Assignment-2

Deadline: 11:59 PM, 8th October 2024

Objective:

In this assignment, you will work with text data from different sources to build models for text summarization using Recurrent Neural Networks (RNNs). You will then apply the trained models to summarize on a new dataset and perform qualitative analysis based on your findings.

Part 1: Data Exploration and Model Building (55 Marks)

Task 1: Data Exploration with CNN/Daily Mail Dataset (15 Marks)

- **Description**: You need to perform an exploratory data analysis on the CNN/Daily Mail dataset to better understand its structure.
  - Inspect the data structure (headlines and stories).
  - Explore the most frequent words and sentence lengths.
  - Perform basic text cleaning (lowercasing, tokenization, removing special characters).
  - Provide visualizations of word frequency distributions and sentence lengths.

  *** For more information on how to load dataset from huggingface refer here**

Task 2: Build and Train Seq2Seq Model for Text Summarization (40 Marks)

- **Description**: You are required to implement a Seq2Seq model using RNNs (or LSTMs) to summarize the texts in the CNN/Daily Mail dataset.
  - **Step 1**: Preprocess the text (tokenization, padding, and vocabulary creation).
  - **Step 2**: Build the Seq2Seq architecture (encoder, decoder, optional attention mechanism). You are free to choose the hyperparameters.
  - **Step 3**: Train the model on the CNN/Daily Mail dataset (train split) and evaluate (on test split) using ROUGE scores (use Rouge-2 and Rouge-L).

Expected Outcome: The trained model should be able to summarize unseen articles with reasonable accuracy.

Part 2: Applying the Model to New Data (20 Marks)

Task 3: Test the Model on Wikipedia Summary Data

- **Description**: Using the same Seq2Seq model trained previously, test the model on the Wikipedia summary dataset (consider the first 10k rows as test data).
    - **Note**: No need for data exploration here. Preprocess the data and report the evaluation results.

Deliverables:

1. **Code (75)**: Submit a well-structured Python notebook with all code, including data preprocessing, model training, and summarization for both datasets.
2. **Report (25)**: Provide a comprehensive report discussing:
    - Data exploration findings.
    - Model training and evaluation results.
    - Observations