# **LỜI CẢM ƠN**

Trong thời gian làm đồ án tốt nghiệp, em đã nhận được nhiều sự giúp đỡ,đóng góp ý kiến và chỉ bảo nhiệt tình của thầy cô, gia đình và bạn bè.

Em xin gửi lời cảm ơn chân thành đến thầy Nguyễn Quốc Khánh, giáo viên khoa công nghệ thông tin - trường học viện kỹ thuật quân sự người đã tận tình hướng dẫn, chỉ bảo em trong suốt quá trình làm đồ án.

Em cũng xin chân thành cảm ơn các thầy cô giáo trong trường Học viện kỹ thuật quân sự nói chung, các thầy cô trong Bộ môn Công nghệ phần mềm nói riêng đã dạy dỗ cho em kiến thức về các môn đại cương cũng như các môn chuyên ngành, giúp em có được cơ sở lý thuyết vững vàng và tạo điều kiện giúp đỡ em trong suốt quá trình học tập.

Cuối cùng, em xin chân thành cảm ơn gia đình và bạn bè, đã luôn tạo điều kiện, quan tâm, giúp đỡ, động viên em trong suốt quá trình học tập và hoàn thành khoá luận tốt nghiệp.

Hà Nội ngày 07/03/2017

Sinh viên thực hiện

Hoàng Thị Lanh

# **Danh mục hình ảnh**

Hình 2.1: [Grammarly.com](http://grammarly.com/)

Hình 2.2: Trang chủ của hệ thống

Hình 2.3 : Màn hình sau khi đăng nhập hệ thống

Hình 2.4: Màn hình kết quả sau khi hệ thống kiểm tra

Hình 2.5: Màn hình Gmail thông báo kết quả sau khi hệ thống kiểm tra

# **Danh mục từ viết tắt**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# LỜI NÓI ĐẦU

Hệ thống chữ tiếng việt là một hệ thống chữ viết linh động, thông minh, có thể dễ dàng biểu thị tiếng nói dưới dạng chữ viết chỉ cần dựa trên những luật nhất định mà không đòi hỏi quá nhiều về mặt kiến thức từ vựng như nhiều ngôn ngữ khác(tiếng Anh, Pháp, Trung…). Tuy nhiên, việc giao tiếp hàng ngày chủ yếu qua tiếng nói, điều này dẫn đến việc có nhiều cách phát âm không giống nhau giữa các vùng miền khác nhau đối với cùng một từ. Việc này dẫn đến mắc lỗi sai chính tả trong khi viết. Thêm vào đó, công nghệ thông tin ngày càng phát triển, việc soạn thảo văn bản bằng hình thức viết tay dần được thay thế bằng soạn thảo văn bản qua hình thức đánh máy. Do đó xuất hiện thêm một loại lỗ chính tả nữa là do lỗi đánh máy gây ra.

Đặc biệt, trong các trường đại học, khi thu nộp luận văn tốt nghiệp thường mất rất nhiều thời gian, do tình trạng sai lỗi chính tả và định dạng văn bản quá nhiều.

Trên khảo sát thực tế, đã có 1 số hệ thống đang thực hiện để khắc phục các vấn đề trên, tuy nhiên vẫn còn nhiều hạn chế. Do đó em xin lựa chọn đề tài: “Xây dựng hệ thống kiểm tra lỗi chính tả và định dạng văn bản” để hoàn thiện những vấn đề đang tồn tại ở trên. Hệ thống mà em xây dựng nhằm mục đích tiết kiệm thời gian, công sức, tạo sự chính xác trong việc kiểm tra, thu nộp các bài luận văn trong nhà trường.

# Chương 1

# KHẢO SÁT HỆ THỐNG

1. Khảo sát về kiểm tra lỗi chính tả và định dạng văn bản khi thu nộp các bài luận văn hiện nay.
2. Bối cảnh

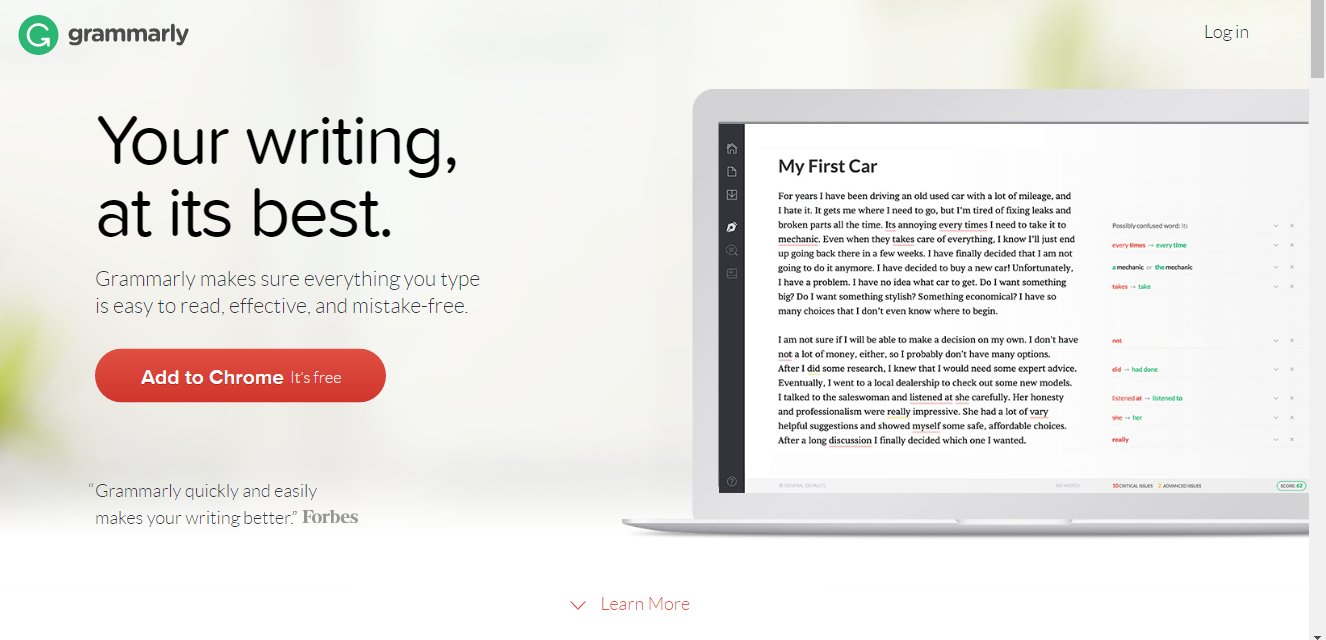
Hiện nay, công tác thu nộp luận văn trong nhà trường nói chung và Học Viện Kĩ Thuật Quân sự nói riêng đang được tiến hành 1 cách thủ công gây mất nhiều thời gian. Cụ thể tại Học Viện Kĩ Thuật quân sự thì quy trình này được thực hiện như sau:

* Đầu tiên: sinh viên tự kiểm tra về lỗi chính tả và định dạng văn bản trong bài luận văn của mình.
* Tiếp theo: Sinh viên nộp luận văn đó cho bộ phận quản lí sinh viên, bộ phận này sẽ kiểm tra 1 cách thủ công về lỗi chính tả và định dạng cho bài luận văn đó.
* Nếu bài luận văn đó đạt yêu cầu thì thu lại
* Nếu sai chính tả hoặc định dạng thì sẽ trả lại cho sinh viên về chỉnh sửa.
* Sinh viên chỉnh sửa xong lại nộp cho cán bộ quản lí.
* Tiếp tục như vậy cho đến khi cán bộ quản lí kiểm tra thấy bài luận văn đó đạt yêu cầu: không còn lỗi và định dạng đã chuẩn theo quy định.

1. Khó khăn của quy trình

* Cách kiểm tra hiện tại được tiến hành 1 cách thủ công, độ chính xác khi kiểm tra lỗi chính tả và định dạng văn bản không cao.
* Tốn rất nhiều thời gian cho sinh viên, cho cán bộ quản lí.
* Lãng phí tiền bạc trong việc in đi in lại quá nhiều lần báo cáo.

1. Nghiên cứu các hệ thống có liên quan
2. Hệ thống [Grammarly.com](http://grammarly.com/)



Hình 2.1: [Grammarly.com](http://grammarly.com/)

Grammarly.com: Là công cụ hỗ trợ phát triển lỗi chính tả, lỗi soạn thảo văn bản và có gợi ý để chỉnh sửa khi bạn viết 1 bài viết tiếng anh 1 cách miễn phí.

Ưu điểm:

* Cho phép người dùng tạo tài khoản miễn phí
* Kiểm tra, phát hiện lỗi chính tả, lỗi soạn thảo 1 cách chính xác và nhanh chóng.

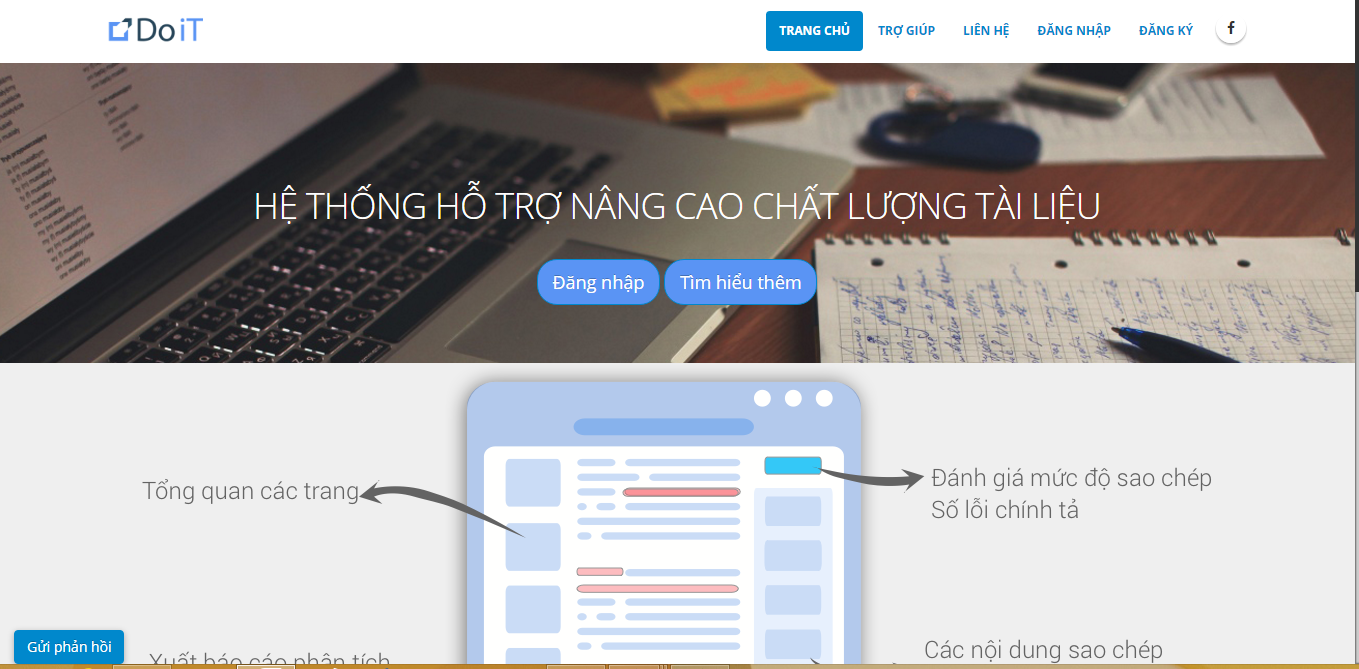
Nhược điểm:

* Chỉ kiểm tra được văn bản bằng tiếng anh, không hỗ trợ tiếng việt.

1. DOIT- Hệ thống hỗ trợ nâng cao chất lượng tài liệu– đại học quốc gia Hà Nội :

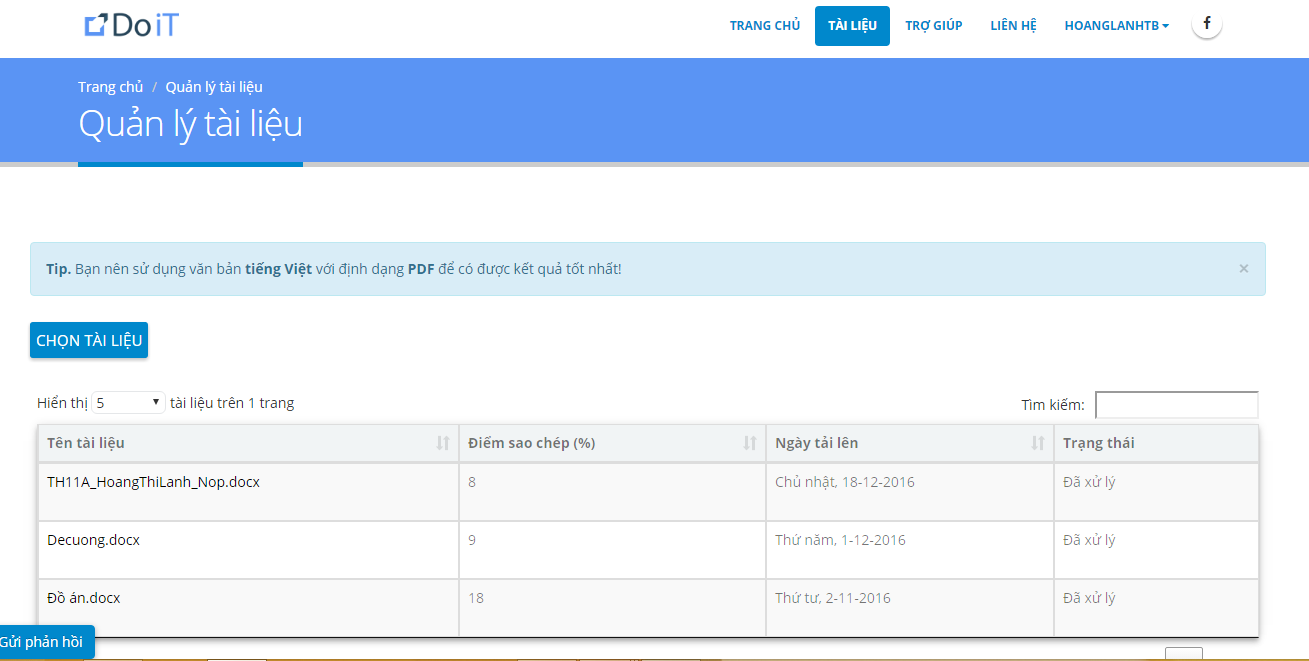
Là hệ thống cho phép kiểm tra lỗi chính tả và phát hiện sao chép.

Khi truy cập đến, giao diện của trang chủ như sau:



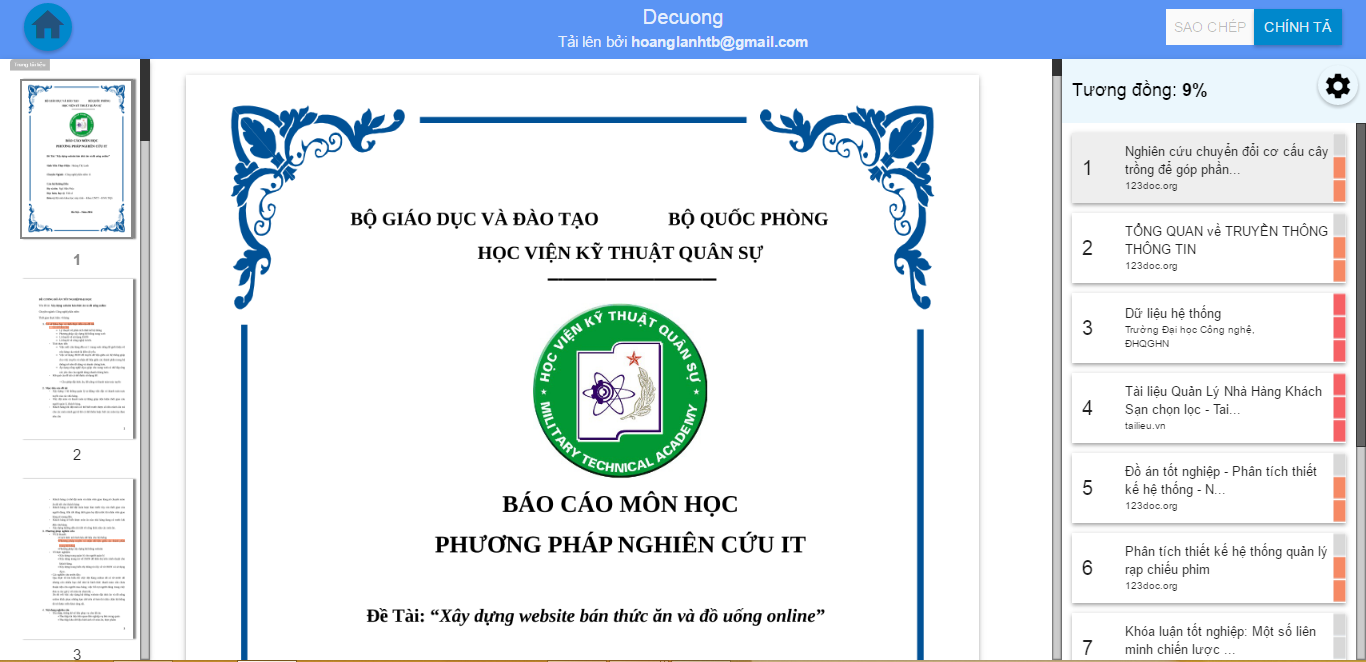
Hình 2.2: Trang chủ của hệ thống

Sau khi đăng nhập vào hệ thống bằng tài khoản đã được đăng kí:

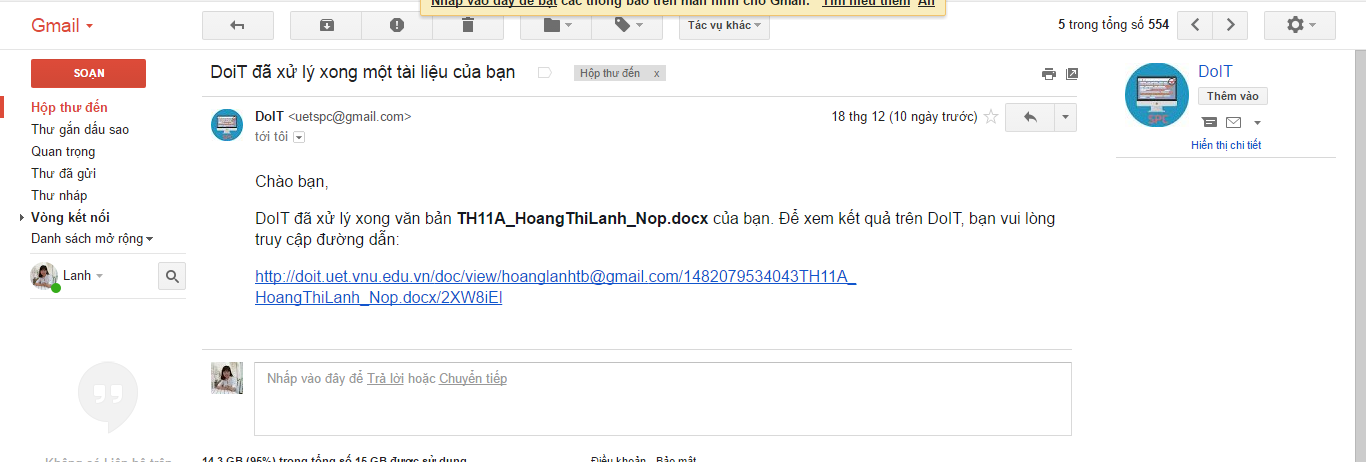


Hình 2.3 : Màn hình sau khi đăng nhập hệ thống

Sau khi xử li xong, hệ thống trả về kết quả tại trang chủ đồng thời gửi kết quả đến mail của người sử dụng.



Hình 2.4: Màn hình kết quả sau khi hệ thống kiểm tra



Hình 2.5: Màn hình Gmail thông báo kết quả sau khi hệ thống kiểm tra

Hệ thống này có các ưu nhược điểm như sau:

Ưu điểm:

* Hệ thống cho phép kiểm tra lỗi chính tả, ngữ pháp, và định dạng của văn bản tương đối chính xác.
* Đưa ra các gợi ý chỉnh sửa cho từ bị sai chính tả.
* Kiểm tra xem nội dung của văn bản có trùng lặp với nội dung của một tài liệu nào đã được công bố trước đó hay không (phát hiện sao chép) và đưa ra độ tương đồng so với văn bản được so khớp.

Nhược điểm:

* Hệ thống xử lí chậm.
* Chỉ đưa ra được 1 từ gợi ý trong khi từ đó chưa thực sự chính xác.
* Chưa được đưa vào dùng chính thức mang tính bắt buộc đối với sinh viên.
* Chỉ được miễn phí 9 tài liệu dùng thử trong vòng 60 ngày.

1. Hệ thống đang được sử dụng tại Học viện Kĩ Thuật Quân Sự

Tại học viện kĩ thuật quân sự hiện nay mới có hệ thống phát hiện sao chép mà chưa có hệ thống kiểm tra lỗi chính tả và định dạng thể thức văn bản. Chính vì vậy mà đang rất cần 1 hệ thống như vậy.

1. Đối tượng sử dụng của hệ thống kiểm tra lỗi chính tả và định dạng văn bản.

* Sinh viên:

+ Có thể đăng kí thành viên

+ Đăng nhập vào hệ thống.

+ Kiểm tra lỗi chính tả.

+ Kiểm tra định dạng văn bản.

* Bộ phận quản lí sinh viên:

+ Đăng nhập vào hệ thống.

+ Kiểm tra lỗi chính tả.

+ Kiểm tra định dạng văn bản.

* Quản trị viên: Quản trị hệ thống

1. Quy trình xử lí của hệ thống kiểm tra lỗi chính tả và định dạng văn bản.

Mỗi sinh viên trong học viện sẽ được cung cấp 1 tài khoản. Với tài khoản này, sinh viên có thể đăng nhập vào hệ thống, có thể tự kiểm tra luận văn của mình trước khi nộp cho bộ phận quản lí. Báo cáo về kết quả kiểm tra lỗi chính tả và đinh dạng văn bản sẽ được hệ thống gửi lại qua mail cho sinh viên.

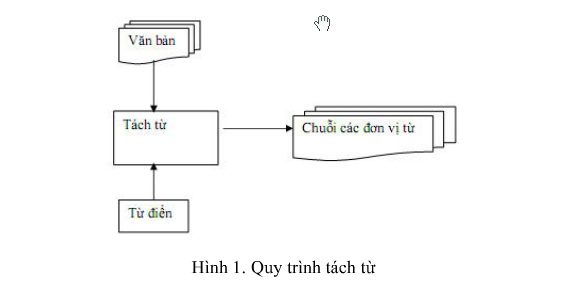
Đến kì phải thu nộp đồ án, bộ phận quản lí sẽ thu bản mềm các bài luận văn của sinh viên. Sau đó sẽ sử dụng hệ thống để kiểm tra xem sinh viên đó đã làm đạt yêu cầu hay chưa. Hệ thống sẽ gửi báo cáo về kết quả kiểm tra lỗi chính tả và định dạng cho bộ phận quản lí. Bộ phận quản lí sẽ gửi kết quả cho sinh viên để sinh viên chỉnh sửa.

1. Tìm hiểu về công cụ tách từ tiếng việt VNTOKENIZER

vnTokenizer là công cụ tách từ tiếng việt được nhóm tác giả Nguyễn Thị Minh Huyền, Vũ Xuân Lương và Lê Hồng Phương phát triển dựa trên phương pháp so khớp tối đa với tập dữ liệu sử dụng là bảng âm tiết tiếng Việt từ điển từ vựng tiếng Việt.

Công cụ được xây dựng bằng ngôn ngữ java, mã nguồn mở. Có thể dễ dàng nâng cấp sửa đổi và tích hợp và các hệ thống phân tích băn bản tiếng Việt khác.

Quy trình thực hiện tách từ theo phương pháp so khớp tối đa:



Đầu vào của công cụ tách từ vnTokenizer là 1 câu hỏi hoặc 1 văn bản lưu dưới dạng tệp.

Đầu ra là 1 chuỗi các từ trong từ điển cũng như các chuỗi số, chuỗi kí tự từ nước ngoài, các hình vị rang buộc (gồm các phụ tố), cac dấu câu và các chuỗi kí tự hỗn tạp khác trong văn bản(ISO, 2008). Các đơn vị từ không chỉ bao gồm các từ có trong từ điển mà cả các từ mới hoặc các từ được phát sinh do theo 1 quy tắc nào đó(như phương thức thêm phụ tối hay phương thức láy) hoặc các chuỗi kí hiệu không được liệt kê trong từ điển.

Công cụ sử dụng tập dữ liệu đi kèm là tập từ điển từ vựng tiêng Việt, danh sách các đơn vị từ mới bổ sung, được biểu diễn bằng otomat tối thiểu hữu hạn trạng thái, tệp chứa các biểu thức chính quy cho phép lọc các đơn vị từ đặc biệt (xâu định dạng số, ngày tháng…), và các tệp chứa các thống kê unigram và bigram trên kho văn bản tách từ mẫu.

Với các đơn vị từ đã có trong từ điển, khi thực hiện tách từ cũng đc xử lí hiện tượng nhập nhằng bằng cách kết hợp với các thống kê unigram và bigram. Chẳng hạn trong tiếng Việt hay gặp những từ nhập nhằng như:

* + Xâu AB vừa có thể hiểu là 1 đơn vị từ, vừa có thể là chuỗi 2 đơn vị từ A-B.
  + Xâu ABC có thể tách thành 2 đơn vị AB-C hoặc A-BC.

1. Tìm hiểu về thư viện docx

Docx là là 1 thư viện .Net cho phép các nhà phát triển thao tác trên các file word 2007/2010/2013, 1 cách dễ dàng trực quan. Docx được viết bởi 1 nghiên cứu sinh -- Cathal Coffey – đại học quốc gia Ireland Maynooth. Docx thì nhanh, nhẹ và tốt mà không yêu cầu Word hay Office nào được cài đặt.

Docx có các tính năng:

* Thêm , loại bỏ, thay thế text trong 1 tài liệu. tất cả các tiêu chuẩn định dạng văn bản có sawnc. Font {Family, Size, color}, Bold, Italic, Underline, Strikethrough, Script{Sub, Super}, Hightlight.
* Thuộc tính của các đoạn văn:

+ Hướng LefttoRight hay RightToLeft

+ Indentation.

+ Alignment

* DocX cũng hỗ trợ

+ Hình ảnh,

+ Các siêu liên kết,

+ Bàn,

+ Đầu trang & chân trang,

+ Thuộc tính tùy chỉnh.

1. Kết luận

Bằng những nhu cầu thực tế và việc nghiên cứu các vấn đề liên quan đến kiểm tra lỗi chính tả và định dạng văn bản thì việc xây dựng 1 hệ thống đáp ứng được các nhu cầu, phát huy các tính năng và xử lí các hạn chế của các hệ thống tương tự là 1 điều cần thiết. Hệ thống này phải đảm bảo:

* Cho phép kiểm tra lỗi chính tả, đưa ra các gợi ý chỉnh sửa 1 cách chính xác.
* Cho phép kiểm tra định dạng.
* Gửi báo cáo sau khi kiểm tra về cho người dùng
* Hệ thống xử lí kết quả 1 cách nhanh chóng, tiết kiệm thời gian của người dùng.
* Giao diện dễ dàng sử dụng, thân thiện với người dùng.

# **Chương 2: Phân tích hệ thống**

## **2.1. Phân tích chức năng nghiệp vụ**

### 2.1.1. Sơ đồ phân rã chức năng

### 2.1.2. Sơ đồ luồng dữ liệu

### 2.1.3. Đặc tả chức năng chi tiết

## **2.2. Phân tích dữ liệu nghiệp vụ**

### 2.2.1. ER mở rộng

### 2.2.2. Chuẩn hóa dữ liệu

### 2.2.3. Đặc tả bảng dữ liệu trong mô hình quan hệ

# **Chương 3: Thiết kế hệ thống**

## **3.1. Thiết kế tổng thể**

### 3.1.1. Phân định công việc giữa người và máy

### 3.1.2. Thiết kế tiến trình hệ thống

## **3.2. Thiết kế kiểm soát**

### 3.2.1. Xác định nhóm người dùng

### 3.2.2. Phân định quyền hạn nhóm người dùng

## **3.3. Thiết kế cơ sở dữ liệu vật lý**

### 3.3.1. Xác định bảng dữ liệu phục vụ bảo mật

### 3.3.2. Mô hình dữ liệu hệ thống

### 3.3.3. Đặc tả bảng dữ liệu

## **3.4. Thiết kế kiến trúc chương trình**

### 3.4.1. Kiến trúc chương trình mức cao

### 3.4.2. Thiết kế module xử lý

## **3.5. Thiết kế giao diện người máy**

### 3.5.1. Thiết kế hệ thống đơn chọn

### 3.5.2. Thiết kế form nhập liệu

### 3.5.3. Thiết kế báo cáo

### 3.5.4. Thiết kế giao diện hỏi đáp

# **Chương 4: Kết quả thử nghiệm và đánh giá**

## 4.1. Kết quả thử nghiệm

## 4.2. Đánh giá

### Kết luận và hướng phát triển

### Một số kết luận

### Hướng phát triển

**Tài liệu tham khảo**

[1] Nguyễn Văn Ba, Phân tích và thiết kế hệ thống thông tin quản lý, NXB-ĐHQG.