be the title of the movie. The node_size parameter should be the in_degree, out_degree, and degree of the network (thus 3 plots). Next to each plot create a table showing the 5 movies with the highest measure (in_dergree, out_degree, or degree) and the magnitude of the measure. At the end write some comments/conclusion about the results. A self contained tutorial on network visualization with plotly can be found here: https://plot.ly/~empet/14683/networks-with-plotly/#/.

Part B

For each movie summary, convert to lowercase and remove the following symbols: "!\"#$ %&()*+-./:;<=>?@[\]^_`{|}~\n". Calculate the TF-IDF statistic. Fit kmeans clustering algorithm with 3 clusters to the result. Calculate the top 3 principal components from TF-IDF. Plot the 3 principal components as a 3d scatterplot using plotly. Points from the same cluster should have the same color. Write a small conclusion about the results.

Part A should be in a .py file and Part B in a .ipynb. You need to create dash app only for part A. The homework should be submitted in one .zip file.

1 https://www.imdb.com/chart/top?ref_=nv_mv_250
2 https://en.wikipedia.org/wiki/Snowball_sampling