Project Requirement

Project Title: Email Subject Line Generation

This project focuses on the task of generating concise and informative email subject lines from the email body text. Unlike conventional tasks such as news summarization or headline generation, email subject line generation presents a unique challenge due to the requirement of producing extremely short, yet meaningful summaries—often limited to just a few words. The task demands identifying key information from the body of an email and distilling it into a compact subject line.

From an implementation standpoint, this project explores the use of generative Natural Language Processing (NLP) models. It also delves into various text generation evaluation strategies, emphasizing how to assess the quality of short-form abstractive outputs.

Reference Paper: This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation

Reference project document: https://aclanthology.org/P19-1043.pdf

Dataset link: The Annotated Enron Subject Line Corpus: https://github.com/ryanzhumich/AESLC

Dataset Description:

- The dataset is a curated subset of the Enron Email Corpus, which includes cleaned, filtered, and deduplicated emails from employee inboxes at the Enron Corporation.
- It is annotated for subject line generation, with up to three human-generated subjects lines per email in the dev/test sets, allowing for robust multi-reference evaluation.

Key Statistics:

- Training / Development / Test splits:
    - Train: 14,436
    - Dev: 1,960
    - Test: 1,906
- Average length of an email body: ~75 words
- Average length of a subject line: ~4 words

Tools: Hugging Face, PyTorch, Tensorflow, Keras, WandB, NLTK

Deployment: FastAPI, Cloud Application Platform | Heroku, Streamlit, Cloud Computing, Hosting Services, and APIs | Google Cloud

Final Submissions:

- Comprehensive technical report detailing methodology, experiments, and outcomes
- Final presentation outlining the problem statement, approach, and results
- Modeling overview covering techniques such as:
    - Sequence-to-sequence transformers (e.g., BART, T5)
    - Fine-tuning strategies and training optimizations
    - Evaluation metrics (e.g., ROUGE, BLEU, METEOR, BERTScore)
- GitHub Repository with complete codebase, instructions, and usage demos