PROJECT TIMELINE FOR NEWS HEADLINE CLUSTERING

DEWANSH SINGH CHANDEL

Data Collection and Preprocessing

- **Tools**: Use **BeautifulSoup**, **Scrapy, Selenium** for scraping Google News Hindi headlines.
- **Data Structuring**: Format data into a structured dataset with headline in text in one column and article link in other column.
- **Text Cleaning**: Apply preprocessing techniques such as stopword removal, tokenization, and lemmatization using **NLTK** or **spaCy**.(required when doing TF-IDF embedding)
- **Dataset Preparation**: Use libraries like **Pandas** to organize and prepare the dataset for training and evaluation.

Model Development

- **Text Representation**:
    - Implement **TF-IDF** or **Bag of Words (BoW)** for initial text vectorization.
    - Can also use **Sentence-BERT** embeddings from **Hugging Face Transformers**
- **Clustering Algorithms**:
    - Begin with unsupervised methods such as **K-Means** or **K-means++** with the help of **Pytorch** and **scikit-learn**.
- **Evaluation Metrics**: Use cosine similarity, silhouette scores, and cluster purity metrics to assess the model's performance.

Visualization and Analysis

- **Dimensionality Reduction**: Apply **t-SNE** or **PCA** for 2D/3D visualization of clusters.
- **Visualization Tools**: Use **Matplotlib**, **Seaborn** for visual representation of clustering patterns and performance.
- **Insights Reporting**: Generate detailed reports on clustering accuracy, notable patterns, and areas for improvement.