**BÁO CÁO: PHÂN LOẠI VĂN BẢN & SỐ VIẾT TAY BẰNG SVM**

**1. Phân tích bài toán**

**1.1. Yêu cầu bài toán**

Bài toán đặt ra là xây dựng mô hình phân loại văn bản tiếng Việt và phân loại chữ số viết tay bằng phương pháp SVM (Support Vector Machine).

Mô hình sẽ được huấn luyện trên tập dữ liệu đã thu thập, sau đó sử dụng để dự đoán nhãn của các văn bản/số viết tay mới.

**1.2. Giới thiệu dữ liệu**

**Dữ liệu văn bản**

Dữ liệu được thu thập từ các bài báo trên VnExpress, được chia thành các danh mục khác nhau như:

* **Kinh tế**
* **Pháp luật**
* **Giải trí**
* **Giáo dục**
* **Thể thao**, v.v.

A screenshot of a computer

AI-generated content may be incorrect.

Số lượng văn bản trong từng danh mục:

A screenshot of a computer code

AI-generated content may be incorrect.

Kết quả :

A screenshot of a computer

AI-generated content may be incorrect.

Dữ liệu được tải vào chương trình bằng load\_files và có Kết quả :

A screenshot of a computer code

AI-generated content may be incorrect.

**2. Phân tích thuật toán**

**2.1. Ý tưởng**

* **Bước 1:** Tiền xử lý văn bản (loại bỏ từ dừng, chuyển văn bản thành vector số).
* **Bước 2:** Chia dữ liệu thành tập huấn luyện (train) và tập kiểm tra (test).
* **Bước 3:** Xây dựng mô hình SVM để huấn luyện trên tập train.
* **Bước 4:** Đánh giá mô hình trên tập test.
* **Bước 5:** Dùng mô hình để dự đoán nhãn của văn bản mới.

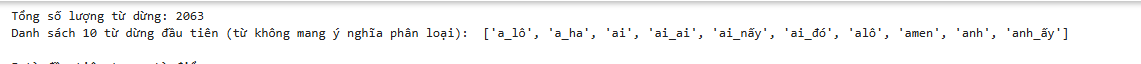
**2.2. Tiền xử lý dữ liệu**

Văn bản cần được chuyển thành dạng ma trận TF-IDF để phục vụ mô hình SVM.

A computer code with text

AI-generated content may be incorrect.

Kết quả :



Chuyển văn bản thành vector TF-IDF

A screen shot of a computer code

AI-generated content may be incorrect.

**2.3. Chia dữ liệu thành tập train và test**

A computer screen shot of a code

AI-generated content may be incorrect.

Kết quả:

**Dữ liệu training = (1071, 12796) (1071,)**

**Dữ liệu testing = (268, 12796) (268,)**

**2.4. Huấn luyện mô hình SVM**

**A screenshot of a computer

AI-generated content may be incorrect.**

**2.5. Đánh giá mô hình**

**A screenshot of a computer

AI-generated content may be incorrect.**

Kết quả :

Độ chính xác: 0.92

**2.6. Dự đoán văn bản mới**

**A screenshot of a computer

AI-generated content may be incorrect.**

Kết quả:

[8] the-thao

**BÀI TẬP BỔ SUNG: PHÂN LOẠI SỐ VIẾT TAY**

**1. Ý tưởng bài toán**

Phân loại số viết tay là một bài toán trong lĩnh vực Machine Learning, thuộc nhóm Nhận dạng mẫu (Pattern Recognition). Mục tiêu của bài toán là xây dựng một mô hình có khả năng nhận diện và phân loại các chữ số viết tay từ hình ảnh.

Bài toán này có thể được giải quyết bằng các mô hình học máy khác nhau, trong đó Support Vector Machine (SVM) là một lựa chọn phổ biến do khả năng phân loại mạnh mẽ.

Dữ liệu được sử dụng trong bài toán này là bộ dữ liệu Digits của Scikit-learn, chứa các hình ảnh số viết tay có kích thước 8x8 pixels. Mỗi hình ảnh sẽ được chuyển thành một vector đặc trưng gồm 64 giá trị (8x8).

**Các bước thực hiện:**

1. **Tải dữ liệu Digits** từ thư viện sklearn.datasets.
2. **Tiền xử lý dữ liệu**: Chuyển đổi hình ảnh thành vector đặc trưng.
3. **Chia tập dữ liệu** thành tập huấn luyện (train) và tập kiểm tra (test).
4. **Huấn luyện mô hình** bằng SVM với kernel='linear'.
5. **Đánh giá mô hình** bằng độ chính xác trên tập kiểm tra.
6. **Dự đoán một số mẫu số viết tay mới**.

**2. Kết quả đạt được**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**2.1. Độ chính xác mô hình**

**Sau khi huấn luyện, mô hình đạt độ chính xác 99.17% trên tập kiểm tra:**

**A computer screen shot of a program

AI-generated content may be incorrect.**

**Nhận xét:**

* Độ chính xác rất cao (99.17%), chứng tỏ mô hình nhận diện số viết tay tốt.
* Số lượng lỗi phân loại rất nhỏ, cho thấy SVM với kernel tuyến tính hoạt động hiệu quả.