**Project Proposal: Building a Data Science Answering Bot**

Arjun, HaeLee, Nayauen

**Problem Selection:**

Our project aims to address the need for an intelligent chatbot capable of providing answers related to data science topics. This chatbot will assist users in finding information, answering questions, and providing guidance on data science concepts. We selected this problem to create a valuable tool for individuals seeking data science knowledge and support.

**Dataset:**

We will use a combination of publicly available data science-related text data, textbooks, research papers, and online resources to create a comprehensive knowledge base for the chatbot. This dataset will include a variety of data science topics, concepts, and questions.

**NLP Methods:**

Text preprocessing for data cleaning, tokenization, and stemming.

Named Entity Recognition (NER) to identify data science concepts and entities.

Text classification for categorizing questions and answers.

Sequence-to-Sequence (Seq2Seq) models for generating responses.

Transformer-based models for understanding context and generating coherent responses.

We may also customize and fine-tune pre-trained language models like GPT-3 to enhance the chatbot's performance.

**Packages:**

NLTK and spaCy for text preprocessing and NER.

TensorFlow or PyTorch for building and training NLP models.

Hugging Face Transformers library for pre-trained language models.

Flask or Django for creating a user interface.

**NLP Tasks:**

The NLP tasks include text preprocessing, NER, text classification, and sequence generation for chatbot responses.

**Performance Evaluation:**

Accuracy for question classification.

BLEU score or ROUGE score for response generation.

**Project Schedule:**

Week 1: Data collection and preprocessing. Implement NER and text classification models.

Week 2: Develop and fine-tune the Seq2Seq model.

Week 3: Integrate a pre-trained language model and build the chatbot interface.

Week 4: Testing, debugging, and user feedback collection.

Week 5: Final refinements and project submission.

Our team believes that this data science answering bot will provide valuable assistance to individuals in the field of data science and improve their access to relevant information and guidance.