EXERCISE 3

Neural Networks for NLP

1. Thông tin chung

Môn học	NLP501 - Natural Language Processing
Trọng số	10% tổng điểm
Ngày phát	Cuối buổi 8
Hạn nộp	1-2 tuần
Hình thức	Cá nhân

2. Yêu cầu

Part A: Sentiment Analysis with RNN/LSTM

Xây dựng mô hình phân loại cảm xúc sử dụng Recurrent Neural Networks.

Yêu cầu:

1. Tiền xử lý dữ liệu: tokenization, padding, vocabulary building
2. Triển khai embedding layer (random hoặc pre-trained)
3. Xây dựng mô hình với LSTM hoặc GRU layers
4. Huấn luyện với appropriate loss và optimizer
5. Vẽ training curves (loss, accuracy)
6. Đánh giá trên test set với classification metrics

Dataset:

- IMDB Movie Reviews (50,000 reviews)
- Hoặc SST-2 (Stanford Sentiment Treebank)

Part B: Named Entity Recognition

Xây dựng NER model để nhận diện entities trong văn bản.

Yêu cầu:

1. Load và preprocess CoNLL-2003 hoặc OntoNotes dataset
2. Xây dựng sequence labeling model (BiLSTM)
3. Triển khai BIO/IOB tagging scheme
4. Đánh giá với entity-level F1 score
5. Phân tích errors theo entity types (PER, LOC, ORG)

Trong đó:

- Nhận diện ít nhất 3 loại entities
- Sử dụng BiLSTM hoặc BiGRU architecture
- Optional: thêm CRF layer cho sequence decoding

Part C: Siamese Network for Question Similarity

Xây dựng Siamese Network để phát hiện câu hỏi duplicate.

Yêu cầu:

1. Load Quora Question Pairs dataset
2. Triển khai Siamese architecture với shared weights
3. Sử dụng LSTM/GRU encoder cho mỗi câu hỏi
4. Triển khai similarity function (cosine, Manhattan, etc.)
5. Huấn luyện với contrastive loss hoặc binary cross-entropy
6. Đánh giá với accuracy và AUC-ROC

3. Sản phẩm nộp

1. full source code
2. Báo cáo (3-4 trang) với training curves và evaluation metrics
3. Trained model weights (upload lên Google Drive nếu file lớn)

4. Tiêu chí đánh giá

Tiêu chí	Điểm	Mô tả
Part A - Sentiment với RNN	**3.5**	Model hoạt động
Part B - NER	**3.5**	F1-score hợp lý, BIO tagging đúng
Part C - Siamese Network	**2.0**	Architecture đúng
Code Quality & Documentation	**1.0**	Code + documents

Tổng điểm: 10.0 (quy về thang 10%)

5. Yêu cầu môi trường

- Python 3.8+
- TensorFlow 2.x hoặc PyTorch 1.x (tùy chọn)
- GPU recommended (có thể sử dụng Google Colab)
- Libraries: numpy, pandas, matplotlib, seaborn