BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**ỨNG DỤNG OPTICAL CHARACTER RECOGNITION  
VÀO HỆ THỐNG QUẢN LÝ CÔNG VĂN**

**Sinh viên: Trần Văn Hòa**

**Mã số: B1809127**

**Khóa: K44**

**Cần Thơ, 12/2022**

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**KHOA TRUYỀN THÔNG ĐA PHƯƠNG TIỆN**



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**CHUYÊN NGÀNH TIN HỌC ỨNG DỤNG**

**Đề tài**

**ỨNG DỤNG OPTICAL CHARACTER RECOGNITION  
VÀO HỆ THỐNG QUẢN LÝ CÔNG VĂN**

**Người hướng dẫn**

**Ths Võ Hải Đăng**

**Sinh viên thực hiện**

**Trần Văn Hòa**

**Mã số: B1809127**

**Khóa: K44**

**Cần Thơ, 12/2022**

**LỜI CẢM ƠN**

# MỤC LỤC

[MỤC LỤC i](#_Toc119610484)

[DANH MỤC HÌNH ẢNH iv](#_Toc119610485)

[DANH MỤC BẢNG v](#_Toc119610486)

[DANH MỤC KÝ HIỆU CHỮ VIẾT TẮT vi](#_Toc119610487)

[TÓM TẮT vii](#_Toc119610488)

[ABSTRACT viii](#_Toc119610489)

[PHẦN GIỚI THIỆU 1](#_Toc119610490)

[1. ĐẶT VẤN ĐỀ 1](#_Toc119610491)

[2. LỊCH SỬ GIẢI QUYẾT VẤN ĐỀ 1](#_Toc119610492)

[3. MỤC TIÊU ĐỀ TÀI 2](#_Toc119610493)

[4. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU 2](#_Toc119610494)

[5. NỘI DUNG NGHIÊN CỨU 2](#_Toc119610495)

[6. NHỮNG ĐÓNG GÓP CHÍNH CỦA ĐỀ TÀI 3](#_Toc119610496)

[7. BỐ CỤC CỦA QUYỂN LUẬN VĂN 3](#_Toc119610497)

[PHẦN NỘI DUNG 4](#_Toc119610498)

[CHƯƠNG 1: MÔ TẢ BÀI TOÁN 4](#_Toc119610499)

[1.1. MÔ TẢ CHI TIẾT BÀI TOÁN 4](#_Toc119610500)

[1.2. CÁC CHỨC NĂNG TRÔNG ĐỢI 5](#_Toc119610501)

[1.3. PHÂN TÍCH VÀ ĐÁNH GIÁ CÁC GIẢI PHÁP CÓ LIÊN QUAN 5](#_Toc119610502)

[1.4. TIẾP CẬN GIẢI QUYẾT VẤN ĐỀ, CHỌN GIẢI PHÁP 5](#_Toc119610503)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 5](#_Toc119610504)

[2.1. Nhận dạng ký tự quan học (Optical Character Recognition) 6](#_Toc119610505)

[2.1.1. Giới thiệu 6](#_Toc119610506)

[2.1.2. Lịch sử của OCR 6](#_Toc119610507)

[2.1.3. Phương pháp nhận dạng ký tự 7](#_Toc119610508)

[2.1.3.1. Các thành phần của hệ thống OCR 7](#_Toc119610509)

[2.1.3.2. Optical scanning (quét quang học) 8](#_Toc119610510)

[2.1.3.3. Location and segmentation (định vị và phân đoạn) 8](#_Toc119610511)

[2.1.3.4. Preprocessing (tiền xử lý) 9](#_Toc119610512)

[2.1.3.5. Feature extraction (khai thác) 9](#_Toc119610513)

[2.1.3.6. Recognition and Post-processing (nhận dạng) 9](#_Toc119610514)

[2.2. Tesseract [4], [5] 9](#_Toc119610515)

[2.3. Giải thuật cây quyết định (Decision tree) 10](#_Toc119610516)

[2.3.1. Giải thuật xây dựng cây quyết định [6] 10](#_Toc119610517)

[2.3.2. Nghi thức kiểm tra giải thuật [11] 11](#_Toc119610518)

[2.3.3. Đo độ hiệu quả của giải thuật [11] 11](#_Toc119610519)

[2.4. Văn bản hành chính [12] 12](#_Toc119610520)

[2.5. Quy trình quản lý công văn 14](#_Toc119610521)

[2.6. Các công nghệ sử dụng 14](#_Toc119610522)

[CHƯƠNG 3: THIẾT KẾ VÀ CÀI ĐẶT THUẬT TOÁN 15](#_Toc119610523)

[3.1. GIẢI PHÁP PHÂN TÍCH VÀ THIẾT KẾ MÔ HÌNH 15](#_Toc119610524)

[3.1.1. Sơ đồ Use Case 15](#_Toc119610525)

[3.1.2. Sơ đồ ER 15](#_Toc119610526)

[3.1.3. Sơ đồ LDM 16](#_Toc119610527)

[3.1.4. Sơ đồ LDM của MongoDB 16](#_Toc119610528)

[3.1.5. Sơ đồ phân rã chức năng 17](#_Toc119610529)

[3.1.6. Sơ đồ hoạt động 18](#_Toc119610530)

[3.1.6.1. Sơ đồ hoạt động chức năng đăng nhập 18](#_Toc119610531)

[3.1.6.2. Sơ đồ hoạt động chức năng thêm công văn đến 18](#_Toc119610532)

[3.1.6.3. Sơ đồ hoạt động chức năng sửa công văn đến 18](#_Toc119610533)

[3.1.6.4. Sơ đồ hoạt động chức năng phê duyệt công văn đến 18](#_Toc119610534)

[3.1.6.5. Sơ đồ hoạt động chức năng phân công xử lý công văn đến 18](#_Toc119610535)

[3.1.6.6. Sơ đồ hoạt động chức năng xử lý công văn đến 18](#_Toc119610536)

[3.1.6.7. Sơ đồ hoạt động chức năng từ chối công văn đến 18](#_Toc119610537)

[3.1.6.8. Sơ đồ hoạt động chức năng xóa công văn đến 18](#_Toc119610538)

[3.1.6.9. Sơ đồ hoạt động thống kê công văn đến 18](#_Toc119610539)

[3.1.6.10. Sơ đồ hoạt động thêm cán bộ 18](#_Toc119610540)

[3.1.6.11. Sơ đồ hoạt động sửa cán bộ 18](#_Toc119610541)

[3.1.6.12. Sơ đồ hoạt động xóa cán bộ 18](#_Toc119610542)

[3.1.7. Sơ đồ tuần tự 19](#_Toc119610543)

[3.1.7.1. Sơ đồ tuần tự chức năng đăng nhập 19](#_Toc119610544)

[3.1.7.2. Sơ đồ tuần tự chức năng thêm công văn đến 19](#_Toc119610545)

[3.1.7.3. Sơ đồ tuần tự chức năng sửa công văn đến 19](#_Toc119610546)

[3.1.7.4. Sơ đồ tuần tự chức năng phê duyệt công văn đến 19](#_Toc119610547)

[3.1.7.5. Sơ đồ tuần tự chức năng phân công xử lý công văn đến 19](#_Toc119610548)

[3.1.7.6. Sơ đồ tuần tự chức năng xử lý công văn đến 19](#_Toc119610549)

[3.1.7.7. Sơ đồ tuần tự chức năng từ chối công văn đến 19](#_Toc119610550)

[3.1.7.8. Sơ đồ tuần tự chức năng xóa công văn đến 19](#_Toc119610551)

[3.1.7.9. Sơ đồ tuần tự thống kê công văn đến 19](#_Toc119610552)

[3.1.7.10. Sơ đồ tuần tự thêm cán bộ 19](#_Toc119610553)

[3.1.7.11. Sơ đồ tuần tự sửa cán bộ 19](#_Toc119610554)

[3.1.7.12. Sơ đồ tuần tự xóa cán bộ 19](#_Toc119610555)

[CHƯƠNG 4: KIỂM THỬ VÀ ĐÁNH GIÁ 19](#_Toc119610556)

[PHẦN KẾT LUẬN 20](#_Toc119610557)

[1. KẾT QUẢ ĐẠT ĐƯỢC 20](#_Toc119610558)

[2. HƯỚNG PHÁT TRIỂN 20](#_Toc119610559)

[TÀI LIỆU THAM KHẢO 21](#_Toc119610560)

[PHỤ LỤC I BIỂU MẪU THƯỜNG GẶP 22](#_Toc119610561)

[PHỤ LỤC II CÁC BẢNG DỮ LIỆU 1](#_Toc119610562)

# DANH MỤC HÌNH ẢNH

[Hình 2.1 Các loại của nhận dạng ký tự 8](#_Toc119658591)

[Hình 2.2 OCR-A (Trên), OCR-B (Dưới) 9](#_Toc119658592)

[Hình 2.3 Các thành phần của hệ thống OCR 10](#_Toc119658593)

[Hình 2.4 Các vấn đề trong việc chọn ngưỡng: Trên: hình ảnh mức xám gốc, Giữa: hình ảnh phân ngưỡng không tốt, Dưới: hình ảnh phân ngưỡng tốt. 10](#_Toc119658594)

[Hình 2.5 Các ký tự bị lỗi 11](#_Toc119658595)

[Hình 2.6 Chuẩn hóa dữ liệu 11](#_Toc119658596)

[Hình 2.7 Các giải thuật được sử dụng nhiều nhất [10] 12](#_Toc119658597)

[Hình 2.8 Sơ đồ bố trí các thành phần thể thức văn bản hành chính 15](#_Toc119658598)

[Hình 3.1 Biểu đồ đường thống kê tỉ lệ phân lớp đúng của 10 lần thực nghiệm 19](#_Toc119658599)

[Hình 3.2 Cây quyết định 20](#_Toc119658600)

[Hình 3.3 Các bước xử lý văn bản 21](#_Toc119658601)

[Hình 3.4 Sơ đồ ER 22](#_Toc119658602)

[Hình 3.5 Sơ đồ LDM 22](#_Toc119658603)

[Hình 3.6 Sơ đồ LDM của MongoDB 23](#_Toc119658604)

[Hình 3.7 Sơ đồ phân rã chức năng 24](#_Toc119658605)

# DANH MỤC BẢNG

[Bảng 2.1 Ma trận confusion 2x2 hay bảng contingancy 13](#_Toc119656719)

[Bảng 3.1 Bảng mô tả thuộc tính của tập dữ liệu 18](#_Toc119656720)

[Bảng 3.2 Thống kê các lần chạy thực nghiệm bằng nghi thức k-fold 19](#_Toc119656721)

[Bảng 3.3 Bảng thông số kết quả chạy giải thuật với k = 10 20](#_Toc119656722)

# DANH MỤC KÝ HIỆU CHỮ VIẾT TẮT

| **STT** | **Từ khóa** | **Diễn giải** |
| --- | --- | --- |
| 1 | OCR | Optical Character Recognition (Nhận dạng ký tự quang học) |
|  |  |  |

# TÓM TẮT

***Từ khóa: NodeJS, Express, ReactJS, nhận dạng ký tự quang học, cây quyết định, quản lý văn bản hành chính.***

# ABSTRACT

**Keyword: NodeJS, Express, ReactJS, optical character recognition.**

# PHẦN GIỚI THIỆU

## ĐẶT VẤN ĐỀ

Bài toán, vấn đề đặt ra đối với đề tài là gì?

Tại sao cần làm đề tài này?

Trong bất kì lĩnh vực nào thì công tác quản lý luôn giữ một vai trò vô cùng  
quan trọng, trong đó có công tác quản lý công văn. Việc quản lý công văn trong các đơn vị hành chính và doanh nghiệp luôn là vấn đề cấp thiết, đòi hỏi có sự  
đầu tư cả về nhân lực và trang thiết bị.

Bên cạnh đó Chính phủ cũng ban hành các văn bản nhằm khuyến kích việc chuyển đổi số trong công tác quản lý công văn như: Nghị định số 30/2020/NĐ-CP ngày 05/3/2020 của Chính phủ về công tác văn thư có quy định về hệ thống quản lý tài liệu điện tử; Nghị quyết số 26/NQ-CP ngày 15/4/2015 của Chính phủ ban hành chương trình hành động của Chính phù thực hiện nghị quyết số 36-NQ/TW ngày 01/7/2014 của Bộ Chính trị Ban Chấp hành Trung ương Đảng về đầy mạnh ứng dụng, phát triển công nghiệp thông tin đáp ứng yêu cầu phát triển bền vững và hội nhập quốc tế;

Trước đây, việc quản lý công văn, giấy tờ thường theo quy cách truyền  
thống là lưu trữ trên giấy tờ, trong các cặp, tủ hồ sơ nên gây ra không ít phiền phức  
trong việc tra cứu, lưu trữ và bảo quản. Gần đây việc áp dụng các hệ thống quản lý vào việc quản lý công văn cũng đã dược sử dụng nhưng đa phần đều phải nhập liệu truyền thống. Do đó cần có một biện pháp quản lý tối ưu hơn trong việc quản lý công văn trong các đơn vị hành chính và doanh nghiệp.

OCR và Data Mining phát triển

Chính vì lý do đó, em quyết định nghiên cứu xây dựng một hệ thống quản lý bằng công nghệ OCR kết hợp với các thuật toán trong Data Mining để giúp việc quản lý công văn diễn ra nhanh chóng, dễ dàng và ít sai sót hơn.

Vì vậy trong luận văn sẽ tập trung tìm hiểu các kỹ thuật, các công nghệ cần thiết để xây dựng hệ thống **Ứng dụng Optical Character Recognition vào hệ thống quản lý công văn** làm đề tài luận văn tốt nghiệp của mình.

## LỊCH SỬ GIẢI QUYẾT VẤN ĐỀ

Trong khoa.

Trong nước.

Ngoài nước.

Đề tài nào đã giải quyết, những vấn đề nào chưa giải quyết được?

## MỤC TIÊU ĐỀ TÀI

Những vấn đề trọng tâm.

Hệ thống được xây dựng với mục tiêu cung cấp đầy đủ các tính năng cho quy trình nhận/gửi công văn của một hệ thống quản lý công văn thông thường tích hợp thêm việc trích xuất và nhận dạng các thành phần của công văn:

* Văn thư dễ dàng tìm kiếm công văn theo các tiêu chí như, ngày ban hành, số ban hành, người ký, … Dễ dàng quản lý trạng thái xử lý công văn cũng như phân loại công văn.
* Hỗ trợ việc nhận dạng và trích xuất các thành phần trên công văn giúp cho việc nhập liệu nhanh chóng và ít sai sót hơn.
* Hỗ trợ việc báo cáo hoạt động quản lý một cách trực quan, rõ ràng, nhanh chóng.
* Hỗ trợ quản trị viên quản lý hệ thống tốt hơn, dễ dàng thiết lập các dữ liệu và việc phân quyền được thực hiện linh hoạt.
* Giao diện hài hòa, thân thiện và hiện đại giúp người dùng dễ dàng sử dụng.

## ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

**Đối tượng nghiên cứu** là một hệ thống quản lý công văn tích hợp các thuật toán Data Mining và công nghệ OCR.

**Phạm vi nghiên cứu** của đề tài bao gồm các tính năng thêm mới công văn, xử lý các nghiệp vụ quản lý như: phê duyệt, phân công xử lý, xử lý, ban hành, …. Đặt biệt là chức năng trích xuất và nhận dạng các thành phần của công văn.

## NỘI DUNG NGHIÊN CỨU

Nghiên cứu về quy trình quản lý công văn, nghiên cứu về công nghệ OCR, xử lý ảnh, các thuật toán về Data Mining.

**Về lý thuyết:**

* Tìm hiểu về Optical Character Recognition.
* Tìm hiểu về các phương pháp xử lý ảnh.
* Tìm hiểu về các thuật toán Data Mining.
* Tìm hiểu về quy trình quản lý công văn theo quy định của nhà nước.

**Về người dùng và chức năng:**

**Phương pháp thực hiện:**

## NHỮNG ĐÓNG GÓP CHÍNH CỦA ĐỀ TÀI

Kết quả chính đạt được

## BỐ CỤC CỦA QUYỂN LUẬN VĂN

Bố cục của luận văn bao gồm các thành phần như sau:

**Phần giới thiệu:** Giới thiệu tổng quan về luận văn, đặt vấn đề, lịch sử giải quyết vấn đề, mục tiêu đề tài, đối tượng và phạm vi nghiên cứu, nội dung nghiên cứu.

**Phần nội dung:** Nội dung gồm 3 chương chính:

**Chương 1:** Mô tả bài toán

**Chương 2:** Thiết kế và cài đặt thuật toán

**Chương 3:** Kiểm thử và đánh giá

**Phần kết luận:** Tổng hợp kết quả đạt được và đề xuất hướng phát triển trong tương lai

# PHẦN NỘI DUNG

## MÔ TẢ BÀI TOÁN

### MÔ TẢ CHI TIẾT BÀI TOÁN

- Mô tả chỉ tiết bài toán: Giúp người đọc hiểu rõ các chức năng/tính năng/đặc điêm của sản phâm/phân mềm/hệ thông/giải pháp là mục đích cân đạt được của đê tài. Nêu là phân mêm thì mô tả các chức năng trông đợi. Nêu là giải pháp thì mô tả những vân đê mà giải pháp sẽ xử lý, cải tiên, khăc phục ...

Người dùng gồm có quản trị viên lãnh đạo đơn vị, văn thư và cán bộ. Mỗi người dùng gồm có các thông tin sau: mã số, chức vụ, tên địa chỉ email, số điện thoại. Mỗi người dùng sẽ thuộc một đơn vị trong cơ quan và chỉ thực hiện các chức năng dành cho cơ quan mình thuộc về. Thông tin của mỗi người dùng sẽ được quản lý tạo sẵn với tài khoản và mật khẩu mặc định sau đó mỗi người dùng sẽ tự thay đổi lại. Mỗi người dùng khi muốn sử dụng phải đăng nhập vào hệ thống mới có thể sử dụng.

Tương ứng với từng người dùng sẽ có các quyền sử dụng riêng cho phép người dùng thực hiện các chức năng khác nhau. Quyền sử dụng sẽ được quản trị hệ thống tạo riêng cho từng nhóm đối tượng. Quyền sử dụng bao gồm các thuộc tính: tên quyền, quyền đọc công văn, quyền thêm mới công văn, quyền cập nhật công văn, quyền đọc danh sách cán bộ, quyền thêm mới thông tin cán bộ, quyền cập nhật thông tin cán bộ, quyền xóa thông tin cán bộ, quyền đọc danh sách quyền, quyền tạo mới quyền, quyền cập nhật quyền, quyền xóa quyền, phạm quy quyền (0: phạm vi hệ thống, 1: phạm vi cơ quan).

Các chức năng của quản trị viên gồm: Cập nhật thông tin của cán bộ, khóa hoặc mở khóa tài khoản của cán bộ, cập nhật thông tin của đơn vị, cập nhật thông tin của loại công văn, tạo quyền quản lý cho cán bộ.

Cán bộ thường có các chức năng sau: Xem và tìm kiếm công văn đi/đến trong đơn vị của mình, xử lí các công văn được phân công.

Cán bộ văn thư đơn vị cũng chính là cán bộ thường vì vậy có các chức năng của cán bộ thường kèm theo các chức năng: Duyệt công văn đến, chuyển công văn đã phê duyệt cho cán bộ xử lý, cập nhật thông tin công văn đến (thêm mới, sửa, xóa), cập nhật thông tin công văn đi (thêm mới, sửa, xóa) và nhắc nhở xử lí công văn.

Tương tự như văn thư, cán bộ lãnh đạo cũng có các chức năng của cán bộ thường và phê duyệt công văn đến, chuyển công văn đã phê duyệt cho cán bộ xử lý, nhắc nhở xử lý công văn, phân quyền văn thư, xem tình hình xử lý công văn.

Đối với chức năng xem danh sách công văn: Mỗi cán bộ chỉ có thể xem được công văn của đơn vị thuộc về, những công văn có độ mật là không, những công văn do cán bộ đó nhập, duyệt hoặc xử lý.

Đối với chức năng tìm kiếm công văn sẽ có 2 loại là tìm kiếm cơ bản theo ký tự có trong một trường nào đó của công văn và tìm kiếm nâng cao theo từng trường của công văn (tìm theo ngày, theo tiêu đề, …).

Mỗi đơn vị bao gồm các thông tin sau: tên đơn vị, mã đơn vị, địa chỉ email đơn vị, đơn vị đó có phải ở bên trong cơ quan không, cơ quan thuộc về.

Công văn gồm các thông tin cơ bản sau: Số công văn, ngày ban hành, trích dẫn, số trang, ý kiến bút phê, tên người ký, chức vụ người ký, hạn giải quyết, danh sách tập tin đính kèm. Đối với công văn đi có thêm số lượng ban hành. Đối với công văn đến kèm theo ngày đến và số đến.

### CÁC CHỨC NĂNG TRÔNG ĐỢI

### PHÂN TÍCH VÀ ĐÁNH GIÁ CÁC GIẢI PHÁP CÓ LIÊN QUAN

- Phân tích đánh giá các giải pháp có liên quan đến bài toán

### TIẾP CẬN GIẢI QUYẾT VẤN ĐỀ, CHỌN GIẢI PHÁP

- Tiếp cận giải quyết vấn đề, chọn lựa giải pháp

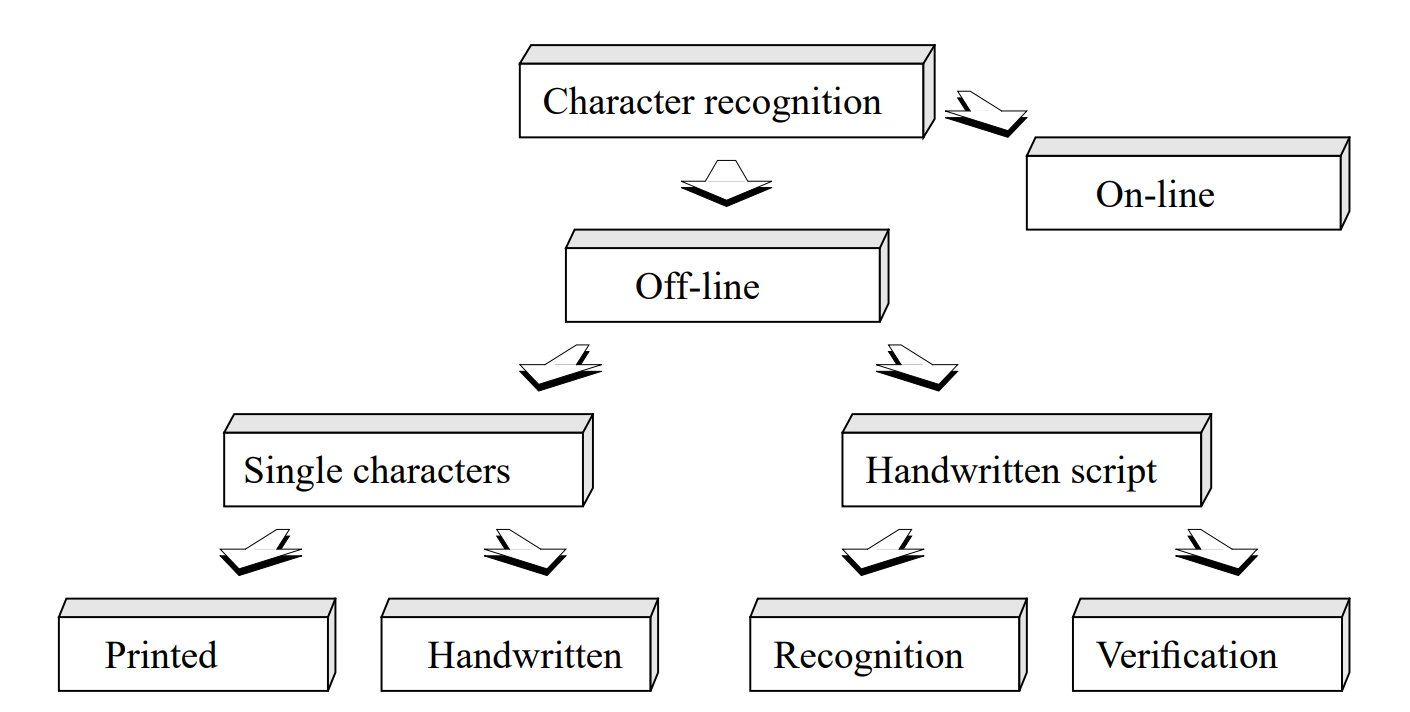
## CƠ SỞ LÝ THUYẾT

### Nhận dạng ký tự quan học (Optical Character Recognition)

#### Giới thiệu

Optical Character Recognition (OCR) là một hệ thống chuyển đổi hình ảnh của các văn bản đầu vào (chữ viết tay hoặc chữ đánh máy) thành mã máy. OCR được hình thành từ một lĩnh vực nghiên cứu về nhận dạng mẫu,  
trí tuệ nhân tạo và machine vision. [1], [2]

Optical Character Recognition giải quyết vấn đề nhận dạng các ký tự được xử lý quang học. Nhận dạng quang học được thực hiện ngoại tuyến sau khi viết hoặc in xong, trái ngược với nhận dạng trực tuyến khi máy tính nhận dạng các ký tự khi chúng được vẽ. Có thể nhận dạng được cả ký tự viết tay và ký tự in nhưng hiệu suất phụ thuộc vào chất lượng của dữ liệu đầu vào. [3]



Hình 2.1 Các loại của nhận dạng ký tự

#### Lịch sử của OCR

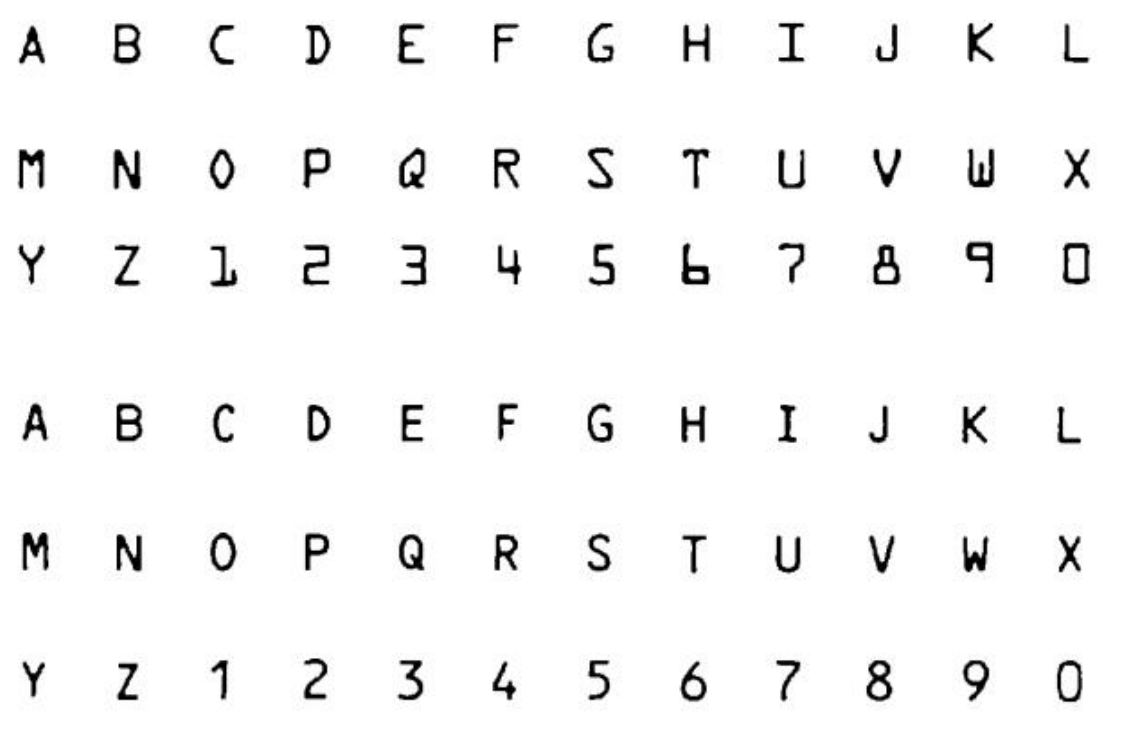
Năm 1870, C.R. Carey ở Boston Massachusetts đã phát minh ra máy quét võng mạc, hệ thống truyền hình ảnh sử dụng các tế bào quang điện.

Năm 1940, phiên bản hiện đại hơn của OCR ra đời.

Năm 1950, Máy OCR đầu tiên ra đời.

Từ 1960 đến 1965, các hệ thống OCR thương mại đầu tiên xuất hiện.

Từ 1965 đến 1975, OCR thế hệ thứ hai với hiệu xuất cao và chi phí thấp hơn ra đời. Font chữ OCR-A và OCR-B được tạo ra để hỗ trợ việc nhận dạng dễ dàng hơn.



Hình 2.2 OCR-A (Trên), OCR-B (Dưới)

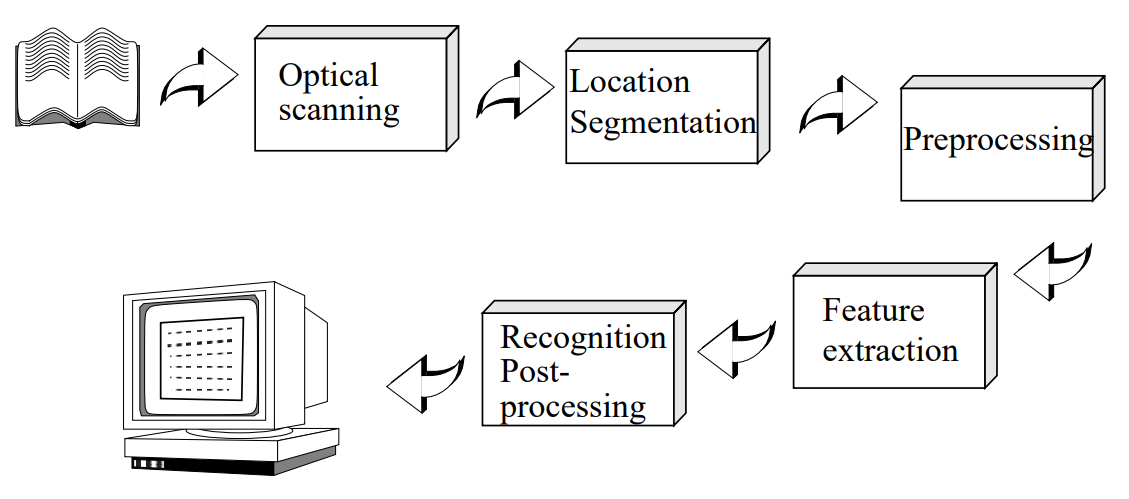
Từ năm 1975 đến 1985, thế hệ thứ ba với các thiết bị OCR đơn giản hơn nhờ sự xuất hiện của máy tính cá nhân và máy in laser.

Ngày nay khi phần cứng ngày càng rẻ hơn và các hệ thống OCR đã có sẵn dưới dạng các gói phần mềm.

#### Phương pháp nhận dạng ký tự

##### Các thành phần của hệ thống OCR

Một hệ thống OCR thông thường sẽ bao gồm một số thành phần như hình 1.3. Bước đầu tiên là số hóa tài liệu bằng máy quét quang học. Khi các vùng chứa văn bản được định vị, mỗi biểu tượng được trích xuất thông qua quá trình phân đoạn. Các biểu tượng được trích xuất sau đó có thể được tiền xử lý, loại bỏ nhiễu, để tạo thuận lợi cho việc trích xuất các trong bước tiếp theo.



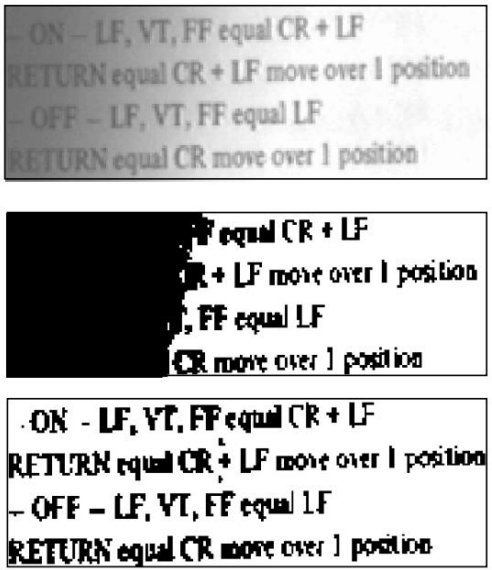
Hình 2.3 Các thành phần của hệ thống OCR

Nhận dạng của từng biểu tượng được tìm thấy bằng cách so sánh các đặc điểm được trích xuất với các mô tả về các lớp biểu tượng thu được qua giai đoạn học tập trước đó. Cuối cùng, thông tin theo ngữ cảnh được sử dụng để tái tạo lại các từ và số của văn bản gốc.

##### Optical scanning (quét quang học)

Hình ảnh được chụp bằng máy quét, sau đó sẽ chuyển đổi cường độ sáng thành các mức xám. Tài liệu in thường bao gồm chữ đen trên nền trắng. Do đó, thông thường là chuyển đổi hình ảnh đa cấp độ thành hình ảnh hai cấp độ đen trắng. Quá trình này được gọi là phân ngưỡng.

Quá trình tạo ngưỡng rất quan trọng vì kết quả nhận dạng hoàn toàn phụ thuộc vào chất lượng của hình ảnh đã phân ngưỡng.

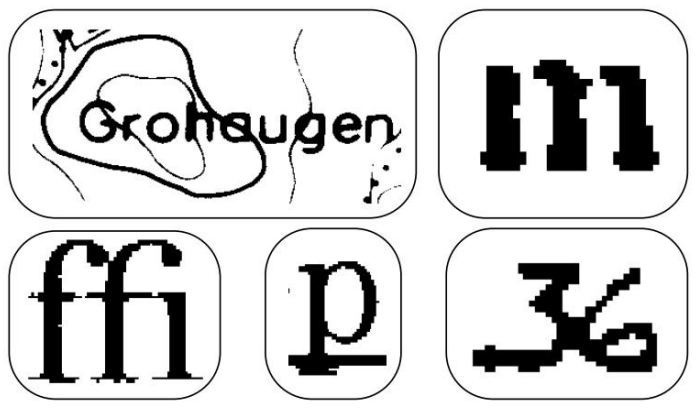


Hình 2.4 Các vấn đề trong việc chọn ngưỡng: Trên: hình ảnh mức xám gốc, Giữa: hình ảnh phân ngưỡng không tốt, Dưới: hình ảnh phân ngưỡng tốt.

##### Location and segmentation (định vị và phân đoạn)

Phân đoạn là quá trình xác định các thành phần của một hình ảnh. Cần xác định vị trí các vùng của tài liệu nơi dữ liệu đã được in ra và phân biệt chúng với các hình và đồ họa.

Áp dụng trong văn bản, phân đoạn là sự cô lập của các ký tự hoặc từ. Kỹ thuật này dễ thực hiện, nhưng sẽ gặp hạn chế nếu các ký tự chạm vào nhau hoặc các ký tự bị phân mảnh và bao gồm nhiều phần.



Hình 2.5 Các ký tự bị lỗi

##### Preprocessing (tiền xử lý)

Hình ảnh thu được từ quá trình quét có thể chứa một lượng nhiễu nhất định tùy thuộc vào máy quét. Các ký tự có thể bị nhòe hoặc bị hỏng. Những lỗi này làm cho việc nhận dạng bị kém đi, có thể được cải thiện bằng cách sử dụng bộ tiền xử lý để làm mịn các ký tự số hóa.



Hình 2.6 Chuẩn hóa dữ liệu

##### Feature extraction (khai thác)

Sử dụng các kỹ thuật như Template-matching and correlation, Feature based để tiến hành trích xuất các đặt trưng. ???

##### Recognition and Post-processing (nhận dạng)

???

### Tesseract [4], [5]

Tesseract là một gói mã nguồn mở được phá triển bởi HP từ những năm 1984 đến 1994.

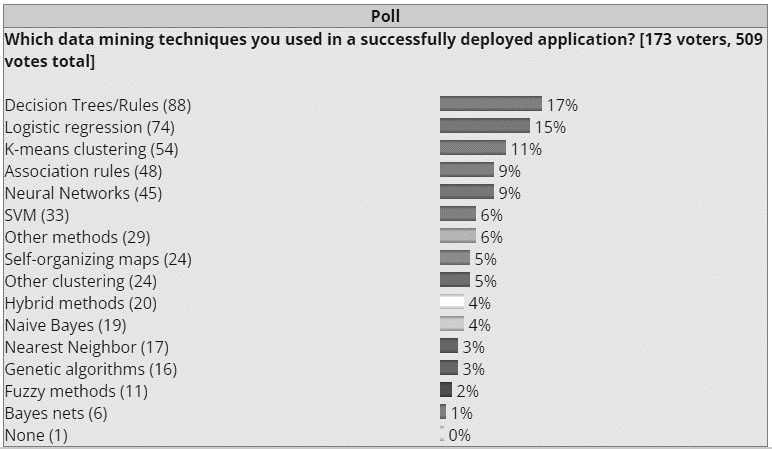
Tesseract ban đầu là một luận án nghiên cứu tiến sĩ trong phòng thí nghiệm HP, Briston. Kết quả của công việc nghiên cứu là xây dựng một phần mềm hoặc phần cứng add on hỗ trợ cho máy quét HP phẳng.

Sau dự án hợp tác giữa HP Labs, Bristol và bộ phận quét HP tại Colorado, độchính xác của Tesseract tăng lên đáng kể theo các công cụ thương mại, nhưng không trở thành sản phẩm thương mại.

Giai đoạn phát triển tiếp theo của Tesseract chủ yếu được thực hiện ở HP Labs, Briston, tập trung nhiều vào việc cải thiện độ chính xác và dự án này kết thúc vào cuối năm 1994. Cuối năm 2005, HP công bồ Tesseract là mã nguồn mở. Hiện nay đã có tại http://code.google.com/p/tesseract-ocr.

### Giải thuật cây quyết định (Decision tree)

Cây quyết định là giải thuật khai khoáng dữ liệu rất phổ biến và hiệu quả [6], [7]. Cây quyết định có cấu trúc là một cây nhị phân, mỗi nút trong tương ứng với việc phân hoạch dựa trên một thuộc tính nào đó, nhánh đại diện cho một quy tắc quyết định. Việc xây dựng cây quyết định phụ thuộc vào việc lựa chọn thuộc tính để phân hoạch. [8], [9]



Hình 2.7 Các giải thuật được sử dụng nhiều nhất [10]

#### Giải thuật xây dựng cây quyết định [6]

Giải thuật học cây quyết định gồm 2 bước lớn: xây dựng cây (Top-down), cắt nhánh (Bottom-up) để tránh học vẹt. Quá trình xây dựng được làm như sau:

* Bắt đầu nút gốc, tất cả các dữ liệu học ở nút gốc.
* Nếu dữ liệu tại 1 nút có cùng lớp thì nút được cho là nút lá và nhãn của nút lá là nhãn của các phần tử trong nút lá (hay luận bình chọn số đông nếu nút lá có chứa các phần tử có lớp khác nhau).
* Nếu dữ liệu ở lớp có chứ các phần tử có lớp rất khác nhau (không thuần nhất) thì nút được cho là nút trong, tiến hành phân hoạch dữ liệu một cách đệ quy bằng việc chọn một thuộc tính để thực hiện phân hoạch tốt nhất có thể.

#### Nghi thức kiểm tra giải thuật [11]

Nghi thức **k-fold**: chia tập dữ liệu thành **k** phần (fold) bằng nhau, lặp lại **k** lần, mỗi lần sử dụng **k-1** folds để học và **1** fold để kiểm tra, sau đó tính trung bình của **k** lần kiểm tra. Khi tập dữ liệu có nhiều phần tử (hơn 300) thường chọn **k = 10**, ngược lại chọn **k = số phần tử**.

Nghi thức **hold-out**: chia tập dữ liệu thành 2 phần (thường là **2/3** tập dữ liệu và **1/3** tập dữ liệu) 1 phần dùng để học phần còn lại dùng để kiểm tra, có thể lặp lại **k** lần rồi tính trung bình.

#### Đo độ hiệu quả của giải thuật [11]

Hiệu quả của giải thuật có thể quan sát dựa vào dự đoán nhãn của giải thuật trên tập dữ liệu. Ma trận confusion cung cấp thông tin dự đoán nhãn của giải thuật, phần tử **[i, j]** của ma trận trình bày số phần tử lớp **I** được dự đoán là lớp **j**.

Bảng 2.1 Ma trận confusion 2x2 hay bảng contingancy

|  |  |  |
| --- | --- | --- |
| **Dự đoán**  **Thực tế** | **Dương** | **Âm** |
| **Dương** | TP | FN |
| **Âm** | FP | TN |

* **TP** tổng phần tử lớp **dương** được dự đoán là lớp **dương**.
* **FN** tổng phần tử lớp **dương** được dự đoán là lớp **âm**.
* **TN** tổng phần tử lớp **âm** được dự đoán là lớp **âm**.
* **FP** tổng phần tử lớp **âm** được dự đoán là lớp **dương**.

Các độ đo thường dùng là:

* Tỉ lệ dương tính thật:
* Tỉ lệ dương tính giả:
* Tỉ lệ số mẫu dự đoán dương đúng trong số những mẫu được dự đoán là dương:
* Tỉ lệ số mẫu dự đoán dương đúng trong số những mẫu dương:
* Tỉ lệ số mẫu dự đoán đúng trong tổng số mẫu:
* Số dung hòa Recall và Precision

### Văn bản hành chính [12]

**“Văn bản”** là thông tin thành văn được truyền đạt bằng ngôn ngữ hoặc ký hiệu, hình thành trong hoạt động của các cơ quan, tổ chức và được trình bày đúng thể thức, kỹ thuật theo quy định.

**“Văn bản hành chính”** là văn bản hình thành trong quá trình chỉ đạo, điều hành, giải quyết công việc của các cơ quan, tổ chức.

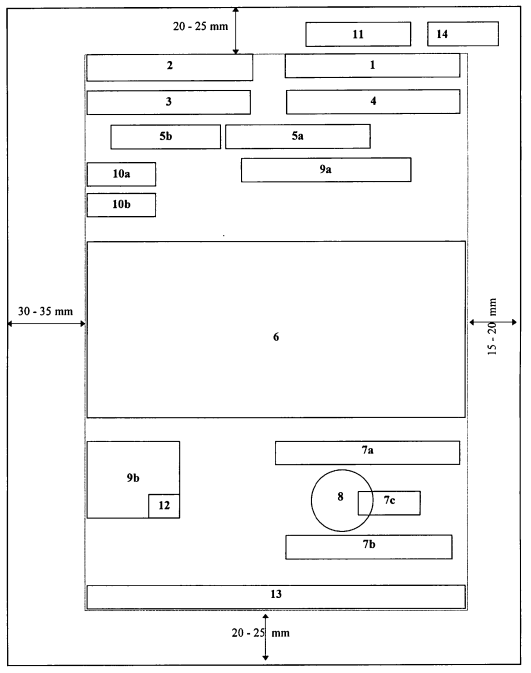
**“Văn bản điện tử”** là văn bản dưới dạng thông điệp dữ liệu được tạo lập hoặc được số hóa từ văn bản giấy và trình bày đúng thể thức, kỹ thuật, định dạng theo quy định.

**“Văn bản đi”** là tất cả các loại văn bản do cơ quan, tổ chức ban hành.

**“Văn bản đến”** là tất cả các loại văn bản do cơ quan, tổ chức nhận được từ cơ quan, tổ chức, cá nhân khác gửi đến.

**Sơ đồ bố trí các thành phần thể thức văn bản hành chính**

Trình bày theo chiều dài của khổ A4 (210 mm x 297 mm). Sơ đồ bố trí các thành phần thể thức văn bản hành chính theo hướng dẫn của **Phụ lục I thể thức kỹ thuật trình bày văn bản hành chính và bản sao văn bản** *(Kèm theo nghị định số 30/2020/NĐ-CP ngày 05 thánh 3 năm 2020 của Chính phủ)*



Hình 2.8 Sơ đồ bố trí các thành phần thể thức văn bản hành chính

**Vị trí trình bày các thành phần thể thức:**

|  |  |  |
| --- | --- | --- |
| **Ô số** | **:** | **Thành phần thể thức văn bản** |
| 1 | : | Quốc hiệu và Tiêu ngữ |
| 2 | : | Tên cơ quan, tổ chức ban hành văn bản |
| 3 | : | Số, ký hiệu của văn bản |
| 4 | : | Địa danh và thời gian ban hành văn bản |
| 5a | : | Tên loại và trích yếu nội dung văn bản |
| 5b | : | Trích yếu nội dung công văn |
| 6 | : | Trích yếu nội dung công văn |
| 7a, 7b, 7c | : | Chức vụ, họ tên và chữ ký của người có thẩm quyền |
| 8 | : | Dấu, Chữ ký số của cơ quan, tổ chức |
| 9a, 9b | : | Nơi nhận |
| 10a | : | Dấu chỉ độ mật |
| 10b | : | Dấu chỉ mức độ khẩn |
| 11 | : | Chỉ dẫn về phạm vi lưu hành |
| 12 | : | Ký hiệu người soạn thảo văn bản và số lượng bản phát hành |
| 13 | : | Địa chỉ cơ quan, tổ chức; thư điện tử; trang thông tin điện tử; số điện thoại; số Fax |
| 14 | : | Chữ ký số của cơ quan, tổ chức cho bản sao văn bản sang định dạng điện tử |

### Quy trình quản lý công văn

### Các công nghệ sử dụng

**HTML (Hyper Text Markup Language)**: …

**CSS (Cascading Style Sheets)**: …

**Javascript**: …

**Bootstrap**: …

**Coreui React**: …

**MongoDb**: …

**React**: …

**Nodejs**: …

**Expressjs**: …

## THIẾT KẾ VÀ CÀI ĐẶT THUẬT TOÁN

Bao gồm thiết kế kiến trúc tổng thể của hệ thống, giải thích chức năng của từng thành phân trong hệ thông, các giải thuật xử lý của hệ thông hoặc của một thành phân hệ thông. thiêt kê cơ sở dữ liệu, ...

Mô tả cách thức cài đặt thiết kế bằng một ngôn ngữ lập trình cụ thể/hệ fflềuẳ Wlý1ýànhlphân cứng .... ị

### PHƯƠNG PHÁP THỰC HIỆN

#### Xử lý và trích xuất các thành phần từ ảnh văn bản

Từ hình ảnh đầu vào sẽ được chuẩn hóa kích thước để dữ liệu ảnh đầu vào được đồng bộ.

Ảnh đầu vào

Chuẩn hóa kích thước ảnh đầu vào

Chuyển đổi thành ảnh xám

Nghịch đảo ảnh

Làm mịn ảnh

Kéo giãn ảnh lần 1

Đóng ảnh

Kéo giãn ảnh lần 2

Mở rộng ảnh

Cắt ảnh thành các phần theo từng biên

Tìm biên

Cắt ngưỡng ảnh

Chuẩn hóa lại kích thước ảnh

Erode

Hình 3.1 Quá trình xử lý ảnh

#### Xây dựng mô hình dự đoán

##### Mô tả tập dữ liệu

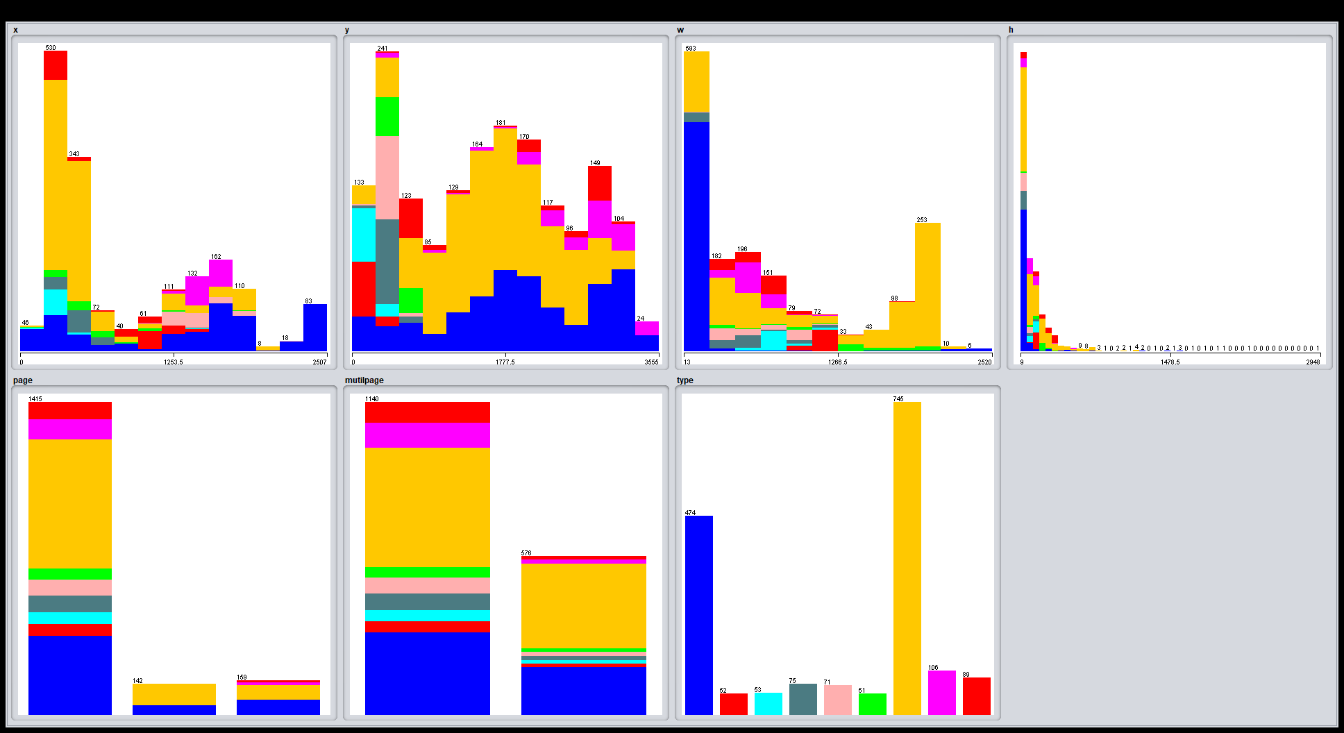
Dữ liệu được sử dụng là 86 tập tin hình ảnh của 56 văn bản hành chính trong đó bao gồm tổng hợp sau. Sau khi tiến hành xử lý bóc tách các thành phần và gán nhãn ta thu được 1716 mẫu dữ liệu.

Tổng số thuộc tính: 7 thuộc tính

* Thuộc tính lớp: type
* Các lớp:
  + 0: Thành phần lỗi
  + 1: Quốc hiệu và Tiêu ngữ
  + 2: Tên cơ quan tổ chức ban hành
  + 3: Số và ký hiệu
  + 4: Địa danh và thời gian ban hành
  + 5: Tên loại và trích yếu
  + 6: Phần nội dung
  + 7: Phần chữ ký
  + 8: Nơi nhận
* Tổng số mẫu tin:
  + Tổng số mẫu tin trong lớp 0: 474
  + Tổng số mẫu tin trong lớp 1: 52
  + Tổng số mẫu tin trong lớp 2: 53
  + Tổng số mẫu tin trong lớp 3: 75
  + Tổng số mẫu tin trong lớp 4: 71
  + Tổng số mẫu tin trong lớp 5: 51
  + Tổng số mẫu tin trong lớp 6: 745
  + Tổng số mẫu tin trong lớp 7: 106
  + Tổng số mẫu tin trong lớp 8: 89

Bảng 3.1 Bảng mô tả thuộc tính của tập dữ liệu

| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| --- | --- | --- | --- |
| 1 | x | Numeric | Tọa độ x của khối văn bản |
| 2 | y | Numeric | Tọa độ y của khối văn bản |
| 3 | w | Numeric | Chiều rộng của khối văn bản |
| 4 | h | Numeric | Chiều cao của khối văn bản |
| 5 | page | Nominal  {0, 1, 2} | Khối văn bản thuộc trang nào.   * 0 nếu khối văn bản thuộc trang đầu. * 1 nếu khối văn bản thuộc các trang ở giữa * 2 nếu khối văn bản nằm ở trang cuối |
| 6 | multiple | Nominal  {0, 1} | Số lượng trang của văn bản   * 0 nếu văn bản có 1 trang * 1 nếu văn bản có nhiều trang |
| 7 | type | Nominal | Loại của khối văn bản |



Hình 3.2 Biểu diễn dữ liệu

##### Khai khoáng dữ liệu

Từ tập dữ liệu ở trên ta tiến hành khai khoáng dữ liệu bằng nghi thức k-fold. Ở đây ta sẽ chạy thực nghiệm 10 lần với giá trị k lần lượt từ 5 đến 14. Bảng 3.2 thống kê kết quả của 10 lần chạy thực nghiệm.

Bảng 3.2 Thống kê các lần chạy thực nghiệm bằng nghi thức k-fold

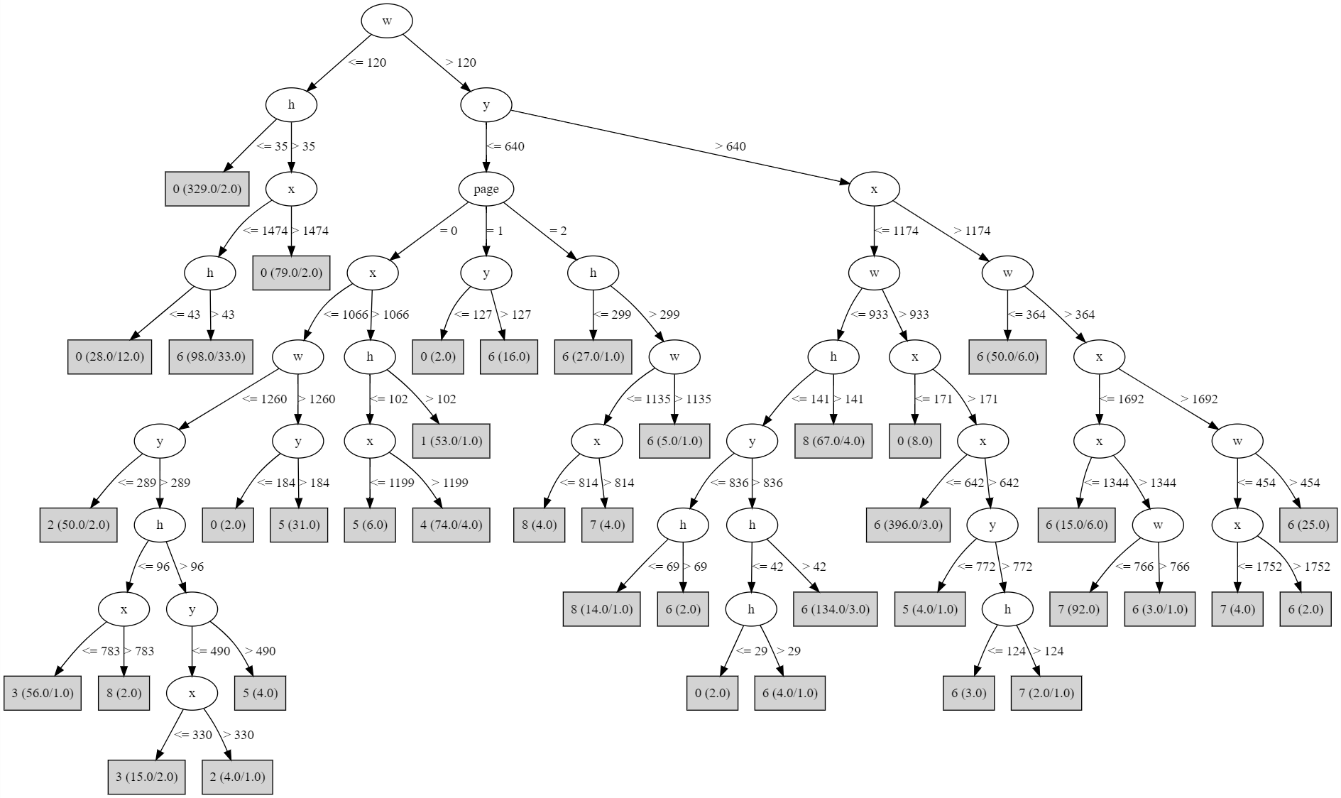
| **Lần chạy** | **k** | **Tổng số bản ghi** | **Số mẫu phân lớp đúng** | **Số mẫu phân lớp sai** | **Tỉ lệ phân lớp đúng** | **Tỉ lệ phân lớp sai** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | **5** | **1716** | **1575** | **141** | **91.7823 %** | **8.2168 %** |
| 2 | 6 | 1716 | 1570 | 146 | 91.4918 % | 8.5082 % |
| 3 | 7 | 1716 | 1572 | 144 | 91.6084 % | 8.3916 % |
| 4 | 8 | 1716 | 1566 | 150 | 91.2587 % | 8.7413 % |
| 5 | 9 | 1716 | 1571 | 145 | 91.5501 % | 8.4499 % |
| **6** | **10** | **1716** | **1575** | **141** | **91.7823 %** | **8.2168 %** |
| 7 | 11 | 1716 | 1559 | 157 | 90.8508 % | 9.1492 % |
| 8 | 12 | 1716 | 1571 | 145 | 91.5501 % | 8.4499 % |
| 9 | 13 | 1716 | 1570 | 146 | 91.4918 % | 8.5082 % |
| 10 | 14 | 1716 | 1564 | 152 | 91.1422 % | 8.8578 % |

Hình 3.3 Biểu đồ đường thống kê tỉ lệ phân lớp đúng của 10 lần thực nghiệm

Dựa trên biểu đồ thống kê ta thấy được k=5 và k=10 cho tỉ lệ phân lớp đúng cao nhất là 91.7823%. Ở đây ta sẽ chọn k=10 để đánh giá thuật toán.

##### Đánh giá kết quả

Từ kết quả phần khai khoáng dữ liệu ta được cây quyết định có kích thước 125 và số lượng lá là 65.



Hình 3.4 Kết quả cây quyết định với k=10

Kết quả ma trận Confusion (Bảng 3.3) là một ma trận vuông 9x9. Nhìn vào ta thấy lớp 0 có 39 mẫu dự đoán sai, tương tự lớp 1 có 1 mẫu, lớp 2 có 5 mẫu, lớp 3 có 11 mẫu, lớp 4 có 3 mẫu, lớp 5 có 13 mẫu, lớp 6 có 48 mẫu, lớp 7 có 6 mẫu và lớp 8 có 15 mẫu dự đoán sai.

Bảng 3.3 Ma trận Confusion

| **Dự đoán**  **Thực tế** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 435 | 0 | 0 | 2 | 0 | 2 | 35 | 0 | 0 |
| **1** | 0 | 51 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 48 | 4 | 1 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 5 | 64 | 1 | 3 | 2 | 0 | 0 |
| **4** | 0 | 1 | 0 | 0 | 68 | 2 | 0 | 0 | 0 |
| **5** | 0 | 1 | 1 | 3 | 2 | 38 | 6 | 0 | 0 |
| **6** | 31 | 0 | 0 | 0 | 2 | 2 | 697 | 5 | 8 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 100 | 1 |
| **8** | 1 | 0 | 0 | 1 | 0 | 2 | 11 | 0 | 74 |

Dưới đây là bảng thống kê chi tiết kết quả của giải thuật cây quyết định.

Bảng 3.4 Bảng thông số kết quả chạy giải thuật

| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **ROC Area** | **PRC Area** | **Class** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.918 | 0.026 | 0.931 | 0.918 | 0.925 | 0.896 | 0.969 | 0.903 | 0 |
| 0.981 | 0.001 | 0.962 | 0.981 | 0.971 | 0.971 | 0.989 | 0.925 | 1 |
| 0.906 | 0.004 | 0.889 | 0.906 | 0.897 | 0.894 | 0.967 | 0.757 | 2 |
| 0.853 | 0.007 | 0.853 | 0.853 | 0.853 | 0.847 | 0.946 | 0.791 | 3 |
| 0.958 | 0.004 | 0.919 | 0.958 | 0.958 | 0.935 | 0.976 | 0.868 | 4 |
| 0.745 | 0.007 | 0.776 | 0.745 | 0.745 | 0.753 | 0.934 | 0.689 | 5 |
| 0.936 | 0.061 | 0.922 | 0.936 | 0.936 | 0.873 | 0.956 | 0.923 | 6 |
| 0.943 | 0.003 | 0.952 | 0.943 | 0.943 | 0.944 | 0.979 | 0.949 | 7 |
| 0.831 | 0.006 | 0.892 | 0.831 | 0.831 | 0.854 | 0.930 | 0.789 | 8 |
| **0.918** | **0.035** | **0.918** | **0.918** | **0.918** | **0.884** | **0.961** | **0.892** | **Weighted Avg.** |

Đánh giá các giá trị quan trọng:

* TP Rate: Cao (91,8%)
* FP Rate: Thấp (3,5%)
* Precision: Cao (91,8%)
* Recall: Cao (91,8%)
* F1 (F-Measure): Cao (91,8%)

#### Phương pháp xử lý văn bản

Mô tả

Hình 3.5 Các bước xử lý văn bản

### PHÂN TÍCH VÀ THIẾT KẾ MÔ HÌNH

#### Sơ đồ Use Case

#### Sơ đồ ER



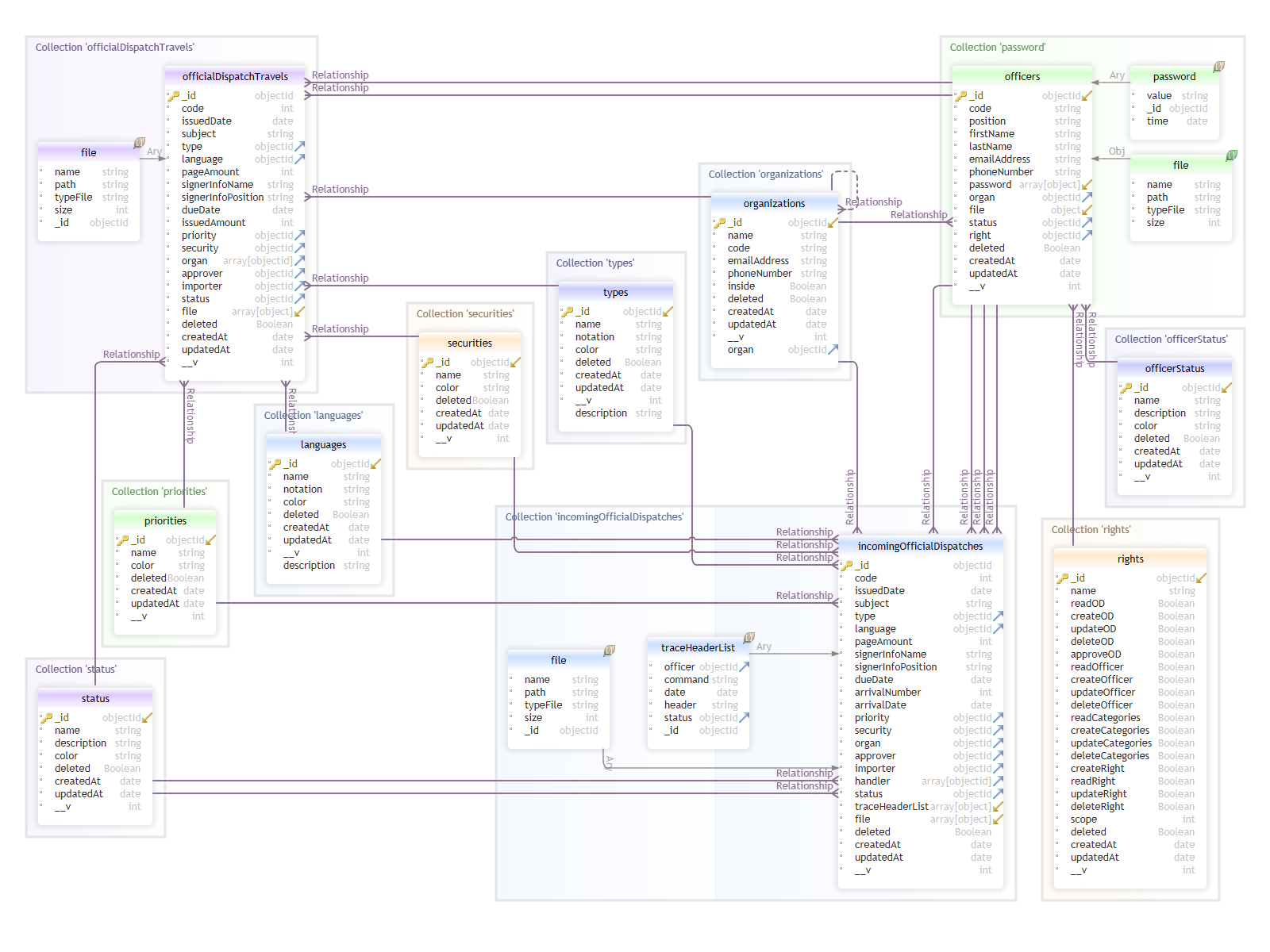
Hình 3.6 Sơ đồ ER

#### Sơ đồ LDM



Hình 3.7 Sơ đồ LDM

#### Sơ đồ LDM của MongoDB



Hình 3.8 Sơ đồ LDM của MongoDB

#### Sơ đồ phân rã chức năng



Hình 3.9 Sơ đồ phân rã chức năng

#### Sơ đồ hoạt động

##### Sơ đồ hoạt động chức năng đăng nhập

##### Sơ đồ hoạt động chức năng thêm công văn đến

##### Sơ đồ hoạt động chức năng sửa công văn đến

##### Sơ đồ hoạt động chức năng phê duyệt công văn đến

##### Sơ đồ hoạt động chức năng phân công xử lý công văn đến

##### Sơ đồ hoạt động chức năng xử lý công văn đến

##### Sơ đồ hoạt động chức năng từ chối công văn đến

##### Sơ đồ hoạt động chức năng xóa công văn đến

##### Sơ đồ hoạt động thống kê công văn đến

##### Sơ đồ hoạt động thêm cán bộ

##### Sơ đồ hoạt động sửa cán bộ

##### Sơ đồ hoạt động xóa cán bộ

#### Sơ đồ tuần tự

##### Sơ đồ tuần tự chức năng đăng nhập

##### Sơ đồ tuần tự chức năng thêm công văn đến

##### Sơ đồ tuần tự chức năng sửa công văn đến

##### Sơ đồ tuần tự chức năng phê duyệt công văn đến

##### Sơ đồ tuần tự chức năng phân công xử lý công văn đến

##### Sơ đồ tuần tự chức năng xử lý công văn đến

##### Sơ đồ tuần tự chức năng từ chối công văn đến

##### Sơ đồ tuần tự chức năng xóa công văn đến

##### Sơ đồ tuần tự thống kê công văn đến

##### Sơ đồ tuần tự thêm cán bộ

##### Sơ đồ tuần tự sửa cán bộ

##### Sơ đồ tuần tự xóa cán bộ

## KIỂM THỬ VÀ ĐÁNH GIÁ

Mô tả mục tiêu kiểm thử, kịch bản kiểm thử và kết quả kiểm thử: có chạy được hay không, chạy đúng không, đạt các mục tiêu đê ra hay không?

# PHẦN KẾT LUẬN

## KẾT QUẢ ĐẠT ĐƯỢC

## HƯỚNG PHÁT TRIỂN

# TÀI LIỆU THAM KHẢO

1. Memon J., Sami M., Khan R.A. và cộng sự. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access*, **8**, 142642–142668.

2. Thông T.T. (2019), *Ứng dụng học máy trong nhận dạng công văn các cơ quan Đảng tỉnh Quảng Bình*, (Luận văn thạc sĩ), Trường Đại học Bách khoa - Đại học Đà Nẵng, Đà Nẵng.

3. Mori S., Nishida H., và Yamada H. (1999), *Optical character recognition*, John Wiley & Sons, Inc.

4. Smith R. (2007). An overview of the Tesseract OCR engine. *Ninth international conference on document analysis and recognition (ICDAR 2007)*, IEEE, 629–633.

5. Hiệp T.Đ. (2013), *Nghiên cứu nhận dạng chữ in trong ảnh scan, ứng dụng vào trích lọc thông tin trích yếu của văn bản hành chính*, (Luận văn thạc sĩ), Trường Đại học Cần Thơ.

6. Nghị Đ.T. (2011). Chương 4 Máy học cây quyết định. *Khai mỏ dữ liệu: Minh họa bằng ngôn ngữ R*. Đại học Cần Thơ, Cần Thơ, 22.

7. Trung L.T. (2010), *Xây dựng cây quyết định trên tập dữ liệu nhỏ*, (Luận văn thạc sĩ), Trường Đại học Cần Thơ, Cần Thơ.

8. Đệ T.C. và Khang P.N. (2012). Phân loại văn bản với máy học Vector hỗ trợ và cây quyết định. *Tạp chí Khoa học Trường Đại học Cần Thơ*, (**21a**), 52–63.

9. Kiên P.T. (2019), *Khai thác và phân tích dữ liệu nhằm quản lý rủi ro trong giao dịch tín dụng*, (Luận văn thạc sĩ), Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, Hà Nội.

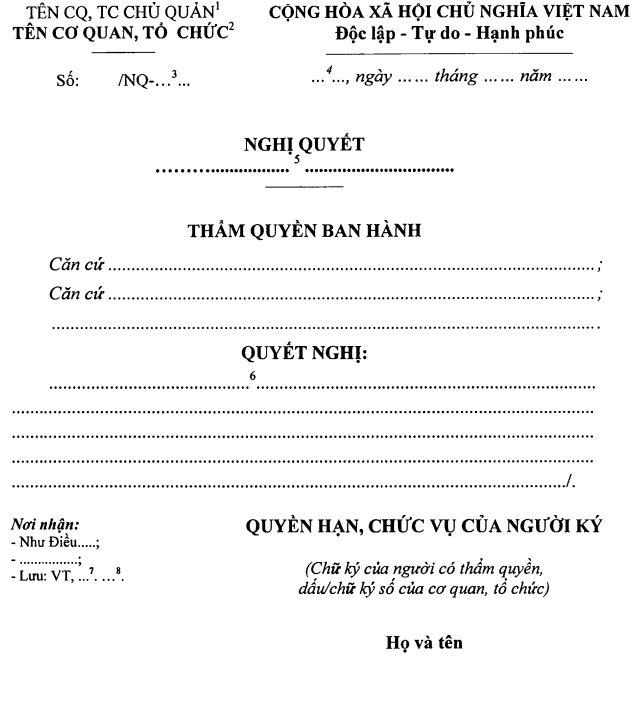
10. Poll: Deployed Data Mining Techniques. <https://www.kdnuggets.com/polls/2004/deployed\_data\_mining\_techniques.htm>, accessed: 16/11/2022.

11. Nghị Đ.T. (2011). Chương 2 Giải thuật k láng giềng. *Khai mỏ dữ liệu: Minh họa bằng ngôn ngữ R*. Đại học Cần Thơ, Cần Thơ, 12.

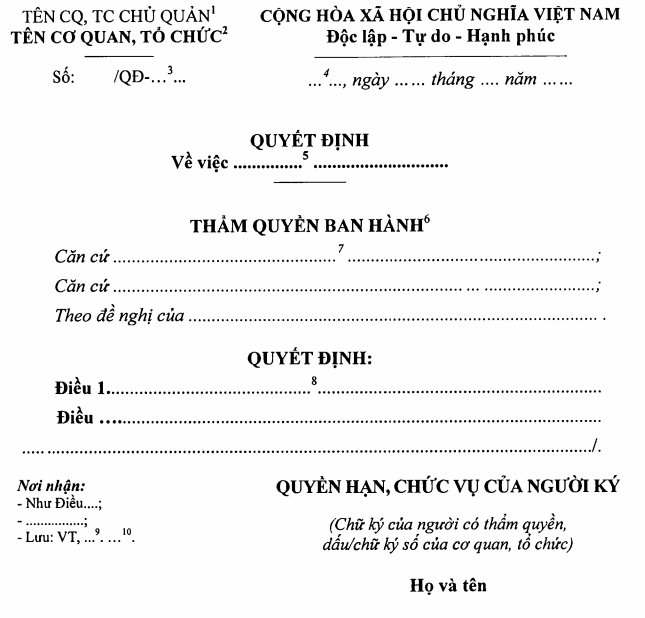
12. Chính phủ (2020), *Về công tác văn thư*, (Số 30/2020/NĐ-CP).

# PHỤ LỤC I BIỂU MẪU THƯỜNG GẶP

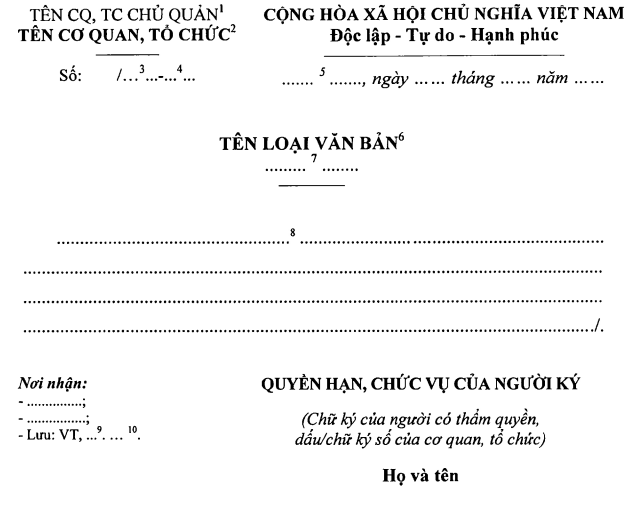
Mẫu 1 Nghị quyết (cá biệt)



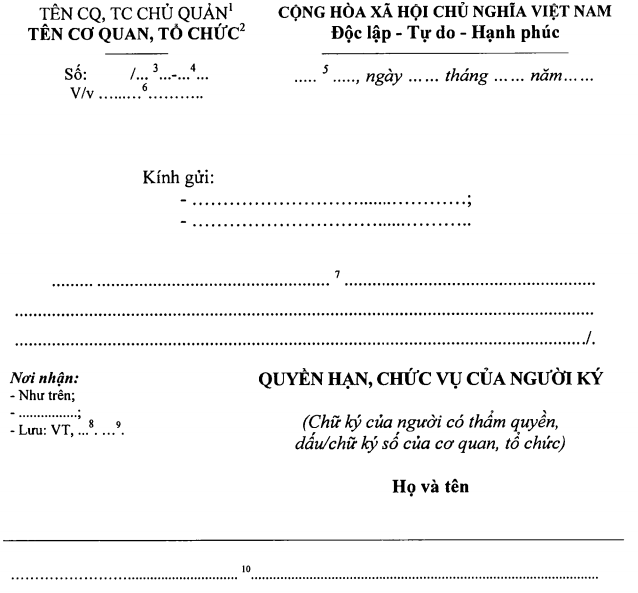
Mẫu 2 Quyết định



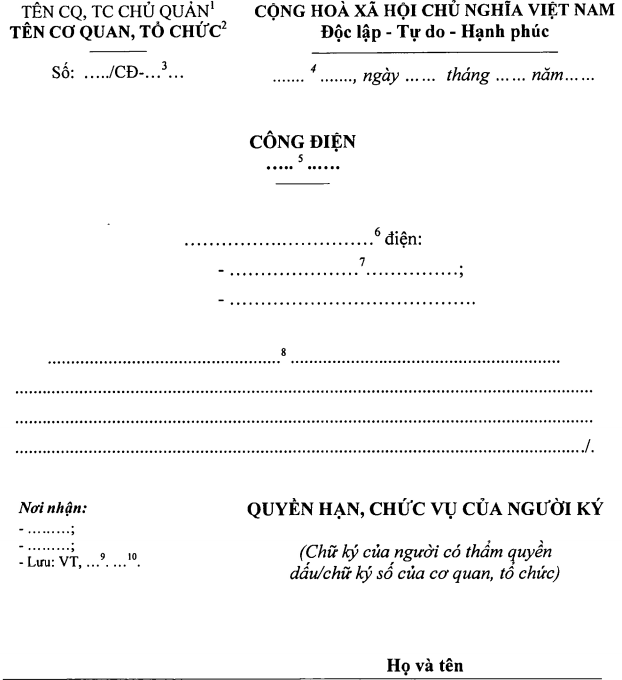
Mẫu 3 Văn bản có tên loại



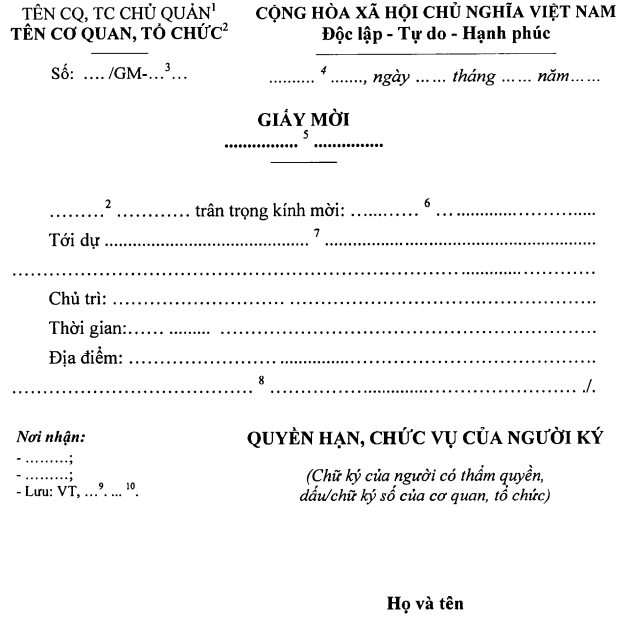
Mẫu 4 Công văn



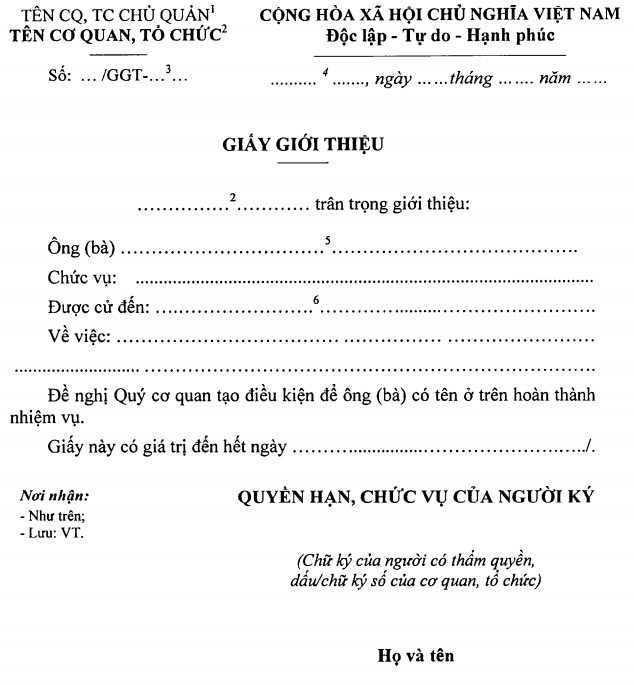
Mẫu 5 Công điện



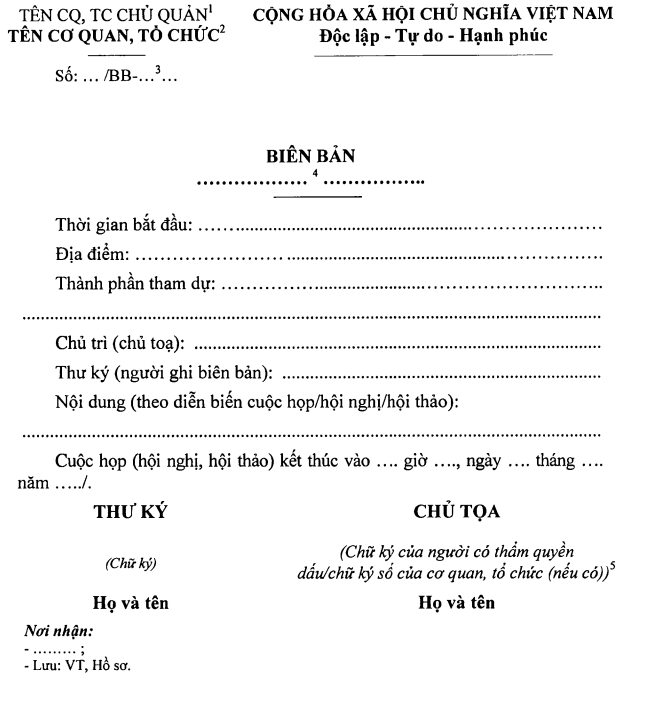
Mẫu 6 Giấy mời



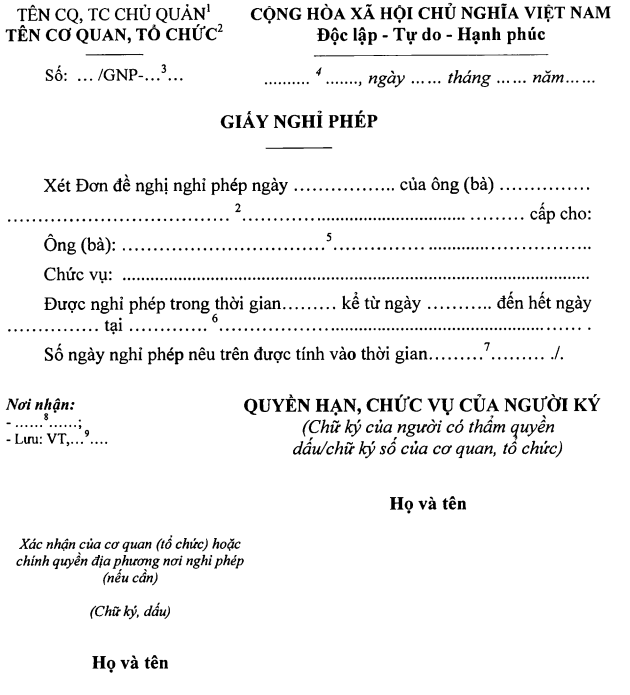
Mẫu 7 Giấy giới thiệu



Mẫu 8 Biên bản



Mẫu 9 Giấy nghỉ phép



# PHỤ LỤC II CÁC BẢNG DỮ LIỆU

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | Languages | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Ngôn ngữ văn bản | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | Notation | varchar | 10 |  |  |  |  |  |  |  | x |  |  | Ký hiệu |
| 4 | Color | varchar | 7 |  | #12B7BC |  | 7 | 7 |  |  | x |  |  | Màu |
| 5 | Description | varchar | 1000 |  |  |  |  |  |  |  |  |  |  | Mô tả |
| 6 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | Type | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Loại văn bản | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | Notation | varchar | 10 |  |  |  |  |  |  |  | x |  |  | Ký hiệu |
| 4 | Color | varchar | 7 |  | #12B7BC |  | 7 | 7 |  |  | x |  |  | Màu |
| 5 | Description | varchar | 1000 |  |  |  |  |  |  |  |  |  |  | Mô tả |
| 6 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | Security | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Độ mật | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | Priority | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Độ khẩn | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | Status | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Trạng thái văn bản | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | Description | boolean | 1000 |  |  |  |  |  |  |  |  |  |  | Mô tả |
| 4 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | OfficerStatus | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Trạng thái cán bộ | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | Color | varchar | 7 |  | #12B7BC |  | 7 | 7 |  |  | x |  |  | Màu |
| 4 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tên bảng** | | Right | | | | | | | | | | | | |
| **Mô tả tên bảng** | | Quyền cán bộ | | | | | | | | | | | | |
| **Thuộc tính** | | | | | | | | | | | | | | |
| STT | Tên thuộc tính | Kiểu | Kích thước | Số thập phân | Trị mặc định | Miền giá trị | Min | Max | Khóa chính | Duy nhất | Not null | Khóa ngoại | Ràng buộc toàn vẹn | Diễn giải |
| 1 | Id | varchar | 12 |  |  |  |  |  | x |  | x |  |  | ID |
| 2 | Name | varchar | 100 |  |  |  |  |  |  |  | x |  |  | Tên |
| 3 | ReadOD | varchar | 7 |  | #12B7BC |  | 7 | 7 |  |  | x |  |  | Màu |
| 4 | Deleted | boolean |  |  | False |  |  |  |  |  | x |  |  | Đã xóa |