MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

LỜI CẢM ƠN

MỞ ĐẦU

Ngày nay, công nghệ và máy móc đang phát triển môt cách nhanh chóng và tiếp cận với hầu hết tất cả mọi người, phục vụ cho con người trong hầu hết tất cả các lĩnh vực. Cùng với đó là lượng thông tin và kiến thức ngày càng lớn làm cho nhu cầu lưu trữ của con người cũng trở nên thiết yếu hơn nhưng thay vì lưu trữ dưới dạng truyền thống là sách báo, hay tài liệu giấy thì ngày nay, chúng ta có thể lưu trữ dúng dưới dạng thông tin số một cách nhanh chóng và tiện lợi hơn nhiều so với tài liệu giấy truyền thống. Tài liệu số có những ưu điểm vượt trội như giảm không gian lưu trữ, dễ dàng sao lưu, quản lý, hạn chế mất mát, dễ dàng tìm kiếm và khai thác thông tin, chi phí vận hành thấp nhưng hiệu quả sử dụng cao. Do đó, việc tài liệu số sẽ thay thế dần cho tài liệu giấy là điều tất yếu/

Đi cùng với nhu cầu lưu trữ thông tin thì nhu cầu khai thác thông tin cũng là một khía cạnh quan trọng phục vụ cho mục đích học tập và nghiên cứu. Thông tin được con lưu trữ là vô cùng lớn và liên quan đến tất cả các lĩnh vực mà con người biết đến, từ khoa học, y tế, kinh tế, giáo dục cho đến văn học nghệ thuật. Việc khai thác thông tin từ tài liệu có sẵn sẽ giúp cho chúng ta tiếp cần với tri thức một cách nhanh chóng và hiệu quả.

Tài liệu số là các dạng tại liệu dạng văn bản, hình ảnh, âm thanh, … được máy tính chấp nhận và lưu trữ dưới dạng các tệp. Trong số các định dạng tài liệu số phổ biến, PDF(Portable Document Format) là một trong những định dạng được sử dụng rộng rãi nhất. PDF được phát triển bởi Adobe Systems với mục đích tạo ra một định dạng tài liệu có khả năng hiển thị nhất quán trên mọi thiết bị và hệ điều hành, có nghĩa là nó sẽ hiển thị giống nhau trên mọi loại thiết bị máy tính, điện thoại di động hay máy tính bảng. Định dạng PDF còn hỗ trợ lưu trữ văn bản và hình ảnh, ngoài ra nó còn hỗ trợ nhúng các tệp âm thanh, video và các đường dẫn tới các tài liệu khác. Những lý do đã biến PDF trở thành định dạng lý tưởng cho mục đích lưu trữ thông tin.

Tuy nhiên, những thông tin được lưu trữ không phải hoàn toàn có ích. Do vậy mà việc trích xuất thông tin từ các tài liệu PDF một cách hiệu quả là vô cùng quan trọng. Việc trích xuất thông tin trong tài liệu PDF sẽ hỗ trợ cho việc tiếp thu thông tin cần thiết một cách nhanh chóng, tìm kiếm, phân tích và sử dụng tài liệu hiệu quả và có giá trị cao nhất, giúp cho quá trình nghiên cứu và học tập hiệu quả hơn. Nhận thấy bài toán trích xuất quản lý và chia sẻ thông tin từ tệp PDF là một nhu cầu cần thiết nên em chọn đề tài “Xây dựng hệ thống trích xuất và quản lý thông tin từ tệp PDF”. Đề tài này không chỉ giải quyết bài toán trích xuất thông tin từ các tệp PDF mà còn nhằm mục đích tạo ra một công cụ có thể chia sẻ và quản lý các tệp PDF và thông tin được trích xuát, phục vụ cho nhiều đối tượng khác nhau và nhiều mục đích khác nhau như chia sẻ kiến thức, thông tin và các tệp PDF.

Nội dung luận văn sẽ được trình bày trong 3 chuơng:

CHƯƠNG 1. TỔNG QUAN VỀ BÀI TOÁN VÀ CƠ SỞ LÝ THUYẾT

Trong chương này luận văn sẽ tìm hiểu và khảo sát các nội dung cơ bản về hệ thông trích xuất thông tin, các kĩ thuật và cơ chế nhận dạng dữ liệu và trích xuất thông tin, xây dựng giải pháp trích xuất thông tin đồng thời đề xuất phương án quản lý và chia sẻ thông tin thu thập được.

CHƯƠNG 2. XÂY DỰNG HỆ THỐNG TRÍCH XUẤT THÔNG TIN TỪ TỆP PDF

Trong chương này luận văn sẽ phân tích thiết kế hệ thống từ cơ sở được đề cập trong chương 1 bao gồm:

* Xây dựng module trích xuất thông tin từ tệp PDF
* Xây dựng máy chủ lưu trữ và quản lý thông tin
* Xây dựng module đẩu cuối khai thác thông tin

CHƯƠNG 3. TRIỂN KHAI VÀ ĐÁNH GIÁ HỆ THỐNG

Trong chương này luận văn sẽ xây dựng, triển khai chi tiết các module đã thiết kế và kết luận những kết quả đã đạt được và đề xuất hướng phát triển.

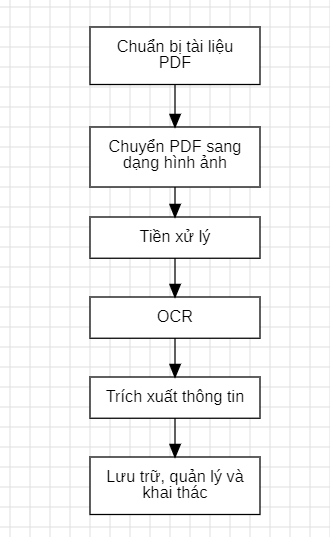
CHƯƠNG 1. TỔNG QUAN VỀ BÀI TOÁN VÀ CƠ SỞ LÝ THUYẾT

* 1. Tổng quan về bài toán nhận dạng và trích xuất thông tin

Bài toán nhận dạng và trích xuất thông tin bao gồm hai bài toán con là bài toán nhận dạng văn bản (Optical Character Recognition - OCR) và bài toán trích xuất thông tin (Information Extraction - IE). Nhận dạng văn bản (Optical Character Recognition - OCR) là quá trình tự động phát hiện và chuyển các tài liệu dạng hình ảnh có chứa văn bản chữ viết tay hoặc tài liệu đánh máy (thường được scan bằng máy) về dạng tài liệu văn bản.

Trích xuất thông tin (Information Extraction - IE) là quá trình tự động trích xuất các thông tin quan trọng từ các tài liệu có cấu trúc hoặc không có cấu trúc. Mục đích chuyển đổi văn bản về dạng thông tin ngắn gọn, cụ thể, dễ dàng tiếp thu, phân tích và xử lý.

Các bước thực hiện sẽ trải qua các giai đoạn theo sơ đồ dưới đây:



Diễn giải:

- Các tài liệu PDF sẽ được chuẩn bị cho quá trình nhận dạng

- Các tệp sẽ được tiền xử lý để chuyển về dang dạng hình ảnh trắng đen và lọc nhiễu, sau đó được nhận dạng bằng công nghệ OCR

- Dữ liệu sau khi được nhận dạng sẽ được trích xuất các thông tin quan trọng

- Các dữ liệu nguyên bản được nhận dạng và dữ liệu được trích xuất sẽ được lưu trữ để phục vụ cho quá trình khai thác.

Quy trình:

Bước 1: Chuẩn bị tài liệu PDF

Bước 2: Chuyển tài liệu về dạng hình ảnh

Bước 3: Tiền xử lý

Bước 4: Nhận dạng văn bản

Bước 5: Trích xuất thông tin

Bước 6: Lưu trữ, quản lý và khai thác thông tin

1.2. Phương pháp nhận dạng và trích xuất thông tin từ file PDF

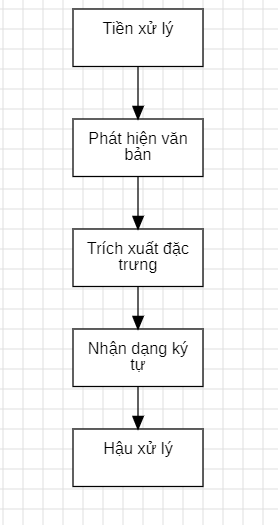
1.2.1. Phương pháp nhận dạng văn bản

1.2.1.1. Phương pháp nhận dạng văn bản

Nhận dạng văn bản hay nhận dạng kí tự quang học (Optical Character Recognition) là kĩ thuật chuyển các hình ảnh chứa văn bản về dạng văn bản có thể chỉnh sửa một cách tự động. Công nghệ này đóng vai trò quan trọng trong nhiều lĩnh vực như số hóa tài liệu, xử lý dữ liệu và tự động hóa công việc văn phòng, giúp tiết kiệm thời gian, công sức so với việc nhập dữ liệu thủ công, tăng hiệu quả trong khai thác dữ liệu.

1.2.1.2. Cơ chế nhận dạng văn bản

Cơ chế chung của quá trình nhận dạng văn (Optical Character Recognition - OCR) bản gồm các bước xử lý với mục đích chuyển hình ảnh chứa văn bản thành văn bản mà máy tính có thể chỉnh sửa. Quá trình bao gồm các bước sau:



Diễn giải:

- Tiền xử lý (Preprocessing): là quá trình chỉnh sửa đầu vào với mục đích tăng chất lượng hình ảnh cho quá trình nhận dạng bao gồm các bước:

* + - * Chuyển đổi thành thang độ xám (Grayscale Conversion): Loại bỏ màu sắc để giảm độ phức tạp của hình ảnh.
      * Chuyển thành dạng nhị phân (Binaryization): Chuyển văn bản về dạng trắng đen để làm nổi bật văn bản lên so với nền. Người ta thường sử dụng môt ngưỡng giới hạn để phân loại cho pixel. Ví dụ: pixel có giá trị lớn hơn 128 thì gán giá trị là 255, ngược lại thì gán giá trị là 0.
      * Lọc nhiễu (Noise Reduction): Loại bỏ nhiễu không mong muốn trong hình ảnh
      * Căn chỉnh (Deskewing): Sửa các hình ảnh bị nghiêng để đảm bảo các dòng văn bản thẳng hàng.
* Phát Hiện Văn Bản (Text Detection): Trong bước này các vùng chứa văn bản sẽ được phát hiện bằng các kĩ thuật như phát hiện biên (Edge Detection).
* Phát hiện biên (Edge Detection): Là phương pháp nhân ma trận ảnh với một ma trận lọc nhằm giữ lại các cạnh biên ngang hoặc biên dọc của các vật thể trong hình ảnh, từ đó phát hiện ra vị trí, hình dạng của vật thể.
* Tìm đường viền (Contour Detection): Là phương pháp xác định đường biên bằng cách nối các điểm liên tiếp có cùng màu hoặc mật độ
* Nhận Dạng Ký Tự (Character Recognition): Giai đoạn này sẽ nhận dạng từng kí tự bằng cách so khớp mẫu, tách từng kí tự và so sánh với mẫu có sẵn hoặc tách các đặc trưng như đường cong, nét thẳng, vòng tròn kín để so sánh với các đăng trưng của các kí tự mẫu từ đó đưa ra dự đoán
* Hậu Xử Lý: Cải thiện kết quả dự đoán được như sửa lỗi chính tả, ghép từ thành câu, …

1.2.1.2. Khảo sát về Tesseract OCR Engine

Tesseract là một công cụ OCR mã nguồn mở được phát triển tại HP từ năm 1984 đến năm 1994. Ban đầu, Tesseract chỉ chấp nhận các hình ảnh văn bản có dạng một cột làm đầu vào, các phiên bản đầu tiên này không có tính năng phân tích bố cục văn bản. Do vậy, khi đầu vào là các hình ảnh văn bản nhiều cột như báo, hóa đơn, … sẽ tạo ra các đầu ra bị chắp vá. Từ phiên bản 3.0, Tesseract đã hỗ trợ tính năng phân tích bố cục và phảt hiện xem hình ảnh đầu vào là đơn cột hay đa cột. Các phiên bản ban đầu cũng chỉ hỗ trợ tiếng anh, từ phiên bản 3.0 trở đi đã có thêm nhiều ngôn ngữ khác được thêm vào kể cả tiếng việt. ngoài ra Tesseract cũng hỗ trợ cả ngôn ngữ đọc từ phải qua trái.

Cơ chế hoạt động của Tesseract gồm các bước như sau:

* Tìm dòng và từ
* Nhận dạng từ
* Phân loại ký tự
* Phân tích ngôn ngữ
* Phân loại tương thích
* Xuất kết quả

Diễn giải

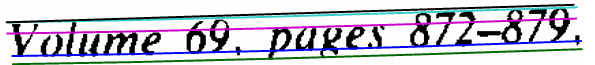
* Giai đoạn tìm dòng được bắt đầu là bước tìm dòng. Thuật toán của Tesseract được thiết kế để có thể tìm được các dòng văn bản bị lệnh mà không cần điều chình độ lệch của hình ảnh, giảm thiểu rủi ro chất lượng hình ảnh bị giảm sút khi phải điều chỉnh. Mục đích chính của bước này là lọc blob (Binary Large Object) và xây dựng các đường thẳng.

Lọc blob (Binary Large Object) sẽ nhận diện các vùng liên tục có giá trị pixel tương tự nhau, thường là các vùng đen trên nền trắng mà khả năng cao chúng sẽ đại diện cho các kí tự của văn bản, tiếp đó, các blob có chiều cao tương tự nhau sẽ được giữ lại, loại bỏ đi các blob có chiều cao lớn hơn hoặc nhỏ hơn nhiều so với giá trị chiều cao trung bình, các giá trị bị loại bỏ này có thể là các dấu câu hoặc các chữ cái đầu đoạn (Drop cap) thường dùng để tạo điểm nhấn trong các bài báo. Sau khi đã loại bỏ và giữ lại các blob có chiều cao tương tự nhau, các kí tự chạm dọc cũng sẽ được loại bỏ để tránh ảnh hưởng đến quá trình nhận dạng. các kí tự chạm dọc là các kí tự kéo dài, chạm hoặc nối liền với nhau ví dụ như chữ “g” ở hàng trên có thể chạm vào chữ “d” ở hàng dưới.

Các blob đã được lọc sẽ là các dòng không chồng chéo nhưng có thể có độ nghiêng, các blob có tọa độ x tương tự nhau sẽ được gán vào một dòng văn bản duy nhất. Trong khi gán, độ nghiêng của các blob sẽ được theo dõi. Để hạn chế việc gán sai dòng, các phép bình phương phương vị nhỏ nhất sẽ được sử dụng để ước lượng lại độ nghiên của văn bản và sau đó, các blob sẽ được sắp xếp lại sao cho đúng với các dòng phù hợp.

Bước tiếp theo của giai đoạn tìm dòng là gán lại các blob đại diện cho dấu câu, các chữ cái đầu đoạn và các kí tự chạm dọc đã bị loại bỏ vào các dòng mà chúng bị loại bỏ.

Bước cuối cùng của quá trình tìm dòng là gán đường cơ sở. Khi các dòng văn bản đã được tìm thấy, các đường cơ sở sẽ được gán vào chính xác hơn.



Hình 3. Ví dụ về đường cơ sở

Hình 3 mô tả hai đường cơ sở chặn ở trên và dưới các dòng và hai đường trung bình nằm giữa hai đường cơ sở

Sau khi tìm dòng và gán đường cơ sở, Tesseract xác định độ rộng của các kí tự và chia nhỏ các blob thành các kí tự theo độ rộng và đánh dấu các từ này để nó không bị cắt trong quá trình nhận dạng từ.

* Tiếp sau đó, Tesseract đo các khoảng cách trống được giới hạn giữa các đường trung bình để tìm ngưỡng khoảng cách giữa các từ, Các khoảng cách gần với ngưỡng ở giai đoạn này được làm mờ, để quyết định cuối cùng có thể được đưa ra sau bước nhận dạng từ.
* Giai đoạn phân loại kí tự, các kí tự cần nhận dạng sẽ được so khớp mẫu với các kí tự đã được huẩn luyện, các đặc trưng như đường viền, độ dài, ngắn, dày sẽ được trích xuất và chuẩn hóa để so sánh với các đặc trưng của mẫu huấn luyện.

Các kí tự mẫu sẽ được rút gọn thành một danh sách mà kí tự chưa biết có thể khớp sau đó sẽ được tính toán độ tương tự để đưa ra kí tự phù hợp nhất.

* Giai đoạn phân tích ngôn ngữ sẽ so sánh từ với các từ mẫu được huần luyện theo thứ tự ưu tiên như từ phổ biến, từ trong từ điển, từ bắt đầu bằng chữ hoa, v.v. Các từ được so sánh trực tiếp với mẫu để xác định ngôn ngữ được sử dụng trong văn bản
* Giai đoạn phân loại tương thích sẽ đưa các ký tự về cùng một kích thước trong một font cụ thể nhằm dễ dàng phân loại chữ hoa và chữ thường cũng như loại bỏ nhiễu.
* Sau khi các bước trên được thực hiện thì kết quả là các đoạn văn bản đã được nhận dạng với độ chính xác lên đến 83%

1.2.2. Phương pháp trích xuất thông tin

1.2.2.1. Phương pháp trích xuất thông tin

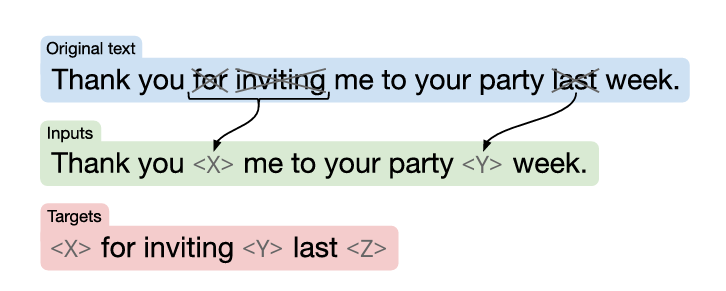
Khai thác văn bản (hay còn gọi là khai phá văn bản, tiếng Anh: text mining hoặc text data mining) là một quá trình xử lý và trích xuất thông tin nằm trong văn bản, quá trình này là một phần của việc phân tích văn bản trong khai thác dữ liệu. Thông tin được thể hiện dưới dạng các mẫu, xu hướng, thứ tự sắp xếp được trích xuất thông qua các luật hoặc thông qua quá trình học dựa trên các mẫu thống kê.

Các bước cơ bản bao gồm chuẩn bị dữ liệu, mã hóa dữ liệu đầu vào thành các chuỗi token, huẩn luyện mô hình, tinh chỉnh và sử dụng mô hình để trích xuất dữ liệu. Dữ liệu được chuẩn bị có dạng hai đoạn văn bản được ghép cặp với nhau như là mẫu và nhãn để sử dụng cho quá trình học. Các văn bản đầu vào gồm cả văn bản mẫu và nhãn sẽ được mã hóa thành các chuỗi token, các mô hình sẽ được huấn luyện để tìm quy luật thay đổi giữa token của văn bản mẫu và token của nhãn và sau đó được tinh chỉnh để có thể sử dụng cho quá trình trích xuất thông tin.

1.2.2.2. Khảo sát model T5 (Text-to-Text Transfer Transformer)

Mô hình T5 (Text-to-Text Transfer Transformer) hay mô hình biến đổi văn bản thành văn bản sử dụng bộ biến đổi Transformer là một mô hình chuyển văn bản thành văn bản dựa trên kiến trúc Transformer và được đào tạo theo phương pháp học chuyển giao bao gồm học không có giám sát trước trên một tập dữ liệu lớn, sau đó được học đa tác vụ có giám sát, cuối cùng được tinh chỉnh lại để đạt kết quả tối ưu. Mô hình T5 được huấn luyện trên một tập dữ liệu vô cùng lớn bao gồm 20TB dữ liệu được thu thập mỗi tháng từ các văn bản trích xuất từ các trang web và được làm sạch. Cấu trúc của T5 là một bộ biến áp mã hóa và giải mã bao gồm các ngăn xếp lớp, bộ mã hóa chuỗi đầu vào, bộ biến áp Transformer và bộ giải mã để tạo chuỗi đầu ra mới được thiết kế để thực hiện một cách tuần tự

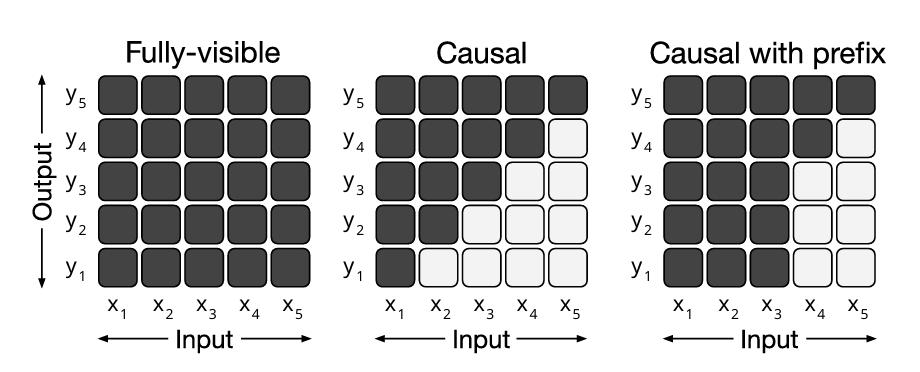
Quá trình học không có giám sát cảu T5 được thể hiện như hình dưới đây:



Hình 4. Ví dụ về học không có giám sát

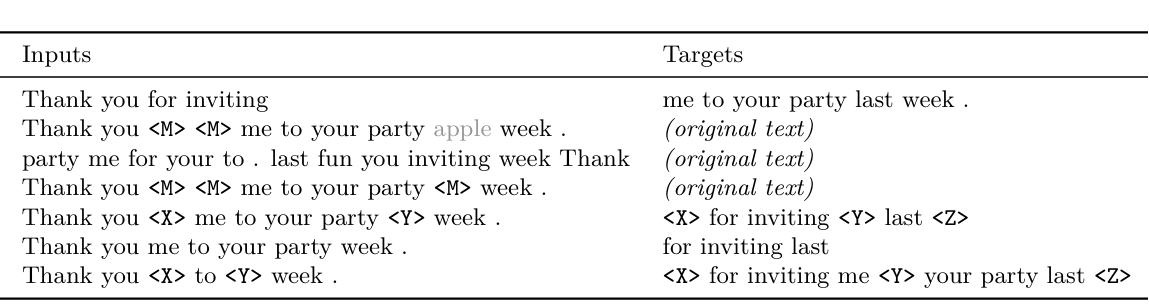
Các văn bản đầu vào được loại bỏ ngẫu nhiên các từ và được thay thế bằng các mã thông báo trọng điểm (Sentinel token) <X> và <Y>. Chuỗi đầu ra sẽ bao gồm các từ bị loại bỏ và các thông báo trọng điểm dùng để thay thế cho chuỗi đầu vào và một mã thông báo trọng điểm được sử dụng để đánh dầu điểm cuỗi <Z>. Mục tiêu của quá trình học này là để mô hình có thể khôi phục lại văn bản gốc bằng cách dự đoán các từ bị thiếu.

Bộ biến áp Transformer của mô hình T5 được dựa trên cơ chế tự chú ý sử dụng các “mặt nạ chú ý” với nhiệm vụ nhận một chuỗi đầu vào và tạo ra một chuỗi đầu ra với kích thước tương tự. mỗi chuỗi đầu ra sẽ được tính bằng trung bình có trọng số của các phần tử trong chuỗi đầu vào với là phần tử thứ j và trọng số vô hướng được tạo ra bởi cơ chế tự chú ý. Công thức tính như sau: . Mặt nạ chú ý sẽ được sử dụng để giảm bớt hoặc loại bỏ một số các trọng số được coi là không cần thiết.



Hình 5. Ví dụ về mặt đạ đầy đủ, mặt nạ nguyên nhân và mặt nạ nguyên nhân kết quả

Quá trình học đa tác vụ hay học đa nhiệm là quá trình học có giám sát dựa trên tập dữ liệu với các đầu vào bị che dấu 1 phần dữ liệu hoặc bị hỏng và sử dụng các nhãn với mục đích khôi phục chuỗi đầu vào hoặc tạo mới hoàn toàn chuỗi đầu vào. Các nhiệm vụ sẽ được huấn luyện cùng một lúc và được kết hợp với quá trình tinh chỉnh để tối ưu hiệu suất.



Hình 6. Ví dụ về mục tiêu học có giám sát

* 1. Xây dựng giải pháp

Dựa vào các kiến thức em đã tìm hiểu được và đã phân tích ở phần trên, giải pháp xây dựng hệ thống trích xuất thông tin từ tệp PDF mà em đưa ra bao gồm ba phần chính:

* Xây dựng module trích xuất thông tin từ tệp PDF sử dụng công cụ Tesseract và mô hình T5
* Xây dựng máy chủ lưu trữ và quản lý thông tin
* Xây dựng module đẩu cuối khai thác thông tin là ứng dụng di động có nhiệm vụ giúp người dùng khai thác thông tin, lưu trữ và chia sẻ tệp PDF
  1. Công nghệ sử dụng
* Ngôn ngữ lập trình: kotlin, python
* Công cụ lập trình: Visual studio code, Android studio
* Cơ sở dữ liệu: MongoDB, Redis
* Webserver: Flask, Flask socket-io

1.4. Tổng kết chương

CHƯƠNG 2. XÂY DỰNG HỆ THỐNG TRÍCH XUẤT THÔNG TIN TỪ TỆP PDF

2.1. Đặc tả yêu cầu hệ thống

2.2. Phân tích, thiết kế hệ thống

2.2.1. Biểu đồ ca sử dụng

2.2.2. Biểu đồ tuần tự/hoạt động

2.2.3. Thiết kế CSDL

2.2.4. Thiết kế giao diện và chức năng

2.5. Tổng kết chương

CHƯƠNG 3. TRIỂN KHAI VÀ ĐÁNH GIÁ HỆ THỐNG

3.3. Xây dựng hệ thống

3.3.1. Kiến trúc hệ thống

3.3.2. Xây dựng các mô-đun chương trình

3.4. Thử nghiệm và đánh giá kết quả

3.5. Tổng kết chương

KẾT LUẬN

TÀI LIỆU THAM KHẢO

PHỤ LỤC