**THỰC HÀNH BUỔI 6**

**BÀI TẬP TỔNG HỢP**

|  |
| --- |
| Họ tên và MSSV:  Lớp: Ngày học: |

Cho đoạn văn bản sau:

"Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed. ^-^"

## **BÀI TẬP 1: TIỀN XỬ LÝ VĂN BẢN (BẮT BUỘC)**

Viết code để thực hiện tiền xử lý văn bản trên đoạn văn bản trên, bao gồm các bước:

* Chuyển tất cả chữ về lowercase.
* Tokenization: Chia văn bản thành các từ riêng lẻ.
* Loại bỏ stopwords.
* Lemmatization để đưa từ về dạng nguyên thể.
* Loại bỏ dấu câu và ký tự đặc biệt.

### **BÀI TẬP 2: **XÂY DỰNG MÔ HÌNH DỰ ĐOÁN** (Chọn một trong các phương pháp sau)**

#### ****Option 1: Sử dụng N-Grams để dự đoán từ tiếp theo****

* Xây dựng mô hình bigram hoặc trigram từ đoạn văn bản trên.
* Viết code để dự đoán từ tiếp theo sau một từ hoặc một cụm từ cho trước.

#### Ví dụ: Nếu nhập vào **"machine learning"**, chương trình có thể dự đoán từ tiếp theo là **"is"**.

#### ****Option 2: Phân loại văn bản bằng Naive Bayes****

* Giả sử đoạn văn bản trên thuộc một tập dữ liệu lớn với hai nhãn: "Tech" (công nghệ) và "Non-Tech" (không thuộc công nghệ)
* Huấn luyện mô hình Naive Bayes với một tập dữ liệu nhỏ để phân loại văn bản.
* Dự đoán xem đoạn văn bản trên có thuộc nhóm "Tech" hay không.

#### ****Option 3: Phân loại văn bản bằng K-Nearest Neighbors (KNN)****

* Chuyển đổi đoạn văn bản trên thành một vector đặc trưng bằng phương pháp TF-IDF hoặc Bag-of-Words.
* Sử dụng mô hình KNN để phân loại văn bản vào một trong hai nhóm: "Tech" hoặc "Non-Tech".
* Thử nghiệm với các giá trị k khác nhau và đánh giá độ chính xác.

Sinh viên nhập câu trả lời tại đây

|  |
| --- |
| **FILE PHẢI NỘP CHO BÀI THỰC HÀNH NÀY** |
| **Một file Word** chứa đáp án các yêu cầu |

*--- Hết ---*