Project Proposal

Text Recovery for Corrupted Sequence Restoration

Team: Toni Grgurević, Marin Bogešić

Dataset:
The dataset consists of pairs of text sequences, where one sequence is the original text, and the other is a corrupted version with certain characters randomly replaced by the # symbol. These pairs serve as the primary training data for the project.

Project Idea:
This project aims to develop a machine learning model capable of restoring corrupted text sequences by predicting missing characters. Using the corrupted-original text pairs as training data, the task is framed as a sequence prediction problem, focusing on capturing dependencies across characters.
We will initially explore statistical approaches such as:

- **n-gram models**
- **Hidden Markov Models (HMMs)**

These models predict missing characters based on probabilistic contextual patterns. Furthermore, we will investigate the application of **Long Short-Term Memory (LSTM) networks**, a type of recurrent neural network (RNN), which excels at capturing longer-range dependencies in sequential data.

Performance Metrics:
To evaluate the model, we will use:

- **Character-level accuracy**
- **BLEU score** (for sequence similarity)
- **Levenshtein distance** (for error quantification).

Software to be Developed:

1. **Data processing scripts**
    - Format and preprocess input sequences for model training.
2. **Model training scripts**
    - Develop statistical models (HMM and n-gram) using libraries like hmmlearn and nltk.
    - Implement LSTM-based models for sequence prediction.
3. **Evaluation scripts**
    - Assess model performance using accuracy, BLEU score, and Levenshtein distance.

Relevant Papers:

1. Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press.
2. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
3. Rabiner, L. R. (1989). *"A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition."* Proceedings of the IEEE.
4. Olah, C. (2015). *"Understanding LSTM Networks."*