EXERCISE 2

Probabilistic Models & Language Modeling

1. Thông tin chung

Môn học	NLP501 - Natural Language Processing
Trọng số	10% tổng điểm
Ngày phát	Cuối buổi 5
Hạn nộp	Đầu buổi 7 (2 tuần)
Hình thức	Cá nhân

2. Yêu cầu chi tiết

Phần 1: Autocorrect System

Xây dựng hệ thống sửa lỗi chính tả tự động sử dụng edit distance.

Yêu cầu:

1. Triển khai Minimum Edit Distance (Levenshtein Distance)
2. Xây dựng vocabulary từ corpus tiếng Anh
3. Tính word frequency để xếp hạng suggestions
4. Kết hợp edit distance và frequency để chọn best correction
5. Tạo interactive demo (command line interface)

Trong đó:

- Hỗ trợ edit distance từ 1-2
- Trả về top 5 suggestions cho mỗi misspelled word
- Xử lý được cả lowercase và uppercase

Phần 2: POS Tagging with HMM

Xây dựng Part-of-Speech Tagger sử dụng Hidden Markov Model.

Yêu cầu:

1. Tính transition probabilities $P(tag_i \mid tag_{i-1})$
2. Tính emission probabilities P(word | tag)
3. Triển khai Viterbi algorithm để decode
4. Xử lý *unknown words* với smoothing
5. Đánh giá accuracy trên test set

Dataset:

- WSJ corpus (Wall Street Journal) từ NLTK
- Hoặc Brown corpus với simplified tagset

Phần 3: N-gram Language Model

Xây dựng N-gram Language Model cho autocomplete.

Yêu cầu:

1. Xây dựng unigram, bigram, và trigram models
2. Triển khai smoothing (Add-k)
3. Tính perplexity trên test set
4. Xây dựng autocomplete function dự đoán next word
5. So sánh hiệu quả giữa các n-gram orders

3. Sản phẩm nộp

Jupyter Notebook (.ipynb) với đầy đủ code và documentation

4. Tiêu chí đánh giá

Tiêu chí	Điểm	Mô tả
Autocorrect	3	Edit distance đúng, suggestions hợp lý
POS Tagging	3	HMM và Viterbi triển khai đúng
Language Model	2	N-gram model hoạt động, perplexity hợp lý
Code	2	Rõ ràng, đầy đủ

Tổng điểm: 10.0