Project Plan

Project name: IntelliBrief News Summarization Project

Project members:

- Muhammad Sulaiman Javed
- Syed Fakhar Abbas Naqvi
- Md Murad Alahi Misu
- Zeshan Hyder

Problem statement:

With the increasing availability of digital news content from multiple sources, it has become difficult for individuals to stay updated efficiently. Manual summarization is time-consuming and often subjective. Therefore, there is a need for an automated system that can generate accurate and concise news summaries, making it easier for users to consume relevant information quickly.

Objectives:

Our goal is to develop a robust news summarization system that can efficiently process a large volume of articles and generate coherent, unbiased summaries. We aim to achieve strong classification performance and meaningful summaries. Success criteria (planned expectations):

- Achieving reasonable classification accuracy.
- Generating summaries with acceptable quality based on standard metrics like ROUGE scores.

Data:

We plan to use an existing dataset of news articles available on Kaggle, covering multiple categories such as business, entertainment, politics, sport, and technology.
Main steps planned:

- Preprocessing the text (cleaning, tokenization, removing stopwords).

- Feature extraction using techniques like Bag-of-Words and word embeddings.
- Dividing the data into training and testing sets.

Methodology:

We intend to combine data science and NLP techniques to build the system. Planned methods include:

- Text Representation: Bag-of-Words and Word Embeddings (such as CBOW and Skip-gram models).
- Summarization Approach: Graph-based summarization techniques.
- Classification Algorithms:
    - Navie Bayes Classifier
    - Decision Tree Classifier

Note: We may also explore comparative analysis between models to select the best-performing approach.

Evaluation:

We plan to evaluate the performance of our system using:

- For classification: Accuracy, confusion matrix.
- For summarization: ROUGE-1, ROUGE-2, and ROUGE-L scores (Precision, Recall, F1-score).

We expect to set target benchmarks after some initial model experiments.

Expected challenges:

- Data Quality: Ensuring the data is clean and diverse enough for training robust models.
- Summarization Quality: Generating summaries that are both concise and preserve essential information.

- Class Imbalance: Some news categories might have fewer examples.
- Computational Resources: Training word embeddings and classifiers may require significant time.

We plan to mitigate these challenges by careful preprocessing, data augmentation if necessary, and tuning model hyperparameters.

Resources and tools:

- Programming Language: Python
- Libraries and Tools:
    - NLTK
    - scikit-learn
    - Gensim (for Word2Vec)
    - Matplotlib (for result visualization)
- Environment: Initially, development will be done on local machines. Each member is supposed to commit to the main Colabs notebook.

Questions for further guidance:

- Should we explore and use ML models not covered in this course but are befitting for our project?