DATA 620 FINAL PROJECT PROPOSAL

Group Members:

- Dhanya Nair
- Dirk Hartog
- Gillian McGovern

Objective: Understanding Product Influence and Customer Behavior through Co-Purchase Networks and NLP-Based Product Analysis.

Guiding question:

How do product characteristics, customer sentiment, and textual similarity influence co-purchasing behavior for Digital_Music on Amazon?

Motivation:

This guiding question allows us to investigate how Amazon recommends their Digital_Music products to customers and gives us experience comparing centrality measures and sentiment metrics across Amazon products.

Project Goals:

- Identify Amazon's most popular products in Digital_Music category.
- Compare centrality measures across Amazon product categories
- Perform topic modeling on product descriptions to group products by latent topics.
- Derive sentiment labels and confidence scores.
- Use sentiment analysis to score product reviews.
- Assign average sentiment per product node and analyze how sentiment relates to centrality (e.g., do more central products have higher sentiment?).
- Identify polarizing products (high centrality, low sentiment).

Data Sources:

Source: https://amazon-reviews-2023.github.io/

This is a large-scale **Amazon Reviews** dataset, collected in **2023** by McAuley Lab, and it includes rich features such as:

1. **User Reviews** (*ratings*, *text*, *helpfulness votes*, etc.);
2. **Item Metadata** (*descriptions*, *price*, *raw image*, etc.);
3. **Links** (*user-item* / *bought together* graphs).

Planned Workflow:

- Read in the data (text file stored on GitHub)
    - Create subset of data
- Read in metadata
- Basic analysis
- Visualize the network
- For each of the nodes in the dataset, calculate:
    - Degree centrality
    - Eigenvector centrality
    - Betweenness centrality
    - Closeness centrality
- Compare centrality measures across Amazon product categories and customers
- Derive sentiment labels and confidence scores
- Present findings (Jupyter Notebook report + GitHub)

Team Work:

Each team member will handle part of the workflow: data set up, graph visualizations, network metrics & comparisons, Text mining, NLP, findings & conclusion