Project 6: Fineli Nutritional Dataset Analysis [Mehrdad]

Consider the Fineli dataset, which contains detailed nutritional information for various ~4K food items. The dataset includes columns such as id, name, energy, calculated (kJ), fat, total (g), carbohydrate, available (g), protein, total (g), and many other nutritional attributes.

Tasks:

1. **Load the Fineli dataset** and preprocess the data to handle any missing values or inconsistencies. Ensure all numerical values are appropriately formatted for analysis.
2. **Generate a nutritional network graph** where nodes represent food items and edges represent similarity based on nutritional content. Define a similarity measure (e.g., cosine similarity) to determine the edges between nodes.
3. **Visualize and plot the degree distribution** of the nutritional network graph. Separate the degree centrality distribution, closeness distribution, and betweenness distribution. Draw three distinct degree distributions accordingly.
4. **Provide the script for drawing power law distributions** for degree, closeness, and betweenness distributions.
5. **Utilize the NetworkX clustering function to calculate the clustering coefficient** for each node within the graph. Generate a histogram with 10 bins based on the clustering coefficient values, displaying the count of nodes within each bin.
6. **Detect communities within the nutritional network** using the Girvan-Newman and Louvain community detection algorithms. Employ suitable plotting techniques to emphasize the distinct communities within the graph. Present a table summarizing key characteristics of each community, including the number of nodes, number of edges, diameter, average path length, and average degree.
7. **Analyze the nutritional composition of each community** by calculating the average values for key nutritional attributes (e.g., energy, fat, carbohydrates, protein). Create a summary table and visualize the differences between communities.
8. **Identify the top-10 most similar food items** within each community based on their nutritional content. Analyze the characteristics of these items and discuss any common trends or patterns.
9. **Convert the nutritional network into a weighted graph** where the weights represent the degree of similarity between food items. Repeat tasks 3-8 for this new weighted graph.
10. **Suggest appropriate metrics to compare the two graphs** (unweighted and weighted) and discuss the differences in their structural properties.
11. **Implement the PageRank algorithm** to identify influential food items within the network. Visualize the top 10 food items based on their PageRank scores.
12. **Analyze the assortativity of the network** to determine if food items with similar nutritional attributes tend to be connected. Provide a script to calculate and visualize the assortativity coefficient.
13. **Use the k-core decomposition technique** to identify the core structure of the network. Visualize the k-core subgraphs and analyze the properties of nodes within different cores.
14. **Implement the HITS algorithm** to identify hubs and authorities within the nutritional network. Visualize the top hubs and authorities.
15. **Analyze the network's robustness** by simulating node removal (both random and targeted) and observing the impact on network connectivity. Provide scripts and visualizations for these simulations.
16. **Identify relevant literature to provide analysis and discussion** on the findings, while also highlighting potential limitations of the study.