Natural Language Processing Final Project Group Proposal
Kunal Inglunkar
Lakshmi Sravya

What problem did you select and why did you select it?

We selected the news summarization problem because it addresses a real-world need and presents a compelling challenge in natural language processing. News summarization involves automatically condensing lengthy news articles or documents into shorter, coherent summaries, extracting the most crucial information while preserving the key facts and context. The reasons for selecting this problem are as follows:

- Relevance: News summarization is highly relevant in today's information-rich world. It helps individuals stay informed about current events without the need to read entire articles, which can be time-consuming.
- Information Overload: With the constant flow of news and information, there is a pressing need for automated tools that can extract and deliver essential news content efficiently. News summarization helps address the problem of information overload.
- Diverse Applications: News summarization has applications in various fields, including journalism, content curation, and information retrieval. It can also benefit individuals who have limited time to keep up with news.

What database/dataset will you use?

https://huggingface.co/datasets/multi_news

Multi-News, consists of news articles and human-written summaries of these articles from the site newser.com. Each summary is professionally written by editors and includes links to the original articles cited.

What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?

We will be using Seq2Seq which is considered a classical model. We'll also implement one pretrained model

What packages are you planning to use? Why?

- PyTorch
  These deep learning frameworks provide the foundation for building and training Seq2Seq models, as well as other NLP models.
- NLTK (Natural Language Toolkit) and spaCy
  These packages are useful for text preprocessing, tokenization, and various text analysis tasks. They help with data cleaning and initial text preparation.
- Scikit-learn

Scikit-learn provides a wide range of machine learning tools and algorithms that can be helpful for any rule-based or traditional machine learning components you plan to use alongside deep learning models.

- Numpy and Pandas
  These libraries are indispensable for data manipulation, handling tabular data, and performing various numerical operations.
- Matplotlib and Seaborn:
  These visualization libraries enable you to create informative plots and charts to better understand your data and model performance.

What NLP tasks will you work on?

- Text Preprocessing:
- Text Summarization:
- Sequence-to-Sequence (Seq2Seq) Modeling:
- Transformer-Based Modeling:
- Fine-Tuning Pretrained Models:
- Model Deployment

How will you judge the performance of the model? What metrics will you use?

- F1 Score

A rough schedule for completing the project.

- Week 1: Project Setup and Data Collection
- Week 2: Data Preprocessing and Model Selection\
- Week 3: Model Training
- Week 4: Model Evaluation and Interpretability
- Week 5: Documentation, Reporting, and Final Evaluation