Group Project Proposal

Team Members:

- Shikha Kumari
- Kalyani Vinayagam Sivasundari

Question - Answering System

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

We selected the problem of question answering using the SQuAD (Stanford Question Answering Dataset) dataset. We chose this problem because question answering is a fundamental task in natural language understanding, and SQuAD is a widely used benchmark dataset for this task. It consists of questions posed on a set of Wikipedia articles, with corresponding answer spans within the text. Solving this problem effectively requires understanding and extracting relevant information from passages to generate accurate answers, which aligns well with the capabilities of NLP models.

What database/dataset will you use?

We will use the SQuAD dataset, which is freely available and widely used for question answering research and benchmarking. The dataset contains questions, passages, and corresponding answer spans from Wikipedia articles.

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 explore using pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly optimized BERT approach), or similar architectures fine-tuned specifically for question answering tasks.

What packages are you planning to use? Why?

We plan to use popular NLP libraries such as Hugging Face's Transformers library for accessing pre-trained models like BERT or RoBERTa, as well as TensorFlow or PyTorch for model training and evaluation. These libraries offer easy-to-use interfaces for working with pre-trained models and implementing custom architectures if necessary.

What NLP tasks will you work on?

The primary NLP task for this project will be question answering, specifically extracting answers from passages given a question. This involves tasks such as passage understanding, question understanding, and answer extraction.

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

We will evaluate the performance of the model using standard metrics for question answering tasks, such as Exact Match (EM) and F1 score. The Exact Match metric measures the percentage of predictions that exactly match the ground truth answers, while the F1 score computes the harmonic mean of precision and recall based on token-level overlap between predicted and true answers. These metrics provide a comprehensive assessment of the model's ability to generate accurate answers.

Provide a rough schedule for completing the project.

- **Week 1: Preprocessing & Exploration** - Load SQuAD data, tokenize passages, questions, and answers for model input.
- **Week 2: Model Selection & Setup** - Choose BERT, RoBERTa, or similar architectures. Fine-tune selected model on SQuAD dataset.
- **Week 3: Training & Optimization** - Experiment with hyperparameters, monitor training using validation metrics.
- **Week 4: Evaluation & Fine-tuning** - Assess model performance using Exact Match & F1 score on test set. Fine-tune if needed. Prepare project report documenting methodology, results, and conclusions.