EXERCISE 1

Sentiment Analysis & Vector Spaces

1. Thông tin chung

Môn học	NLP501 - Natural Language Processing
Trọng số	10% tổng điểm
Ngày phát	Cuối buổi 3
Hạn nộp	Đầu buổi 5 (2 tuần)
Hình thức	Cá nhân

2. Mục đích:

- Hiểu và triển khai các thuật toán phân loại văn bản cơ bản (Logistic Regression, Naïve Bayes)
- Xây dựng và sử dụng word vectors để tính toán độ tương đồng ngữ nghĩa
- Áp dụng TF-IDF và cosine similarity trong bài toán tìm kiếm văn bản
- Đánh giá và phân tích kết quả của các mô hình NLP

3. Yêu cầu chi tiết

Phần 1: Sentiment Classification (40%)

Xây dựng bộ phân loại cảm xúc (positive/negative) sử dụng Logistic Regression hoặc Naïve Bayes.

Yêu cầu:

1. Chọn và tiền xử lý dataset (Twitter Sentiment hoặc IMDB Reviews, hoặc bất kỳ dữ liệu văn bản)
2. Triển khai feature extraction: word frequency, TF-IDF features
3. Huấn luyện mô hình Logistic Regression HOẶC Naïve Bayes
4. Đánh giá với accuracy, precision, recall, F1-score
5. Phân tích confusion matrix và các lỗi phổ biến

Trong đó:

- Chia train/test set với tỷ lệ 80/20 hoặc 70/30
- Tiền xử lý: lowercase, remove punctuation, tokenization
- So sánh ít nhất 2 phương pháp feature extraction

Phần 2: Word Vectors & Similarity (30%)

Xây dựng word vectors từ corpus và tính toán độ tương đồng giữa các từ.

Yêu cầu:

1. Xây dựng co-occurrence matrix từ corpus
2. Áp dụng PCA để giảm chiều word vectors
3. Tính cosine similarity giữa các cặp từ

4. Visualize word embeddings trong không gian 2D
5. Thực hiện word analogy tests (king - man + woman = queen)

Trong đó:

- Sử dụng context window size (3, 5, etc.)
- Áp dụng PCA giảm về số chiều nhỏ hơn (30, 50, etc.)
- Test với ít nhất 10 cặp từ có quan hệ ngữ nghĩa

Phần 3: Document Search (30%)

Triển khai hệ thống tìm kiếm văn bản đơn giản sử dụng TF-IDF.

Nhiệm vụ:

1. Xây dựng TF-IDF vectors cho tập documents
2. Triển khai search function với cosine similarity
3. Xếp hạng kết quả theo độ liên quan
4. Tạo simple interface (command line hoặc notebook)
5. Đánh giá cho các query mẫu

Dataset gợi ý:

- AG News Dataset (news articles)
- 20 Newsgroups Dataset
- Tự tạo corpus từ Wikipedia articles

4. Sản phẩm nộp

1. Jupyter Notebook (.ipynb) với đầy đủ code và giải thích
2. Báo cáo ngắn (PDF/Docs) gồm: phương pháp, kết quả, phân tích
3. File requirements.txt

5. Tiêu chí đánh giá

Tiêu chí	Điểm	Mô tả
Kết quả code	**3.0**	Code chạy đúng, kết quả hợp lý
Chất lượng code	**2.0**	Code sạch, có comments, tổ chức tốt
Phân tích kết quả	**2.0**	Giải thích kết quả, nhận xét sâu sắc
Báo cáo	**2.0**	Trình bày rõ ràng, đầy đủ thông tin
Sáng tạo	**1.0**	Có cải tiến hoặc thử nghiệm thêm

Tổng điểm: 10.0

6. Lưu ý

- Được phép sử dụng thư viện: numpy, pandas, sklearn, nltk, gensim
- Không sử dụng pre-trained embeddings cho Part B
- Bài nộp trễ sẽ bị trừ 20% điểm mỗi ngày