Gen-AI Bootcamp 24
Natural Language Processing Course Assignment

Part 1: Text Collection and Loading

Objective: *Collect and load a text dataset from a selected domain into a suitable format for processing.*

Tasks:

- Choose a domain you are interested in (e.g., healthcare, sports, e-commerce).
- Identify a dataset from your chosen domain. Some sources for datasets include Kaggle, UCI Machine Learning Repository, government open data portals, and APIs from relevant websites.
- Write code to load your dataset into a suitable format for processing, such as CSV, JSON, or plain text files.
- Ensure the dataset is loaded correctly by displaying the first few rows.

Part 2: Text Preprocessing

Objective: *Gain hands-on experience with text preprocessing techniques.*

Tasks:

- Choose a text corpus from the NLTK library, which is publicly available and widely used in text processing.
- Perform the following preprocessing steps on the corpus:
- Tokenization: Split the text into words and sentences.
- Stemming: Reduce words to their root form.
- Lemmatization: Further reduce the stemmed words by considering their context.
- Stop Word Removal: Eliminate common words that may not be useful for analysis.
- Each step should be commented properly.
- Discuss the impact of each preprocessing step on the corpus.

Part 3: Feature Extraction Techniques

Objective: *Understand and apply text data transformation into machine-readable vectors.*

Tasks:

- Using the pre-processed text from **Part 2**, implement the following feature extraction methods:
- Bag-of-words
- TF-IDF
- n-grams
- When one method might be preferred over the others? Explain the reason.
- Visualize the most common terms with each method using a word cloud.

Part 4: Word Embeddings

Objective: *Explore word embeddings and their applications.*

Tasks:

- Using a pre-trained model from Gensim's Data repository (https://github.com/piskvorky/gensim-data), apply Word2Vec, GloVe, and FastText embeddings to a sample text and your dataset as well.
- Visualize the word embeddings using t-SNE to see clusters of similar words.

Part 5: Model Training and Evaluation

Objective: *Understand RNNs and their ability to handle sequence data.*

Tasks:

- Utilize the TensorFlow and Keras libraries to construct a simple RNN model.
- Train an LSTM and a GRU network on a your problem using the loaded dataset.
- Compare the performance of LSTM and GRU networks on the task.
- Analyze the long-term dependencies captured by each model

Part 6: Visualization and Interpretation

Objective: *Visualize and interpret the results to gain insights from the model's performance.*

Tasks:

- Use visualizations to understand the data and model outputs (e.g., word clouds, confusion matrices).
- Provide human-readable interpretations of the model's predictions and decisions. Discuss the implications of the findings in the context of the selected domain.

Submission Guidelines:

- Submit a detailed report including code, outputs, and your analysis of the results. The assignment is due by [insert date here].