**<u>Project Proposal</u> (Group: G5)**

**Comparing Louvain Modularity, Leiden Modularity, and Dynamic Programming Algorithms for Graph Based News Article Clustering**

**Project Description:**

This project aims to compare the effectiveness and efficiency of two distinct approaches for clustering news articles: the Louvain Modularity, Leiden Modularity method and Dynamic Programming algorithms. Clustering news articles is vital for content organization, recommendation systems, and information retrieval. By conducting a comprehensive analysis, we will determine which method is better suited for this task.

**Project Plan:**

1. *Data Preparation:*

- We will begin by obtaining a dataset of news articles from reputable sources.
- Data preprocessing steps will include tokenization, lowercasing, stop word removal,and TF-IDF vectorization to prepare the text data for clustering.

2. *Graph Construction:*

- We will construct a weighted graph based on semantic similarity using the TF-IDF vectors of the articles.
- Edges in the graph will represent the strength of the semantic relationship between articles.

3. *Clustering Algorithms:*

- We will apply the Louvain Modularity algorithm [1] to the graph to perform clustering based on modularity optimization. We plan to compare this modularity algorithm with Leiden Modularity algorithm [2] by implementing it.
- Additionally, we plan to implement the Dynamic Programming algorithm [3], defining specific objectives and constraints for clustering.
- Both algorithms will be applied to the same graph to ensure a fair comparison.

4. *Evaluation:*

- To evaluate the concentration of relationships within clusters compared with a random distribution of relationships, we will use metrics such as modularity and the runtime complexities.
- The results from both algorithms will be compared and statistically analyzed (if necessary) to determine significant differences in clustering quality.

**Bibliography:**

[1] Blondel Ghosh, S., Halappanavar, M., Tumeo, A., & Kalyanarainan, A. (2019). Scaling and Quality of Modularity Optimization Methods for Graph Clustering. *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, 1-6.

[2]Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

[3] Li, Y., Zhao, X. & Qu, Z. (2020). A Dynamic Programming Framework for Large-Scale Online Clustering on Graphs. *Neural Processing Letters*, 52, 1613–1629.

[4] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[5] Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), Article 026113

**Programming Language:**

Python will be the primary programming language used for this project. Python offers a wide range of libraries and tools for natural language processing (NLTK, spaCy), graph analysis (NetworkX), data manipulation (Pandas), and visualization (Matplotlib), making it well-suited for the implementation of the algorithms and data analysis.

**Project Outcomes:**

By the end of this project, we aim to have a comprehensive comparison of the Louvain Modularity, Leiden modularity, and Dynamic Programming algorithms in the context of news article clustering. This comparison will provide insights into their strengths and weaknesses, potentially guiding their use in applications for organizing news content.

**Team Members**

- Anurag Deotale
- Sanket Kulkarni