TÀI LIỆU ĐẶC TẢ TOOL PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

BẰNG PHƯƠNG PHÁP NAÏVE BAYES

Mục lục

[1.Giới thiệu thiết kế Tool phân loại văn bản 1](#_Toc343985676)

[2.Thuật toán Naïve Bayes. 3](#_Toc343985677)

[3.Áp dụng cho bài toán phân loại văn bản. 4](#_Toc343985678)

**Phân công nhiệm vụ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ tên** | **Nhiệm vụ** |
| 1 | Tạ Tuấn Anh | Viết method tính xác suất của từ trong Category, FrmTraining, FrmMain.  Viết tài liệu đặc tả. |
| 2 | Lưu Hùng Cường | Viết method phân loại văn bản mới, FrmClassify, FrmImport.  Thiết kế giao diện.  Viết tài liệu đặc tả. |
| 3 | Nguyễn Thị Ngọc Lan | Viết method tính xác suất của Category, Thiết kế Database.  Viết tài liệu đặc tả.  Tổng hợp báo cáo. |

# 1.Giới thiệu thiết kế Tool phân loại văn bản

Thiết kế tool gồm 3 tầng: Tầng Giao diện (class PhanLoaiVanBan.Application), tầng xử lý văn bản (class PhanLoaiVanBan.document), tầng truy xuất vào CSDL (class PhanLoaiVanBan.dataaccess).

1. Tầng Giao diện (class PhanLoaiVanBan.Application)

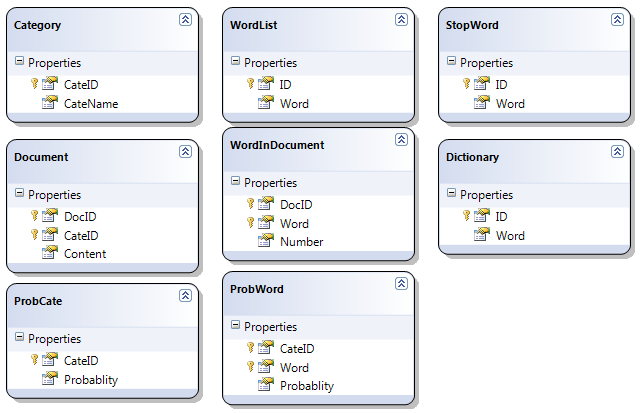
* FrmImport: nhập mới các document để training.
* FrmTraining: thực hiện huấn luyện dựa trên tập dữ liệu.
* FrmClassify: phân loại văn bản.
* FrmMain: giao diện chính của chương trình.

2. Tầng xử lý văn bản (class PhanLoaiVanBan.document)

* Class Document analyzer: thực hiện phân tích văn bản và đưa vào CSDL (thống kê các từ trong văn bản và đưa vào các bảng trong CSDL tương ứng).
* Class NaiveBayes: gồm các phương thức Training để tính toán xác suất của các phân lớp, xác suất của các từ trong Document tương ứng với các Category.

3. Tầng truy xuất vào CSDL (class PhanLoaiVanBan.dataaccess): thực hiện thao tác đọc và ghi vào CSDL.

4. Thiết kế cơ sở dữ liệu: gồm các bảng



|  |  |
| --- | --- |
| **Tên bảng** | **Mô tả** |
| Category | Chứa danh mục các lớp của văn bản |
| WordList | Chứa các tử của văn bản đã được huấn luyện |
| StopWord | Chứa các từ dừng, cần phải loại bỏ khi xử lý văn bản |
| WordInDocument | Chứa các từ trong văn bản |
| Document | Chứa dữ liệu văn bản dùng để huấn luyện |
| ProbCate | Lưu thông tin xác suất của từng phân lớp |
| Dictionary | Chứa từ điển các từ, dùng để tách từ |
| ProbWord | Lưu thông tin xác suất của từ của văn bản tương ứng với từng phân lớp |

# 2.Thuật toán Naïve Bayes.

Thuật toán Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

Áp dụng trong bài toán phân loại, các dữ kiện gồm có:

* D: tập dữ liệu huấn luyện đã được vector hóa dưới dạng
* Ci: phân lớp i, với i = {1,2,…,m}.
* Các thuộc tính độc lập điều kiện đôi một với nhau.

Theo định lý Bayes:

Theo tính chất độc lập điều kiện:

Trong đó:

* là xác suất thuộc phân lớp i khi biết trước mẫu X.
* xác suất là phân lớp i.
* xác suất thuộc tính thứ k mang giá trị xk khi đã biết X thuộc phân lớp i.

**Các bước thực hiện thuật toán Naïve Bayes:**

Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính và

Bước 2: Phân lớp , ta cần tính xác suất thuộc từng phân lớp khi đã biết trước Xnew. Xnew được gán vào lớp có xác suất lớn nhất theo công thức

# 3.Áp dụng cho bài toán phân loại văn bản.

**Ý tưởng:** Sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng…

**Hướng dẫn cài đặt:**

Mô tả vector đặc trưng của văn bản: Là vector có số chiều là số đặc trưng trong toàn tập dữ liệu, các đặc trưng này đôi một khác nhau. Nếu văn bản có chứa đặc trưng đó sẽ có giá trị 1, ngược lại là 0.

Thuật toán gồm 2 giai đoạn huấn luyện và phân lớp:

1. Huấn luyện: tính và

Đầu vào:

* Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
* Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.

Đầu ra:

* Các giá trị xác suất và .

Công thức tính đã làm trơn Laplace

Trong đó:

* |docsi|: số văn bản của tập huấn luyện thuộc phân lớp i.
* |total docs|: số văn bản trong tập huấn luyện.
* m số phân lớp

Cài đặt:

* Khởi tạo mảng A, B có kích thước m.
* Duyệt qua các văn bản trong tập dữ liệu, đếm số văn bản trong mỗi phân lớp lưu vào A.
* Tính xác suất cho từng phân lớp theo công thức trên và lưu vào mảng B.

Công thức tính đã làm trơn Laplace:

Trong đó:

* : Số văn bản trong trong phân lớp i có đặc trưng thứ k mang giá trị xk. (hay số văn bản trong lớp i, có xuất hiện/không xuất hiện đặc trưng k)
* : Số văn bản của tập huấn luyện thuộc phân lớp i.
* Số giá trị có thể có của đặc trưng thứ k

Cài đặt:

* Với vector đặc trưng như mô tả bên trên, dk ở đây mang giá trị là 2, tương ứng với xuất hiện và không xuất hiện. Do chỉ có 2 giá trị, ta có thể tính nhanh xác suất không xuất hiện theo công thức
* Khởi tạo mảng 3 chiều C, chiều 1 có kích thước là m (số phân lớp), chiều 2 có kích thước là N (số đặc trưng), chiều 3 có kích là 2 (dk) để lưu các giá trị .
* Duyệt qua các văn bản trong tập dữ liệu, tiến hành thống kê các chỉ số cần thiết để tính xác suất theo công thức trên và lưu vào mảng C.

1. Phân lớp:

Đầu vào:

* Vector đặc trưng của văn bản cần phân lớp.
* Các giá trị xác suất và .

Đầu ra:

* Nhãn/lớp của văn bản cần phân loại.

Công thức tính xác suất thuộc phân lớp i khi biết trước mẫu X

Dựa vào vector đặc trưng của văn bản cần phân lớp, áp dụng công thức trên tính xác suất thuộc từng phân lớp cho văn bản, và chọn ra lớp có xác suất cao nhất.