**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**LÊ TRẦN LÂM AN – DIỆP HẢI BÌNH**

**NGUYỄN THANH BÌNH – LÊ ĐỨC THUẬN**

**NGUYỄN HOÀNG TUẤN CƯỜNG**

**XÂY DỰNG HỆ THỐNG PHÁT HIỆN VĂN BẢN TƯƠNG ĐỒNG**

**TRONG KHO NGỮ LIỆU CỤC BỘ VÀ INTERNET**

**KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT**

**CHƯƠNG TRÌNH CHÍNH QUY**

**Tp. Hồ Chí Minh, tháng 07/2021**

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

**KHOA CÔNG NGHỆ THÔNG TIN**

**LÊ TRẦN LÂM AN – 1612008**

**DIỆP HẢI BÌNH – 1712291**

**NGUYỄN THANH BÌNH – 1712295**

**NGUYỄN HOÀNG TUẤN CƯỜNG – 1712309**

**LÊ ĐỨC THUẬN - 1712805**

**XÂY DỰNG HỆ THỐNG PHÁT HIỆN VĂN BẢN TƯƠNG ĐỒNG**

**TRONG KHO NGỮ LIỆU CỤC BỘ VÀ INTERNET**

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

CHƯƠNG TRÌNH CHÍNH QUY

**GIÁO VIÊN HƯỚNG DẪN**

|  |  |
| --- | --- |
| TS. | ĐINH ĐIỀN |
| ThS. | LÊ THÀNH NGUYÊN |

Tp. Hồ Chí Minh, tháng 07/2021

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

TpHCM, ngày ….. tháng ..… năm …..

Giáo viên hướng dẫn

[Ký và ghi rõ họ tên]

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

TpHCM, ngày ….. tháng ..… năm …..

Giáo viên hướng dẫn

[Ký và ghi rõ họ tên

# Lời cảm ơn

Để hoàn thành đề tài đồ án tốt nghiệp này, nhóm xin chân thành gửi lời cảm ơn đến thầy … người đã đồng hành và trực tiếp hướng dẫn nhóm trong suốt quá trình thực hiện đồ án. Thấy đã có những góp ý và định hướng để nhóm có thể hoàn thành được đề tài được giao. Một lần nữa xin chân thành cảm ơn thầy.

Em cũng xin gửi lời cảm ơn chân thành đến với các Thầy Cô trường Đại học Khoa học tự nhiên - Đại học Quốc Gia Tp. Hồ Chí Minh đã truyền đạt những kiến thức nền tảng trong suốt thời gian học tập tại trường, không chỉ phục vụ cho đồ án tốt nghiệp mà còn ứng dụng vào công việc thực tế trong tương lai.

Tuy nhiên do kiến thức của nhóm còn hạn chế và thiếu kinh nghiệm nên đồ án của nhóm không tránh khỏi những thiếu sót, nhóm rất mong nhận được những góp ý chân thành từ quý thầy cô để có thể rút kinh nghiệm và tự hoàn thiện bản thân.

Và một lần nữa xin chân thành cảm ơn quý Thầy Cô.

Thành phố Hồ Chí Minh, ngày ….. tháng ….. năm 2021

Nhóm thực hiện

Lê Trần Lâm An

Diệp Hải Bình

Nguyễn Thanh Bình

Nguyễn Hoàng Tuấn Cường

Lê Đức Thuận

# Đề cương chi tiết

|  |
| --- |
| ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP |
| Tên đề tài: Xây dựng hệ thống phát triển văn bản tương đồng trong kho ngữ liệu cục bộ và internet |
| Người hướng dẫn:  PGS.TS. Đinh Điền (Khoa Công Nghệ Thông Tin)  ThS. Lê Thành Nguyên (Khoa Công Nghệ Thông Tin) |
| Nhóm sinh viên thực hiện:  Lê Trần Lâm An (MSSV:1612008)  Diệp Hải Bình (MSSV:1712291 )  Nguyễn Thanh Bình (MSSV:1712295 )  Nguyễn Hoàng Tuấn Cường (MSSV:1712309 )  Lê Đức Thuận (MSSV:1712805 ) |
| Loại đề tài: Ứng dụng |
| Thời gian thực hiện: Từ 12/2020 đến 7/2021 |

## Kế hoạch thực hiện

|  |  |  |  |
| --- | --- | --- | --- |
| **Từ ngày** | **Đến ngày** | **Công việc** | **Thực hiện** |
| 25/12/2020 | 18/01/2021 | Nghiên cứu framework, các ngôn ngữ và công nghệ sử dụng trong đồ án. | Cả nhóm |
| 25/02/2021 | 6/03/2021 | Thiết lập sourcecode, quản lý mã nguồn trên github.  Thêm các thành viên nhóm và các branch cho từng thành viên.  Cài đặt các framework và thử nghiệm. | Cả nhóm |
| 07/03/2021 | 20/04/2021 | Xây dựng hệ thống tiền xử lý, kết nối database.  thiết kế hệ thống, luồng hoạt động.  Dựng giao diện và kết nối api.  Xây dựng tính năng upload nhiều file, tìm kiếm từ khóa trên internet, sửa các lỗi kết nối giữa frontend và backend. | Cả nhóm |
| 21/04/2021 | 11/05/2021 | Xử lý truy xuất dữ liệu từ database, tăng tốc độ truy vấn.  Xử lý các loại định dạng văn bản, nâng cấp khả năng tìm kiếm từ khóa chính xác.  Thử nghiệm tích hợp học sâu. | Cả nhóm |
| 12/05/2021 | 28/05/2021 | Tối ưu hóa các tính năng, sửa lỗi, điều chỉnh giao diện phù hợp.  Chạy thử hệ thống. | Cả nhóm |
| 30/05/2021 | 22/06/2021 | Soạn thảo và hoàn thành cuốn báo cáo.  Xác định nghiệp vụ của hệ thống. | Cả nhóm |

# Mục lục

[**Lời cảm ơn 1**](#_heading=h.gjdgxs)

[**Đề cương chi tiết 2**](#_heading=h.1fob9te)

[Kế hoạch thực hiện 2](#_heading=h.3znysh7)

[**Danh sách các hình 9**](#_heading=h.tyjcwt)

[**Danh sách các bảng 11**](#_heading=h.3dy6vkm)

[**Tóm tắt 12**](#_heading=h.1t3h5sf)

[**Chương 1**](#_heading=h.4d34og8) **Giới thiệu 13**

[1.1](#_heading=h.2s8eyo1) Giới thiệu 13

[1.2](#_heading=h.edr80cp4lrbb) Lý do chọn đề tài 13

[1.3](#_heading=h.249hh7gyijb1) Mục tiêu 14

[1.4](#_heading=h.pz1t0piaotqv) Phạm vi đề tài 14

[1.5](#_heading=h.bx2lljxx360j) Cấu trúc cuốn luận 14

[1.6](#_heading=h.ldjfz8i0rr03) Tổng kết chương 1 15

[**Chương 2**](#_heading=h.17dp8vu) **Tổng quan 16**

[2.1](#_heading=h.3rdcrjn) Khảo sát các ứng dụng tương tự 16

[2.1.1](#_heading=h.26in1rg) Khảo sát hệ thống kiểm tra đạo văn DoIT 16

[2.1.2](#_heading=h.1ksv4uv) Sản phẩm của nhóm 18

[2.1.3](#_heading=h.2jxsxqh) Tương quan giữa 2 sản phẩm 20

[2.2](#_heading=h.z337ya) Khảo sát các nền tảng xây dựng ứng dụng 20

[2.2.1](#_heading=h.3j2qqm3) Server 20

[2.2.2](#_heading=h.1y810tw) Database 21

[2.2.3](#_heading=h.4i7ojhp) Front-end 21

[2.3](#_heading=h.2xcytpi) Tổng kết chương 2 21

[**Chương 3**](#_heading=h.1ci93xb) **Phân tích và thiết kế hệ thống 22**

[3.1](#_heading=h.3whwml4) Kiến trúc hệ thống Plagiarize 22

[3.2](#_heading=h.qsh70q) Mô hình dữ liệu Database 23

[3.3](#_heading=h.ihv636) Sơ đồ usecase của hệ thống Plagiarize 27

[3.3.1](#_heading=h.32hioqz) Tổng quan hệ thống 27

[3.3.2](#_heading=h.41mghml) Hệ thống quản lý tài khoản 27

[3.3.3](#_heading=h.vx1227) Kiểm tra tương đồng 29

[3.3.4](#_heading=h.1v1yuxt) Xem nội dung report 29

[3.3.5](#_heading=h.2u6wntf) Nhóm chức năng dành cho admin 30

[3.4](#_heading=h.3tbugp1) Prototype 31

[3.5](#_heading=h.46r0co2) Tổng kết chương 3 33

[**Chương 4**](#_heading=h.2lwamvv) **Phương pháp sử dụng 34**

[4.1](#_heading=h.9ys8t7f4kult) Tiền xử lý 34

[4.1.1](#_heading=h.7enb86czd655) Rút trích văn bản 35

[4.1.2](#_heading=h.8uk4tu5ffdrw) Tách câu, tách từ 35

[4.1.3](#_heading=h.aq0011smtdpf) Rút trích từ khóa 35

[4.2](#_heading=h.22km3985mcri) Tìm kiếm internet thông qua từ khóa 38

[4.3](#_heading=h.qx7jblit9df6) Kiểm tra tương đồng 38

[4.3.1](#_heading=h.111kx3o) Thuật toán tính khoảng cách Levenshtein 38

[4.3.2](#_heading=h.2zbgiuw) Mô hình học sâu Siamese 41

[4.4](#_heading=h.1egqt2p) Tổng kết chương 4 42

[**Chương 5**](#_heading=h.3ygebqi) **Giới thiệu các công nghệ sử dụng 43**

[5.1](#_heading=h.6x0qisieav2) Database và cơ sở dữ liệu MySQL 43

[5.1.1](#_heading=h.7x3yer13hw54) Giới thiệu MySQL 43

[5.1.2](#_heading=h.sqyw64) Ưu điểm của MySQL 44

[5.1.3](#_heading=h.3cqmetx) Nhược điểm MySQL 45

[5.2](#_heading=h.1rvwp1q) Framework xây dựng backend - Django 45

[5.2.1](#_heading=h.4bvk7pj) Giới thiệu về Django 45

[5.2.2](#_heading=h.2r0uhxc) Tại sao lại chọn Django? 45

[5.2.3](#_heading=h.1664s55) Lợi ích của Django 46

[5.3](#_heading=h.25b2l0r) Ngôn ngữ lập trình Python 46

[5.3.1](#_heading=h.1288jvv2f08s) Giới thiệu về Python 46

[5.3.2](#_heading=h.rhrrn8635ypa) Lợi ích của Python 47

[5.4](#_heading=h.nx0cxw2358mn) Framework xây dựng frontend - Angular 47

[5.4.1](#_heading=h.edtdx3ken0rq) Giới thiệu về angular 47

[5.4.2](#_heading=h.hc98vxd73r9m) Lợi ích của angular 47

[5.5](#_heading=h.kgcv8k) Xác thực bằng json web token - JWT 48

[5.5.1](#_heading=h.34g0dwd) Giới thiệu về JWT 48

[5.5.2](#_heading=h.1jlao46) Giải thích về JWT 49

[5.5.3](#_heading=h.2iq8gzs) Trường hợp cần sử dụng: 49

[5.5.4](#_heading=h.xvir7l) Cơ chế hoạt động: 50

[5.5.5](#_heading=h.1x0gk37) Lợi ích của JWT 50

[5.6](#_heading=h.cqcd9ox04jy4) Tổng kết chương 5 51

[**Chương 6**](#_heading=h.4h042r0) **Triển khai và đánh giá sản phẩm 52**

[6.1](#_heading=h.hsbkikho7yzj) Kết quả đạt được 52

[6.2](#_heading=h.2z623dprjwww) Mô tả chức năng 53

[6.2.1](#_heading=h.qmo66iaqxi6w) Chức năng đăng ký và đăng nhập 53

[6.2.2](#_heading=h.3vac5uf) Nhóm chức năng kiểm tra tương đồng văn bản 55

[6.3](#_heading=h.haapch) Nhóm chức năng của admin 59

[6.3.1](#_heading=h.319y80a) Chức năng khóa tài khoản người dùng 59

[6.3.2](#_heading=h.1gf8i83) Chức năng file lên database của hệ thống. 60

[6.4](#_heading=h.2fk6b3p) Tổng kết chương 6 61

[**Chương 7**](#_heading=h.upglbi) **Tổng kết và hướng phát triển 62**

[7.1](#_heading=h.gvshrx3xh547) Những thuận lợi, khó khăn gặp phải trong quá trình thực hiện và hướng khắc phục 62

[7.1.1](#_heading=h.qbmhvzesv1jb) Thuận lợi 62

[7.1.2](#_heading=h.1u7ugdf7sc8x) Khó khăn 62

[7.1.3](#_heading=h.fmryl87392ur) Hướng khắc phục 63

[7.2](#_heading=h.usufn2u8wq8z) Hạn chế 63

[7.3](#_heading=h.70v6ax4hpl0h) Hướng phát triển 63

[7.4](#_heading=h.8d18ueett42w) Quản lý, phân chia công việc 64

[7.4.1](#_heading=h.5wv3aqy3snt1) Phương pháp quản lý công việc 64

[7.5](#_heading=h.3ep43zb) Tổng kết chương 7 64

[**Danh mục công trình của tác giả 65**](#_heading=h.1tuee74)

[**Tài liệu tham khảo 65**](#_heading=h.2szc72q)

[**Phụ lục 65**](#_heading=h.4du1wux)

# Danh sách các hình

[Hình 2.1 Chức ngăng kiểm tra tương đồng 17](#_heading=h.lnxbz9)

[Hình 2.2 Chức năng kiểm tra chính tả 18](#_heading=h.35nkun2)

[Hình 2.3 Giao diện đăng ký tài khoản của hệ thống 19](#_heading=h.44sinio)

[Hình 3.1 Sơ đồ kiến trúc hệ thống của Plagiarize 22](#_heading=h.2bn6wsx)

[Hình 3.2 Class diagram cho chức năng kiểm tra tương đồng của văn bản 26](#_heading=h.3o7alnk)

[Hình 3.3 Class diagram cho chức năng xem report 26](#_heading=h.23ckvvd)

[Hình 3.4 Use case diagram cho hệ thống xác thực và các tính năng chính của ứng dụng Plagiarize 27](#_heading=h.1hmsyys)

[Hình 3.5 Use case dìagram cho quy trình đăng nhập, xác thực, phân quyền 28](#_heading=h.2grqrue)

[Hình 3.6 Use case diagram cho chức năng kiểm tra tương đồng của văn bản 29](#_heading=h.3fwokq0)

[Hình 3.7 Use case diagram cho chức năng xem nội dung report của phiên kiểm tra 30](#_heading=h.4f1mdlm)

[Hình 3.8 Use case diagram cho chức năng của admin 31](#_heading=h.19c6y18)

[Hình 3.9 Màn hình chính của người dùng 31](#_heading=h.28h4qwu)

[Hình 3.10 Màn hình chính của admin 32](#_heading=h.nmf14n)

[Hình 3.11 Màn hình so sánh kết quả 32](#_heading=h.37m2jsg)

[Hình 3.12 Màn hình tải tệp lên để so sánh tương đồng 33](#_heading=h.1mrcu09)

[Hình 4.1 Ma trận tách các từ thành các kí tự 39](#_heading=h.3l18frh)

[Hình 4.2 Ma trận giá trị D(i,j) 40](#_heading=h.206ipza)

[Hình 4.3 Ma trận kết qủa 40](#_heading=h.4k668n3)

[Hình 5.1 Mô tả mối quan hệ của ORM 46](#_heading=h.3q5sasy)

[Hình 5.2 Nguyêm tắc mã hoá JWT 49](#_heading=h.43ky6rz)

[Hình 5.3 Cấu trúc JWT 50](#_heading=h.3hv69ve)

[Hình 6.1 Giao diện đăng ký tài khoản mới](about:blank) 54

[Hình 6.2 Giao diện nhập mã xác thực](about:blank) 54

[Hình 6.3 Giao diện đăng nhập 55](#_heading=h.1baon6m)

[Hình 6.4 Giao diện chính của hệ thống 56](#_heading=h.2afmg28)

[Hình 6.5 Các tuỳ chọn của người dùng 56](#_heading=h.pkwqa1)

[Hình 6.6 Người dùng tải file cần kiểm tra lên hệ thống 57](#_heading=h.39kk8xu)

[Hình 6.7Ngườ dùng up các file làm database để so sánh 57](#_heading=h.48pi1tg)

[Hình 6.8 Tuỳ chọn cho các chức năng kiểm tra 58](#_heading=h.1302m92)

[Hình 6.9 Màn hình kết quả 59](#_heading=h.2250f4o)

[Hình 6.10 Giao diện quản lý tài khoản của admin](about:blank) 60

[Hình 6.11 Giao diện cập nhật database của admin 60](#_heading=h.40ew0vw)

# Danh sách các bảng

[Bảng 3.1 Database User 23](#_heading=h.3as4poj)

[Bảng 3.2 B̫ảng database DataDocument 24](#_heading=h.1pxezwc)

[Bảng 3.3 B̫ảng database DataDocumentContent 24](#_heading=h.49x2ik5)

[Bảng 3.4 B̫ảng database DocumentSession 25](#_heading=h.2p2csry)

[Bảng 3.5 B̫ảng database ReportDocument 25](#_heading=h.147n2zr)

[Bảng 5.1 Kiểu dữ liệu và kích thước 44](#_heading=h.2dlolyb)

[Bảng 6.1 Danh sách chức năng 53](#_heading=h.2w5ecyt)

# Tóm tắt

Đồ án được thực hiện nhằm mục tiêu tạo ra một trang web giúp tính toán độ tương đồng giữa các văn bản với nhau. Các chức năng kiểm tra tương đồng chính bao gồm: so sánh với dữ liệu của hệ thống, của người dùng tải lên, với dữ liệu internet. Từ đó cho người dùng biết được độ tương đồng giữa các văn bản, giúp cho người dùng kiểm tra được một văn bản có bị đạo hay không. Tùy vào loại tài khoản mà người dùng có các chức năng khác nhau.

Nếu là tài khoản dành cho quản trị viên sẽ có được gần như là toàn quyền sử dụng các tính năng như: cấp quyền sử dụng cho người dùng thông thường; cập nhật, thêm hoặc xóa dữ liệu hệ thống; được xem toàn bộ thông tin của người dùng.

Nếu là tài khoản dành cho người dùng thông thường sẽ chỉ được sử dụng duy nhất tính năng kiểm tra tương đồng, tùy thuộc vào quản trị viên sẽ có thể giới hạn số lượt kiểm tra.

Những dữ liệu đã tải lên sẽ được hiển thị đầy đủ thông tin tại trang chủ của người dùng bao gồm: tên phiên. ngày tải lên, tổng số tệp, trạng thái, loại so sánh.

# Giới thiệu

*Chương mở đầu nhóm sẽ giới thiệu khái quát về hoàn cảnh của đề tài, nêu lý do lựa chọn đề tài, đề tài này giải quyết vấn đề gì mang tính thực tiễn, mô tả phạm vi của đề tài và mục tiêu của nhóm đặt ra khi thực hiện đề tài này. Bên cạnh đó trình bày cấu trúc của cuốn luận.*

## Giới thiệu

Trong thời đại công nghệ phát triển như hiện nay, khi mà các nội dung ghi chép đang được dần chuyển sang dạng lưu trữ thông tin điện tử, đem đến nhiều ưu điểm giải quyết các vấn đề của việc sao lưu vật lý. Thì bên cạnh đó, nó cũng dấy lên những thách thức mới cho tầng lớp giảng dạy và học tập về vấn đề trình bày sự hiểu biết, tư duy phản biện dưới dạng văn bản. Việc tìm kiếm và sao chép nội dung tài liệu được viết và xây dựng từ các nguồn lưu trữ văn bản trên mạng đang là vấn đề rất được quan tâm đối với ngành giáo dục ở nhiều các quốc gia trên thế giới trong đó có cả Việt Nam. Đây là một vấn đề nan giải đối với ngành giáo dục cần phải có những cách giải quyết chuẩn mực.

Xuất phát chính của vấn đề này chủ yếu là do sự thiếu siêng năng trong việc tìm hiểu và nghiên cứu nội dung của đối tượng được tiếp cận. Từ đó tiến hành sao chép các văn bản được viết bởi người khác thay vì trình bày nội dung đó theo sự hiểu biết của bản thân. Đây là hành vi đã và đang tồn tại trong hầu hết mọi trường học ở mọi quốc gia trên thế giới. Để hạn chế và phát hiện một cách thủ công là một việc không hề dễ dàng, rất mất thời gian, công sức nên việc phát triển một hệ thống giúp phát hiện ra các hành vi này là thật sự cần thiết trong bối cảnh mà máy tính có thể thay cho sức người.

Vì thế, xây dựng hệ thống phát hiện văn bản tương đồng trên dữ liệu cục bộ và internet được chọn làm hướng nghiên cứu chính. Báo cáo đồ án sẽ giới thiệu cũng như trình bày về hệ thống mà nhóm sinh viên đã xây dựng được.

## Lý do chọn đề tài

Vì các hệ thống phát hiện đạo văn dành cho tiếng Việt chưa được phổ biến tại Việt Nam nên nhóm có mong muốn đóng góp cho người Việt, doanh nghiệp, trường học thêm một hệ thống mới cũng như có nhiều sự lựa chọn hơn. Ngoài chức năng chính là phát hiện tương đồng thì hệ thống còn có khả năng cho phép người dùng tự so sánh với chính tài liệu riêng của họ. Ngoài ra, nhóm cũng đã có những kỹ thuật xử lý, cải tiến nhằm nâng cao hiệu năng, tăng độ chính xác và giảm thời gian xử lý của hệ thống nhiều nhất có thể.

## Mục tiêu

Mục tiêu cuối cùng mà nhóm hướng tới vẫn là phát hiện những hành vi đạo văn và thậm chí là tìm kiếm những tài liệu gốc bị sao chép mà không có sự cho phép của tác giả.

## Phạm vi đề tài

Các chức năng chính của hệ thống bao gồm:

* So sánh với dữ liệu cục bộ
* So sánh với dữ liệu internet
* So sánh với dữ liệu cục bộ và internet
* So sánh với dữ liệu do người dùng tải lên

## Cấu trúc cuốn luận

Đây là phần tóm tắt nội dung cuốn luận được trình bày sau đây:

* Chương 1: Giới thiệu
  + Trình bày sơ lược hoàn cảnh thực tế, lý do chọn lựa đề tài, mục tiêu cần đạt và phạm vi của đồ án.
* Chương 2: Tổng quan
  + Ở chương này, sẽ tiến hành khảo sát các ứng dụng có cùng đề tài hiện có, bên cạnh đó giới thiệu nền tảng sử dụng cho đồ án, trình bày sơ lược lí do vì sao chọn những nền tảng này, đến các chương kế tiếp sẽ trình bày chi tiết hơn.
* Chương 3: Phân tích và thiết kế hệ thống
  + Chương 3 sẽ trình bày chi tiết về các sơ đồ kiến trúc hệ thống, các sơ đồ use case và bản prototype mẫu được phác thảo trước khi đi vào xây dựng hệ thống.
* Chương 4: Phương pháp sử dụng
  + Trong chương 4 sẽ trình bày cách tiếp cận bài toán đặt ra, các cách giải quyết chức năng chính của đề tài, đi vào chi tiết nội dụng cũng như các bước xử lý văn bản để cho ra kết quả tương đồng.
* Chương 5: Các công nghệ sử dụng
  + Trình bày chi tiết về các công nghệ đã sử dụng trong đồ án, lý do chọn lựa công nghệ đó.
* Chương 6: Triển khai và đánh giá sản phẩm
  + Trình bày thành quả của đồ án, thống kê và đánh giá các chức năng của đồ án
* Chương 7: Tổng kết và hướng phát triển
  + Tổng kết quá trình thực hiện đồ án của nhóm, những khó khăn trở ngại gặp phải, đề xuất hướng khắc phúc cũng như các hướng cải tiến cho tương lai.

## Tổng kết chương 1

Thông qua chương giới thiệu này, nhóm đã khái quát sơ lược về đề tài bao gồm: các mục tiêu mà đề tài này hướng đến, những lợi ích đề tài mang lại cũng như phạm vi của đề tài. Các kỹ thuật, chức năng chính của đề tài sẽ được nhóm trình bày chi tiết hơn ở các chương tiếp theo.

# Tổng quan

*Trong chương 2 nhóm sẽ giới thiệu tổng quan về việc khảo sát ứng dụng tương tự với nội dung đề tài, nêu lý do lựa chọn các công nghệ, framework để phục vụ cho việc xây dựng hệ thống. Bên cạnh đó trình bày prototype của đồ án mà nhóm thực hiện.*

## Khảo sát các ứng dụng tương tự

### Hệ thống kiểm tra đạo văn DoIT

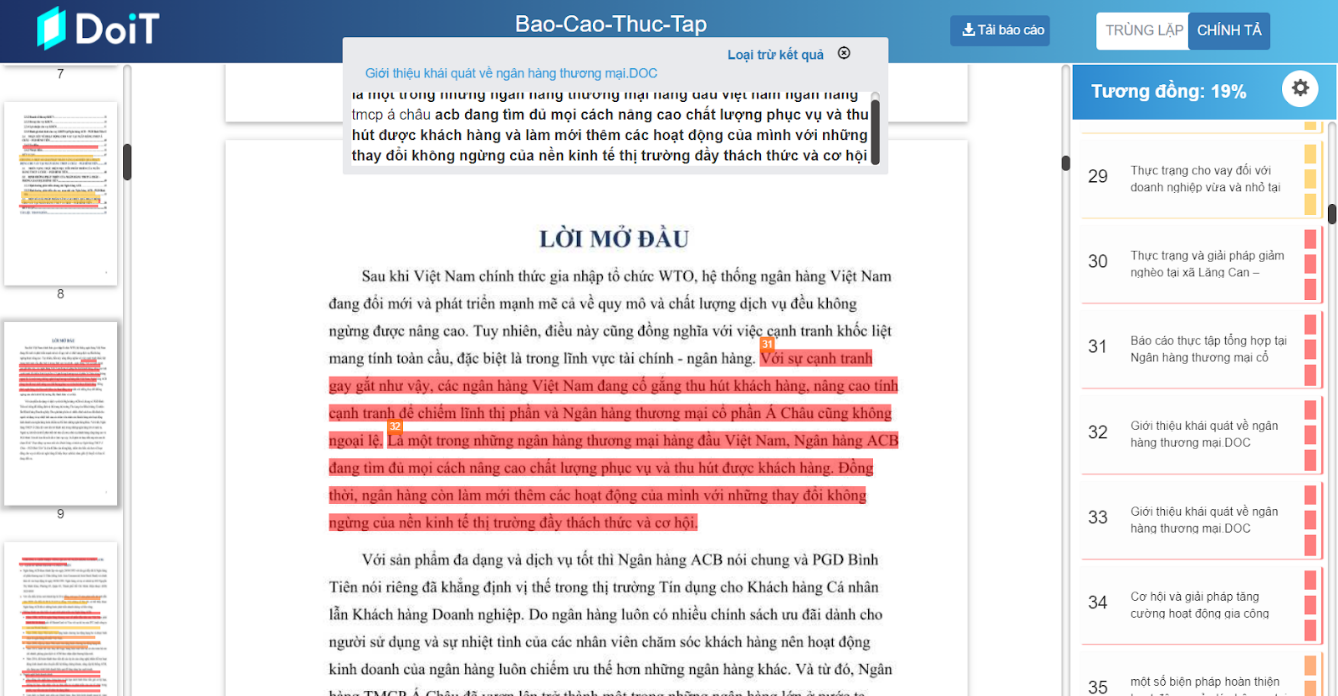
Hệ thống kiểm tra đạo văn DoIT (Document Improvement Tool) là sản phẩm dịch vụ được cung cấp bởi Trường Đại học Công nghệ, dành cho đối tượng cá nhân, các trường, tổ chức giáo dục.

DoIT được phát triển bởi Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội. Bằng việc ứng dụng các nghiên cứu trong lĩnh vực xử lý ngôn ngữ tự nhiên và xử lý dữ liệu lớn, DoIT đã hỗ trợ việc kiểm tra trùng lặp văn bản và sửa lỗi chính tả thông minh, nhanh chóng.

Để sử dụng được, người dùng bắt buộc phải có tài khoản và tài khoản của người dùng thông thường chỉ có thể kiểm tra tối đa 2 lần.

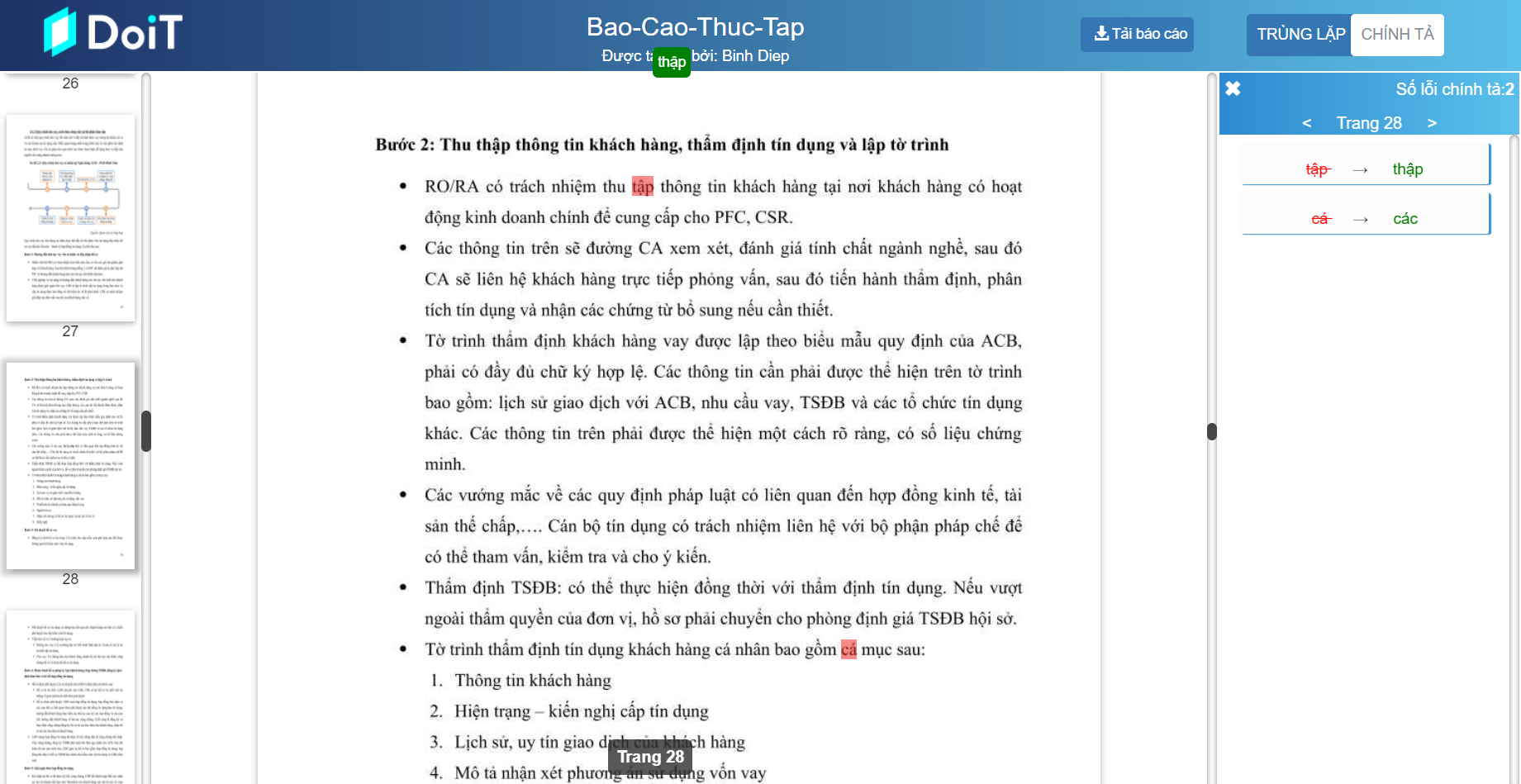
Hệ thống có 2 chức năng chính là:

* Kiểm tra tương đồng
  + Tại mỗi trang, hệ thống sẽ tính toán độ tương đồng của từng trang so với các kết quả mà hệ thống tìm được trên cơ sở dữ liệu của hệ thống lẫn trên internet.
  + Với mỗi vị trí có sự trùng lặp, hệ thống sẽ tô đậm và gắn kèm link mà hệ thống tìm thấy.
  + Sau khi đã xác định được mức độ trùng lặp tại mỗi trang, hệ thống sẽ có những tính toán để xác định độ trùng lặp cho toàn bộ văn bản như trong hình 2.1.



**Hình 2.1 Chức năng kiểm tra tương đồng của DoIT**

* Kiểm tra chính tả
  + Bằng các phương pháp máy học, hệ thống sẽ phát hiện những từ bị sai chính tả, sai định dạng như trong hình 2.2.
  + Tuy nhiên, việc kiểm tra chính tả này sẽ không đạt được độ chính xác tuyệt đối nên một số từ có thể vẫn đúng nhưng hệ thống vẫn xem là lỗi.



**Hình 2.2 Chức năng kiểm tra chính tả của DoIT**

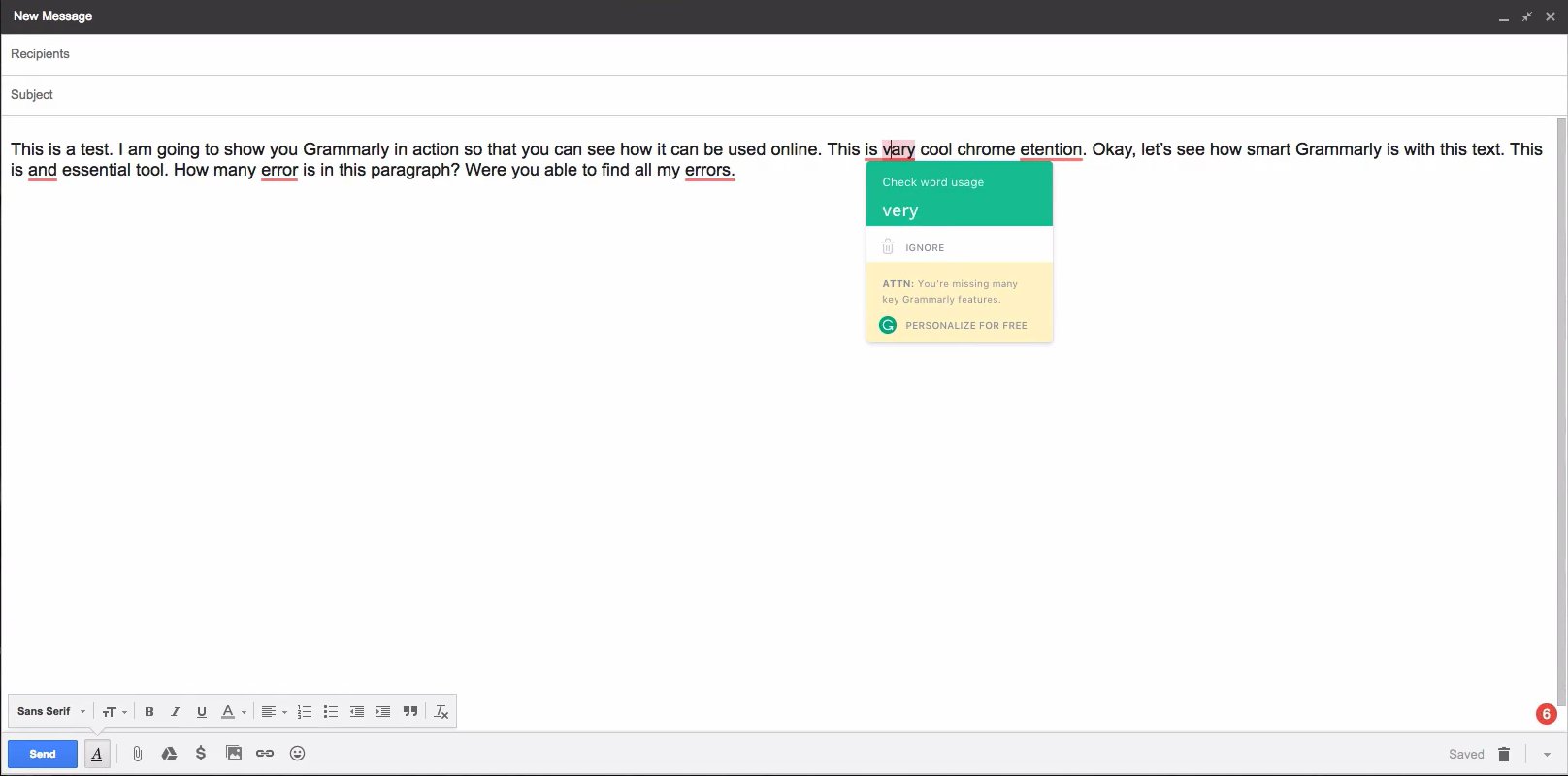
**<http://doit.uet.vnu.edu.vn/>**

**2.1.2. Grammarly**

Grammarly là sản phẩm ứng dụng xử lý ngôn ngữ tự nhiên (NLP) và trí tuệ nhân tạo để phát hiện và chỉnh sửa lỗi ngữ pháp trong tiếng Anh có trụ sở tại San Francisco. Grammarly sẽ đề xuất thay thế các từ sao cho hợp ngữ cảnh cho người dùng xem xét thay đổi. Grammarly cũng hỗ trợ phát hiện tương đồng của một văn bản tiếng Anh.

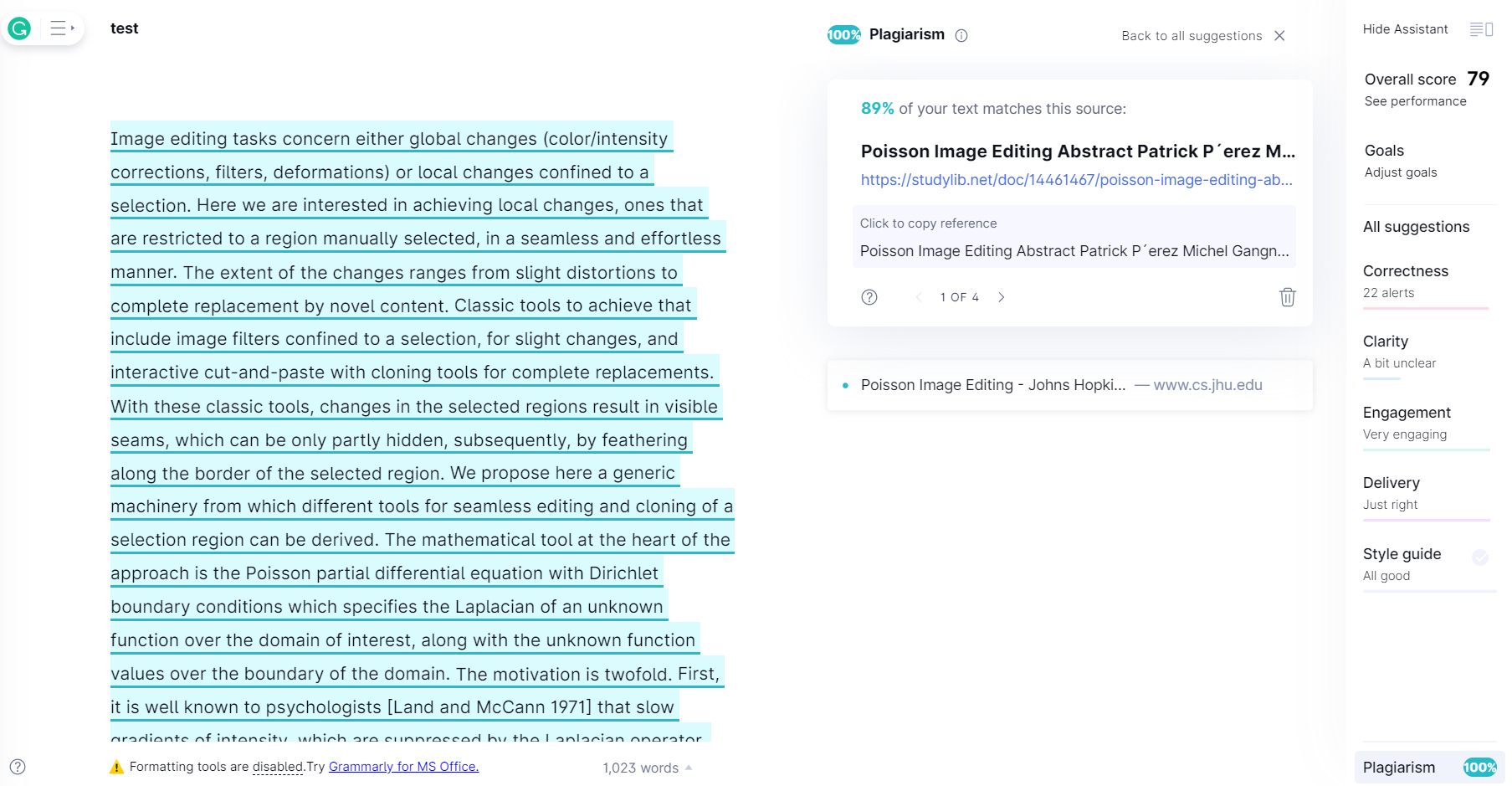
Grammarly hỗ trợ trên nhiều nền tảng như: Browser, Web, Microsoft Office, Mobile application…

* Nếu sử dụng tại Browser, hệ thống không bắt buộc người dùng phải tạo tài khoản vì hệ thống cho sử dụng miễn phí. Tuy nhiên, giới hạn chỉ được sử dụng để kiểm tra chính tả.



**Hình 2.3 Chức năng kiểm tra chính tả của Grammarly**

* Nếu người dùng có tài khoản thì sẽ được sử dụng thêm chức năng kiểm tra tương đồng nhưng chỉ với văn bản tiếng Anh. Hạn chế của hệ thống là chỉ có thể upload được các file: doc, docx, odt, rtf, htm, html, txt.



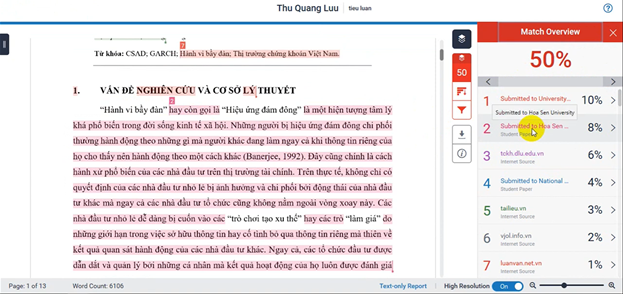
**Hình 2.4 Chức năng kiểm tra tương đồng của Grammarly**

**2.1.3 Turnitin**

Turnitin là một dịch vụ phát hiện đạo văn trên Internet được điều hành bởi công ty Turnitin, có trụ sở chính tại Mỹ.

Tương tự các hệ thống khác, Turnitin cũng yêu cầu có tài khoản mới có thể sử dụng chức năng kiểm tra đạo văn. Tài khoản khi đăng ký sẽ có ba loại:

* Student: tài khoản dành cho học sinh, sinh viên
* Teaching Assistant: tài khoản dành cho việc giảng dạy
* Instructor: tài khoản dành cho giảng viên



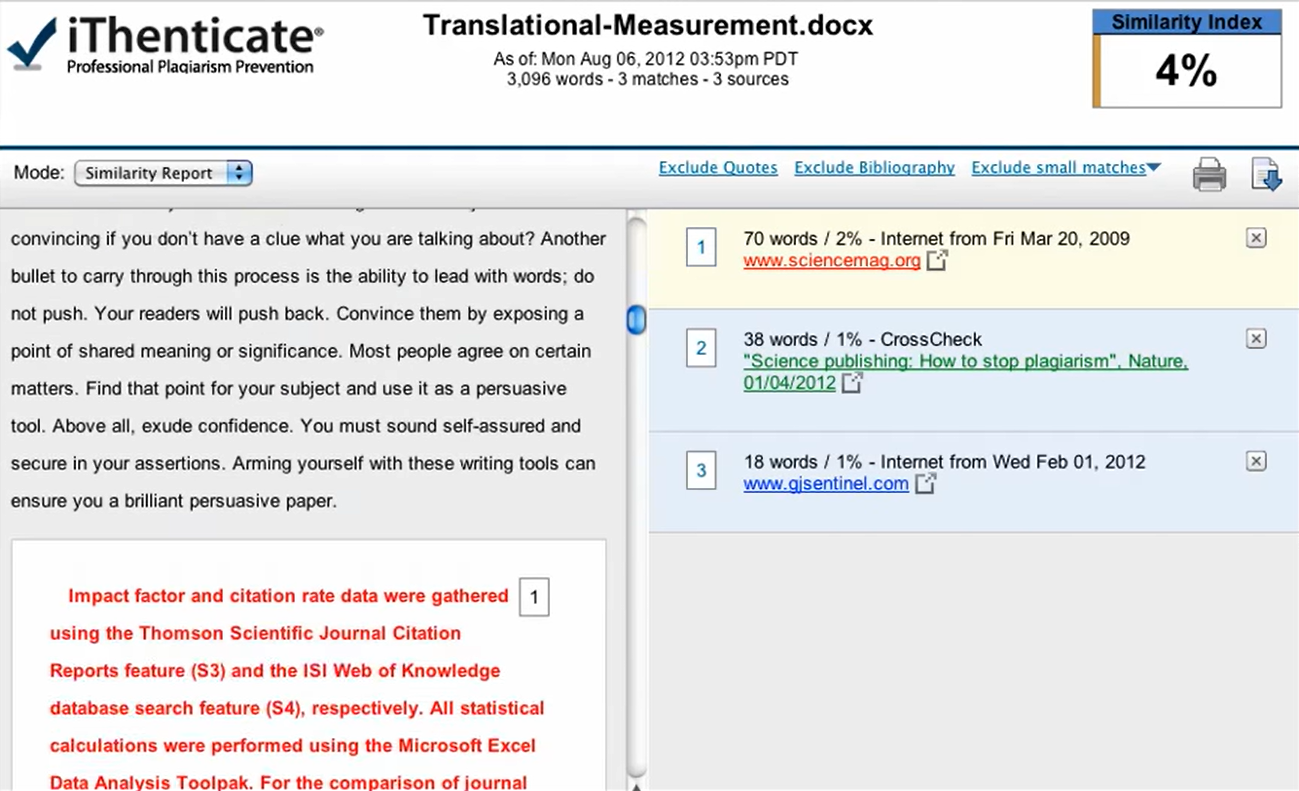
**Hình 2.4 Chức năng kiểm tra tương đồng của Turnitin**

[**https://en.wikipedia.org/wiki/Turnitin**](https://en.wikipedia.org/wiki/Turnitin)

**2.1.4 Ithenticate**

iThenticate cũng là một dịch vụ phát hiện đạo văn dành cho thị trường doanh nghiệp thuộc Turnitin. iThenticate hỗ trợ 30 ngôn ngữ bao gồm các ngôn ngữ phổ biến như: Anh, Trung, Tây Ban Nha, Nga,...

* Với người dùng thông thường, cần trả phí và đăng ký tài khoản mới có thể sử dụng. Người dùng sẽ có 2 gói dịch vụ để lựa chọn: gói 100$/năm và 300$/năm
* Gói 100$/năm: Cho phép kiểm tra 1 lần 1 văn bản giới hạn 25000 từ.
* Gói 300$/năm: Cho phép kiểm tra nhiều văn bản cùng lúc với mỗi văn bản giới hạn 75000 từ.
* Với người dùng là một tổ chức, trường học thì sẽ phải thương lượng riêng với công ty.

****

**Hình 2.5 Chức năng kiểm tra tương đồng của iThenticate**

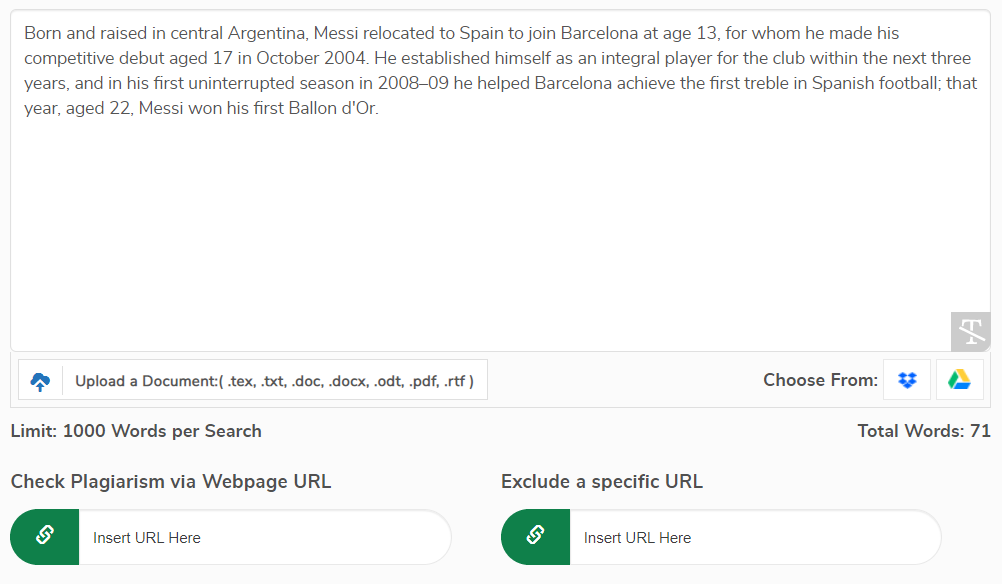
[**https://en.wikipedia.org/wiki/IThenticate**](https://en.wikipedia.org/wiki/IThenticate)

[**https://www.ithenticate.com/**](https://www.ithenticate.com/)

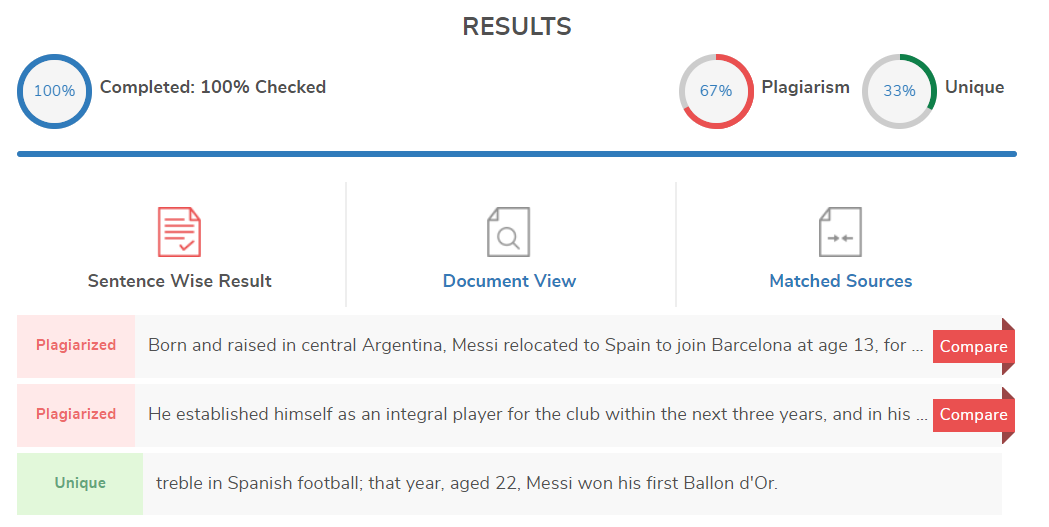
**2.1.5 Small seo tools**

Small seo tools cũng là một website giúp kiểm tra chính tả và kiểm tra tương đồng.

* Ưu điểm:
* Small seo tools hỗ trợ hầu hết các định dạng file thông dụng như: txt, doc, docx, pdf,...
* Ngoài upload các file cá nhân, hệ thống cũng cho phép upload một URL.
* Được sử dụng miễn phí, không cần đăng ký tài khoản.
* Nhược điểm: hệ thống chỉ cho phép giới hạn 1000 từ cho mỗi lần kiểm tra.

****

**Hình 2.6 Giao diện upload nội dung cần kiểm tra của Small seo tools**

****

**Hình 2.7 Giao diện hiển thị kết quả của Small seo tools**

[**https://smallseotools.com/plagiarism-checker/**](https://smallseotools.com/plagiarism-checker/)

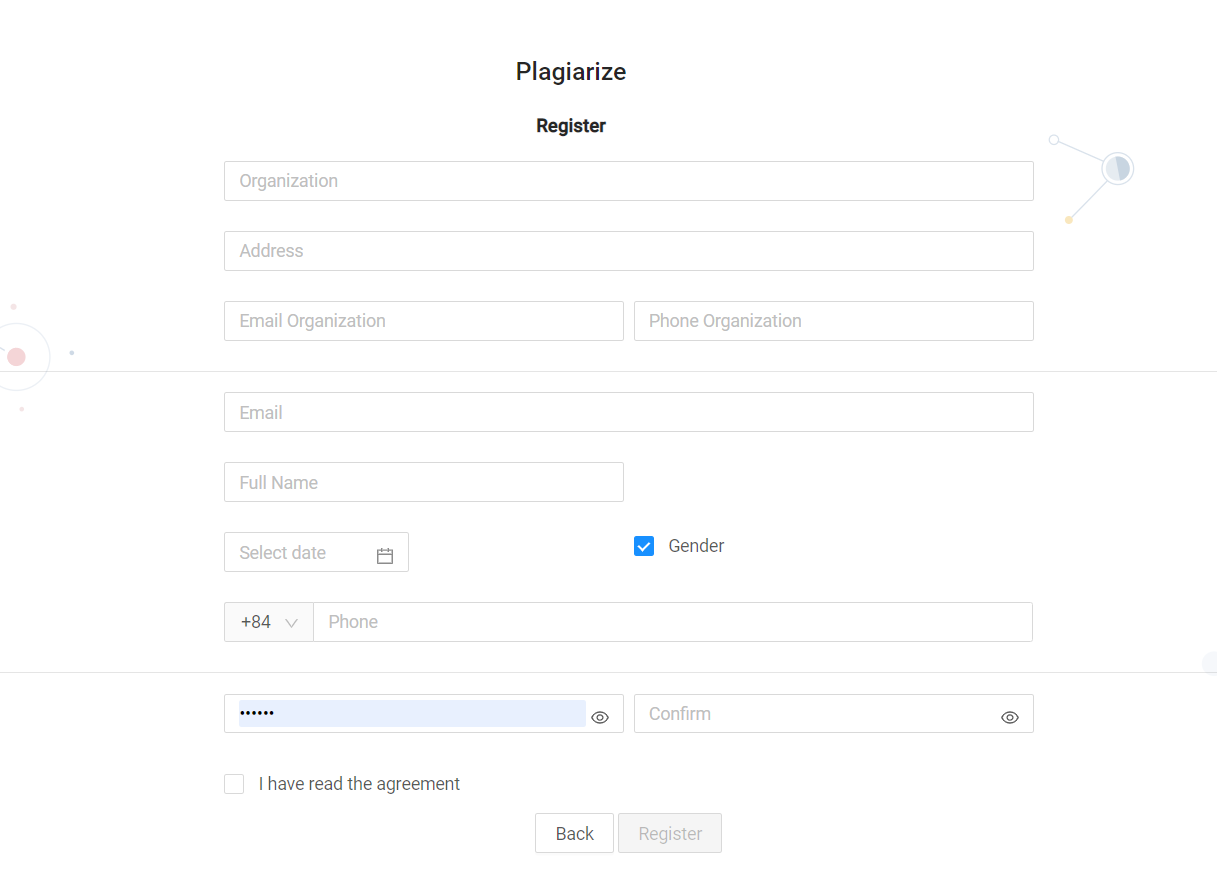
### Plagiarize - Sản phẩm của nhóm

Cũng tương tự các hệ thống trên, để có thể sử dụng hệ thống kiểm tra đạo văn thì bắt buộc phải đăng ký tài khoản.

Có 2 loại tài khoản:

* Tài khoản administrator: là tài khoản dành cho quản trị viên, có toàn quyền quản lý cơ sở dữ liệu của hệ thống cũng như cấp quyền sử dụng cho người dùng. Tài khoản này không thể trực tiếp đăng ký ngay trên hệ thống.
* Tài khoản user: là tài khoản do người dùng đăng ký để được sử dụng các chức năng kiểm tra đạo văn của hệ thống. Loại tài khoản này chỉ được sử dụng những tính năng trong giới hạn mà admin đã cấp quyền.

Sau đây là giao diện đăng ký tài khoản của hệ thống:



**Hình 2.8 Giao diện đăng ký tài khoản của hệ thống**

Những vấn đề chưa tìm ra hướng giải quyết còn tồn tại trong hệ thống này bao gồm:

* Độ chính xác trong quá trình xác định độ tương đồng giữa 2 câu: Vì dữ liệu còn hạn chế, nên mô hình học sâu được sử dụng trong hệ thống chưa được huấn luyện đủ tốt, dẫn đến độ chính xác chưa cao.
* Khai thác dữ liệu từ internet: Khó khăn ở đây chính là các website không thể truy cập tại Việt Nam mà đòi hỏi phải chuyển vùng mới có thể truy cập, dẫn đến bị thiếu sót một số lượng lớn nguồn tham khảo tìm được trên internet

### 

### Tương quan giữa các sản phẩm

**Giống nhau:**

* Ngoài Small seo tools, các hệ thống còn lại đều cần đăng ký tài khoản để sử dụng.
* Đều có chức năng chính là xác định mức độ tương đồng giữa các văn bản với nhau.
* Có tính năng thông báo qua email người dùng về chi tiết kết quả khi đã kiểm tra xong.
* Kiểm tra ngay cả trên cơ sở dữ liệu của hệ thống và trên internet.

**Khác nhau:**

* Sự khác biệt giữa sản phẩm của nhóm và các hệ thống đã khảo sát ở trên chủ yếu ở chức năng cho phép người dùng tùy chọn lưu hoặc không lưu văn bản của mình vào cơ sở dữ liệu của hệ thống.
* Sản phẩm của nhóm chỉ dừng lại ở mức kiểm tra tương đồng và không có chức năng kiểm tra chính tả.
* Ngoài các chi tiết chính, cũng có một số điểm khác nhau sau:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Plagiarize** | **Turnitin** | **Small seo tools** | **Grammarly** | **iThenticate** | **DoIT** |
| **Cập nhật cơ sở dữ liệu** | **✓** |  |  |  |  |  |
| **Sửa lỗi chính tả** |  | **✓** |  | **✓** | **✓** | **✓** |
| **Giới hạn độ dài văn bản** |  |  | **✓** |  | **✓** |  |
| **Tốn phí** |  | **✓** |  | **✓** | **✓** |  |

### 

## Khảo sát các nền tảng xây dựng ứng dụng

### Server

Sau các cuộc họp trao đổi và thảo luận, theo những đề xuất từ giáo viên hướng dẫn, nhóm sinh viên quyết định chọn framework Django. Django là một framework để xây dựng ứng dụng hiệu quả. Nó sử dụng ngôn ngữ bậc cao phổ biến, Python, phân chia ứng dụng dưới dạng các component riêng biệt giúp cho developer dễ dàng phát triển.

### Database

Với sự phổ biến của MySQL và kinh nghiệm làm việc của thành viên nhóm với ngôn ngữ truy vấn này, nhóm sinh viên quyết định chọn đây là cơ sở dữ liệu để sử dụng. Ngoài ra MySQL còn hỗ trợ khả năng lập chỉ mục phục vụ cho việc xử lý truy vấn trong cơ sở dữ liệu, tăng tốc độ truy xuất với dữ liệu lớn.

### Front-end

Với khả năng xử lý nhanh,giảm kích thước ứng dụng, khả năng tạo ra một cấu trúc rõ ràng cộng với lượng tài liệu hướng dẫn phong phú mà nhóm quyết định chọn Angular đảm nhiệm vai trò front end của ứng dụng.

## Tổng kết chương 2

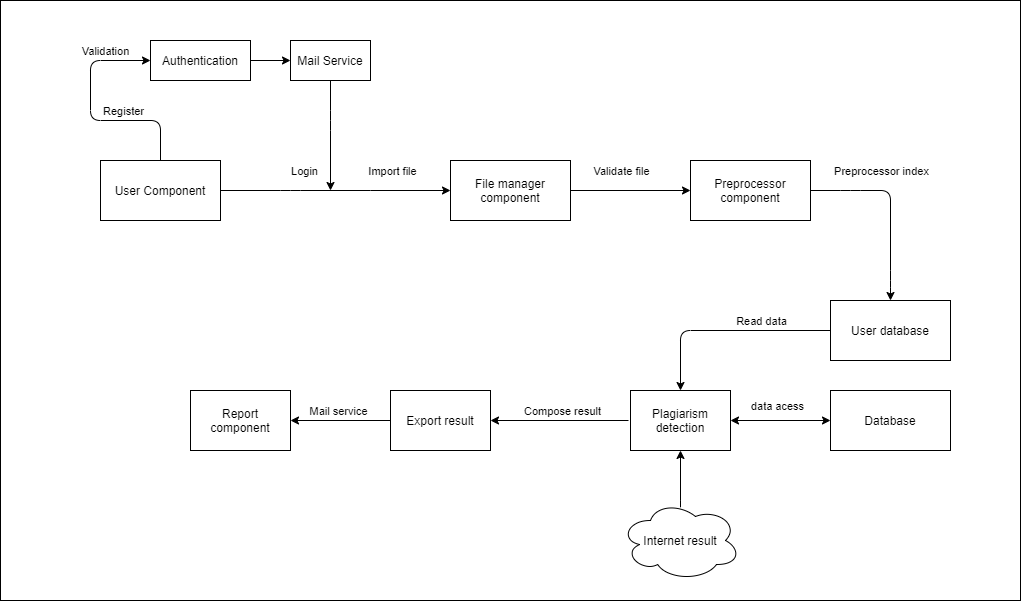
Thông qua chương 2 nhóm đã trình bày một cách tổng quan, một cái nhìn sơ lược về đồ án nhóm sẽ thực hiện, xác định được hướng triển khai đồ án.

# Phân tích và thiết kế hệ thống

*Trong chương 3 nhóm sẽ giới thiệu về cấu trúc của hệ thống được xây dựng, các use case của người dùng, cách thức tổ chức dữ liệu trong cơ sở dữ liệu.*

## Kiến trúc hệ thống Plagiarize

Hình 3.1 dưới đây là mô tả tổng quát hệ thống Plagiarize. Chi tiết từng thành phần của hệ thống sẽ được nhóm trình bày kĩ hơn ở chương 5.



**Hình 3.1 Sơ đồ kiến trúc hệ thống của Plagiarize**

## Mô hình dữ liệu Database

Ở phần này nhóm sẽ trình bày các bảng cơ sở dữ liệu được xây dựng trong đồ án.

|  |  |  |
| --- | --- | --- |
| **Mô Tả** | **Lưu trữ thông tin của User** | |
| **id** | int (11) | Khóa chính |
| **UserName** | varchar (30) |  |
| **Password** | varchar (150) |  |
| **Name** | varchar (30) |  |
| **EmailOrganization** | varchar (30) |  |
| **DateOfBirth** | datetime (6) |  |
| **is\_active** | tinyint (1) |  |
| **is\_admin** | tinyint (1) |  |
| **phone** | varchar (15) |  |
| **is\_lock** | tinyint (1) |  |
| **last\_login** | datetime (6) |  |

**Bảng 3.1 Database User**

|  |  |  |
| --- | --- | --- |
| **Mô Tả** | **Lưu trữ thông tin của Document** | |
| **id** | int (11) | Khóa chính |
| **DataDocumentAuthor** | int (11) | Khóa ngoại bảng User |
| **DataDocumentType** | varchar (20) |  |
| **DataDocumentName** | varchar (1000) |  |
| **DocumentStatus** | tinyint (1) |  |
| **SessionId** | int (11) |  |
| **DataDocumentFile** | varchar (1000) |  |

**Bảng 3.2 B̫ảng database DataDocument**

|  |  |  |
| --- | --- | --- |
| **Mô Tả** | **Lưu trữ các câu của Document** | |
| **id** | int (11) | Khóa chính |
| **DataDocumenNo** | int (11) | Khóa ngoại bảng DataDocument |
| **DataDocumentSentence** | varchar (500) |  |
| **DataDocumentSentenceLength** | int (11) |  |

**Bảng 3.3 B̫ảng database DataDocumentContent**

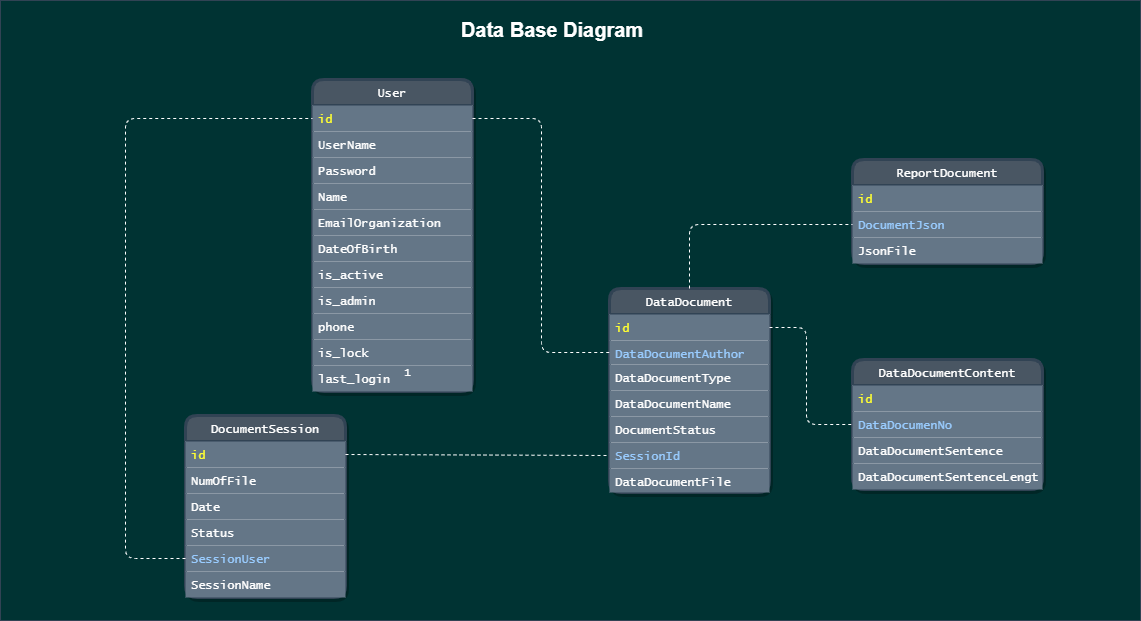
|  |  |  |
| --- | --- | --- |
| **Mô Tả** | **Lưu trữ thông tin của Session** | |
| **id** | int (11) | Khóa chính |
| **NumOfFile** | int (11) |  |
| **Date** | datetime (6) |  |
| **Status** | tinyint (1) |  |
| **SessionUser** | int (11) | Khóa ngoại bảng User |
| **SessionName** | varchar (50) |  |

**Bảng 3.4 B̫ảng database DocumentSession**

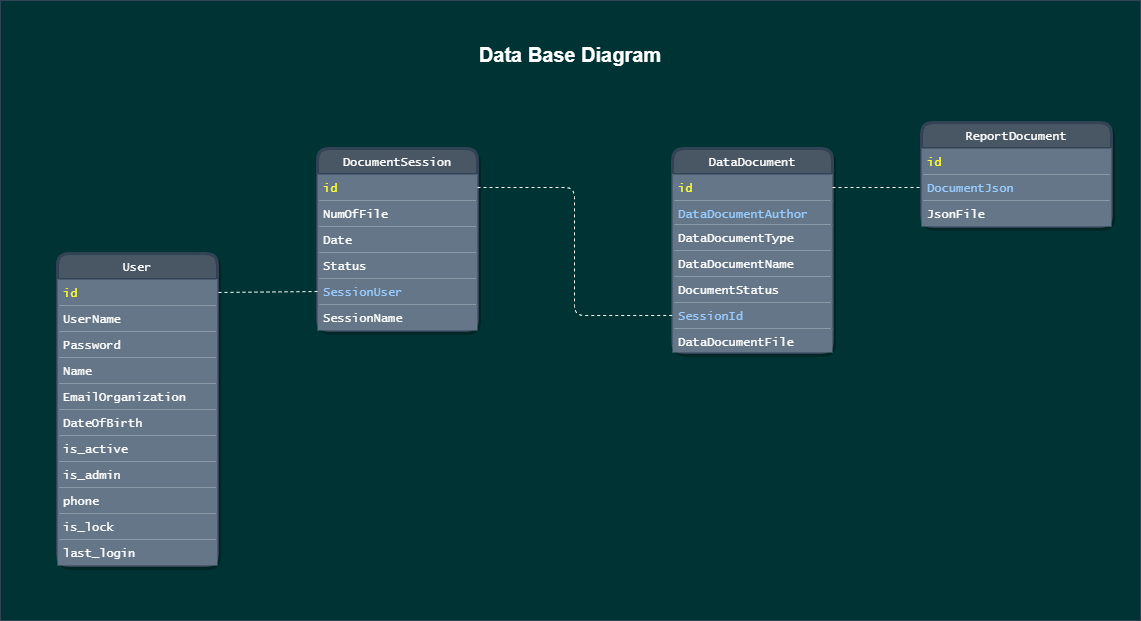
|  |  |  |
| --- | --- | --- |
| **Mô Tả** | **Lưu trữ thông tin của Report** | |
| **id** | int (11) | Khóa chính |
| **DocumentJson** | int (11) | Khóa ngoại bảng Document |
| **JsonFile** | longtext (0) |  |

**Bảng 3.5 B̫ảng database ReportDocument**

Sau đây nhóm xin trình bày các hình từ Hình 3.2 và 3.3 để mô tả các bảng database được sử dụng và mối quan hệ giữa chúng trong các tính năng của đồ án.



**Hình 3.2 Class diagram cho chức năng kiểm tra tương đồng của văn bản**



**Hình 3.3 Class diagram cho chức năng xem report**

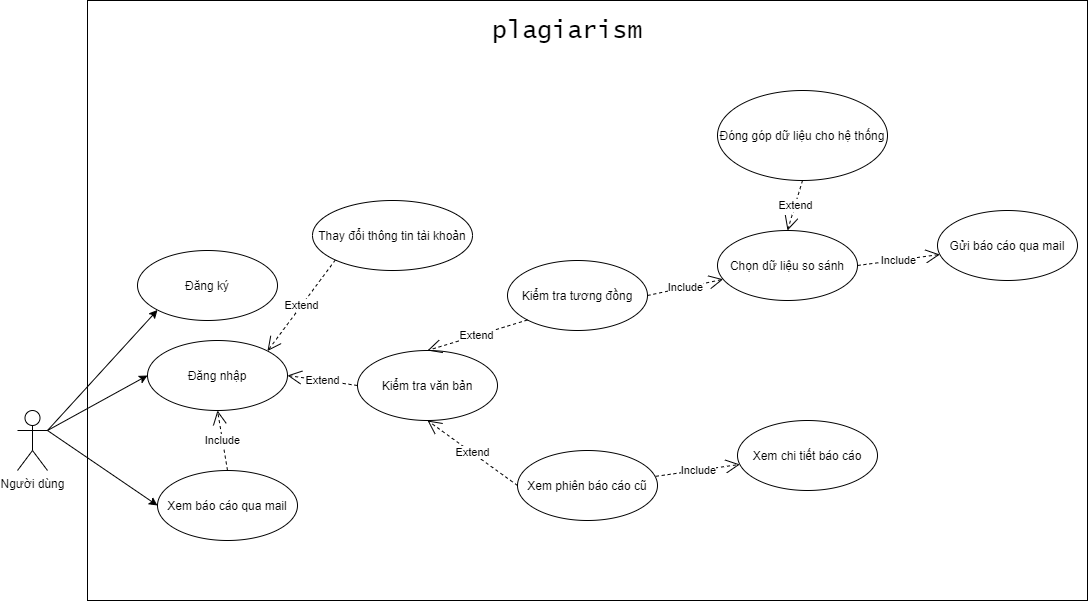
## Sơ đồ usecase của hệ thống Plagiarize

### Tổng quan hệ thống

Hệ thống của Plagiarize với chức năng kiểm tra tương đồng của văn bản và xem báo cáo cần người dùng cung cấp những thông tin cơ bản cũng như đăng ký tài khoản để có thể sử dụng.

Luồng hoạt động của hệ thống ở mức tổng quan cho thấy các bước chính trong quá trình truy cập.

Các use case trong hệ thống Plagiarize được thể hiện trong hình



**Hình 3.4 Use case diagram cho hệ thống xác thực và các tính năng chính của ứng dụng Plagiarize**

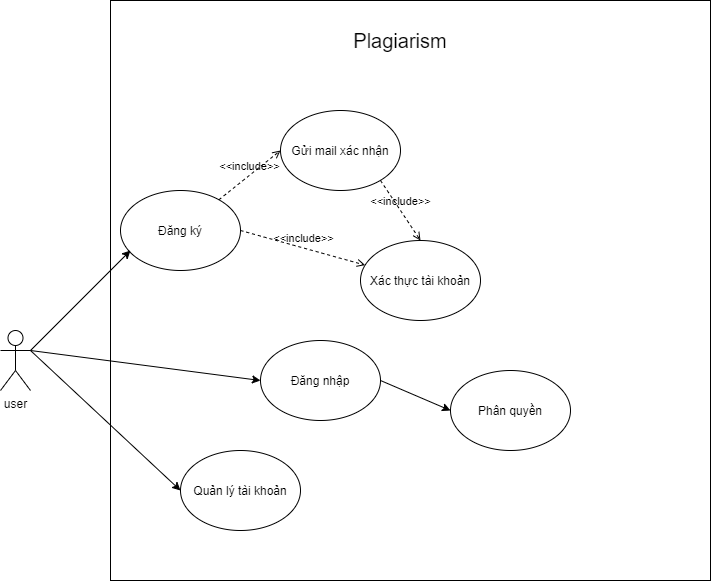
### Hệ thống quản lý tài khoản

Hệ thống quản lý tài khoản của Plagiarize cho phép người dùng sử dụng đầy đủ những chức năng cơ bản của người dùng như đăng ký, đăng nhập, đổi thông tin cá nhân và gửi email xác minh trong trường hợp người dùng quên mật khẩu.

Hầu hết các chức năng của Plagiarize đều được thực hiện dựa trên email của người dùng, nên hệ thống yêu cầu người dùng phải sử dụng đúng email của bản thân để xác minh mới cho phép người dùng tạo lập tài khoản.

Hệ thống có sử dụng Json Web Token để tăng tính bảo mật cho người dùng.

Hệ thống có thể tạo ra những tài khoản admin có chức năng quản lý người dùng, vậy nên nếu ai đăng nhập bằng một tài khoản admin thì người đó sẽ được chuyển tới màn hình quản lý của admin.



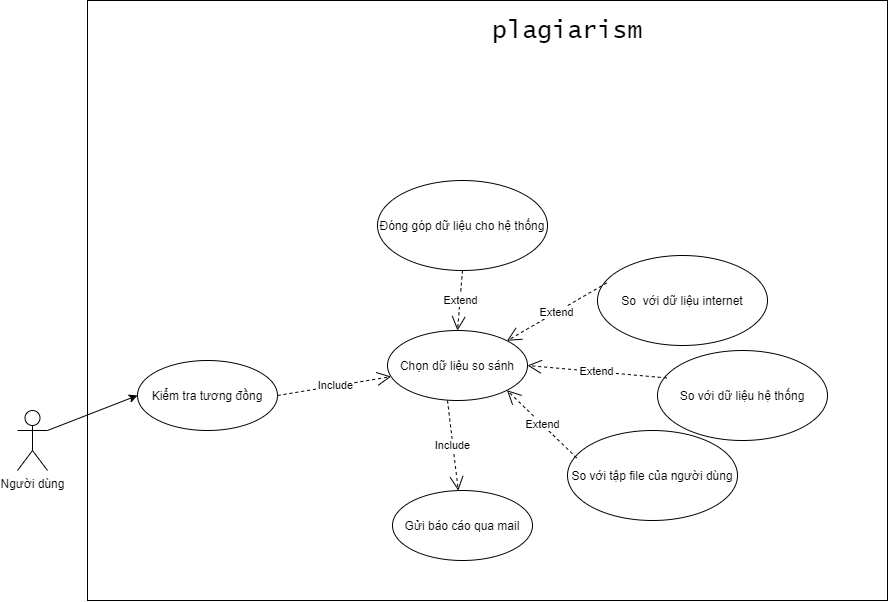
**Hình 3.5 Use case dìagram cho quy trình đăng nhập, xác thực, phân quyền**

### Kiểm tra tương đồng

Tính năng kiểm tra tương đồng là chức năng cho phép người dùng tiến hành kiểm tra độ tương đồng của văn bản tải lên. Với nhiều lựa chọn đa dạng về đối tượng dữ liệu mà người dùng muốn so sánh với như : so với tập dữ liệu của người dùng tải lên, dữ liệu lưu trữ trên hệ thống, dữ liệu tìm kiếm trên internet.

Người dùng có lựa chọn đóng góp dữ liệu cho hệ thống với các bước đơn giản.

Use case diagram thể hiện chi tiết tính năng kiểm tra tương đồng được thể hiện trong hình 3.6.



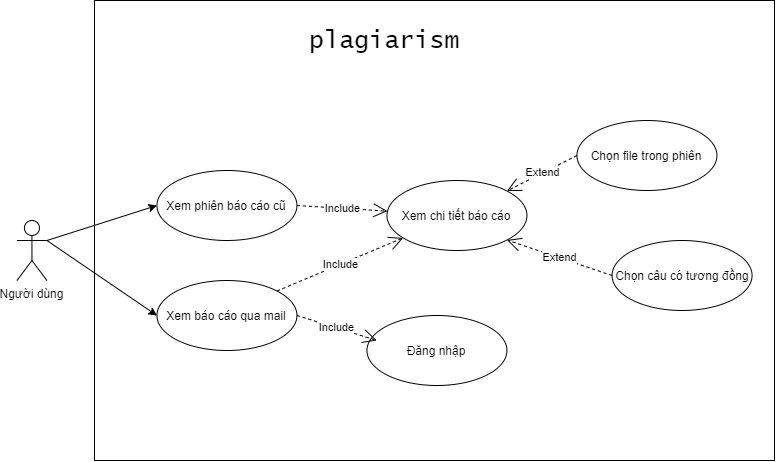
**Hình 3.6 Use case diagram cho chức năng kiểm tra tương đồng của văn bản**

### Xem nội dung report

Tính năng xem nội dung report(báo cáo) là chức năng cho phép người dùng xem nội dung của các phiên kiểm tra tương đồng văn bản. Trong mỗi phiên, hệ thống cho phép người dùng xem chi tiết về mức độ tương đồng của văn bản cần so sánh với các đối tượng dữ liệu đưa ra.

Người dùng có thể xem và đối chiếu tỉ lệ tương đồng với từng văn bản một, từng câu có độ tương đồng tương ứng dễ dàng.

Use case diagram thể hiện chi tiết tính năng xem report ở trong hình.



**Hình 3.7 Use case diagram cho chức năng xem nội dung report của phiên kiểm tra**

### Nhóm chức năng dành cho admin

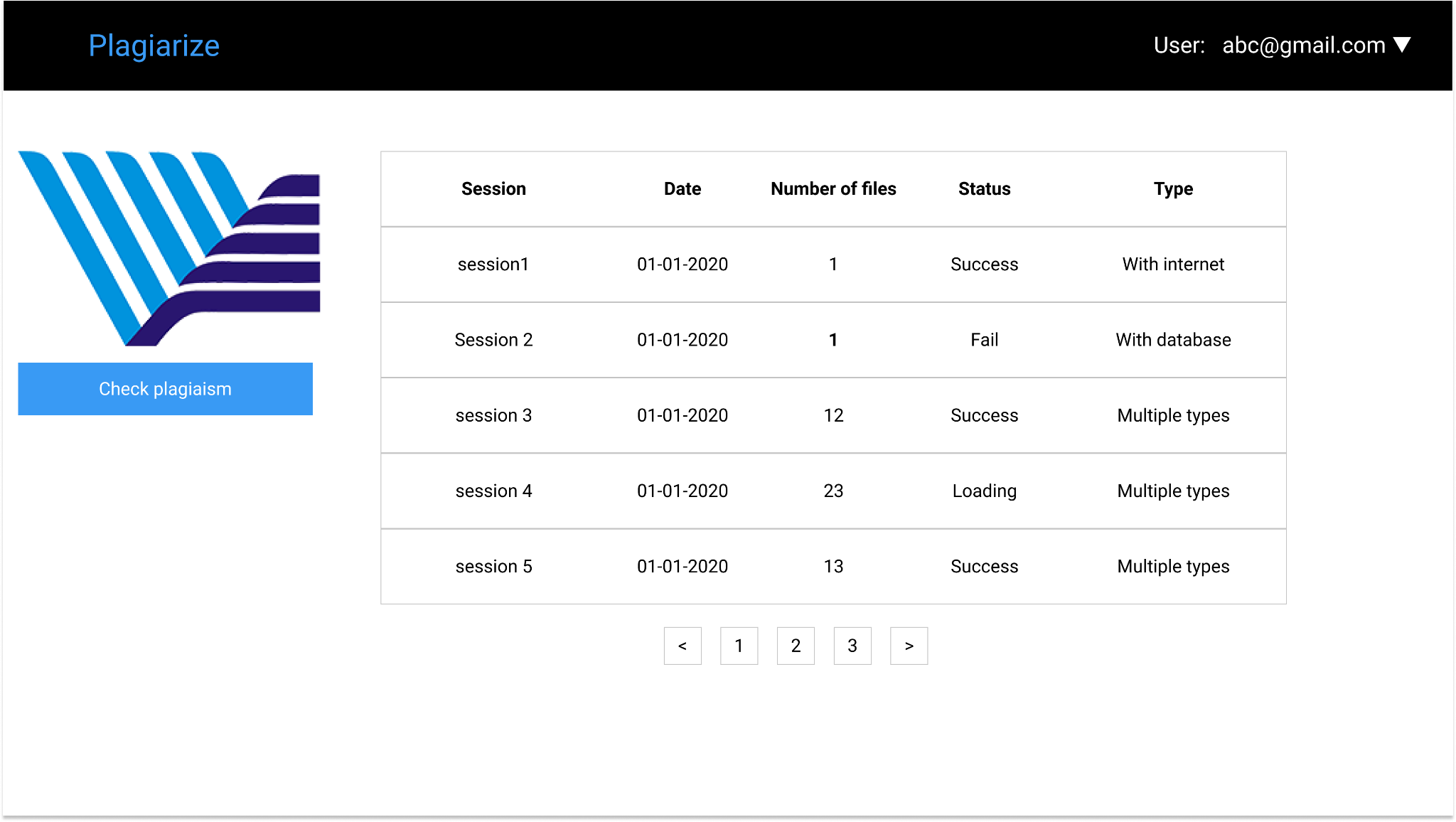
Tính năng quản lý người dùng: admin có quyền khóa 1 tài khoản của một người dùng bất kỳ, tài khoản bị khóa sẽ không thể đăng nhập vào được ứng dụng.

Tính năng upload database: admin có thể bổ sung database của hệ thống bằng cách tương tác trực tiếp với giao diện để upload số lượng file mong muốn miễn là toàn bộ file đó phù hợp với quy định.

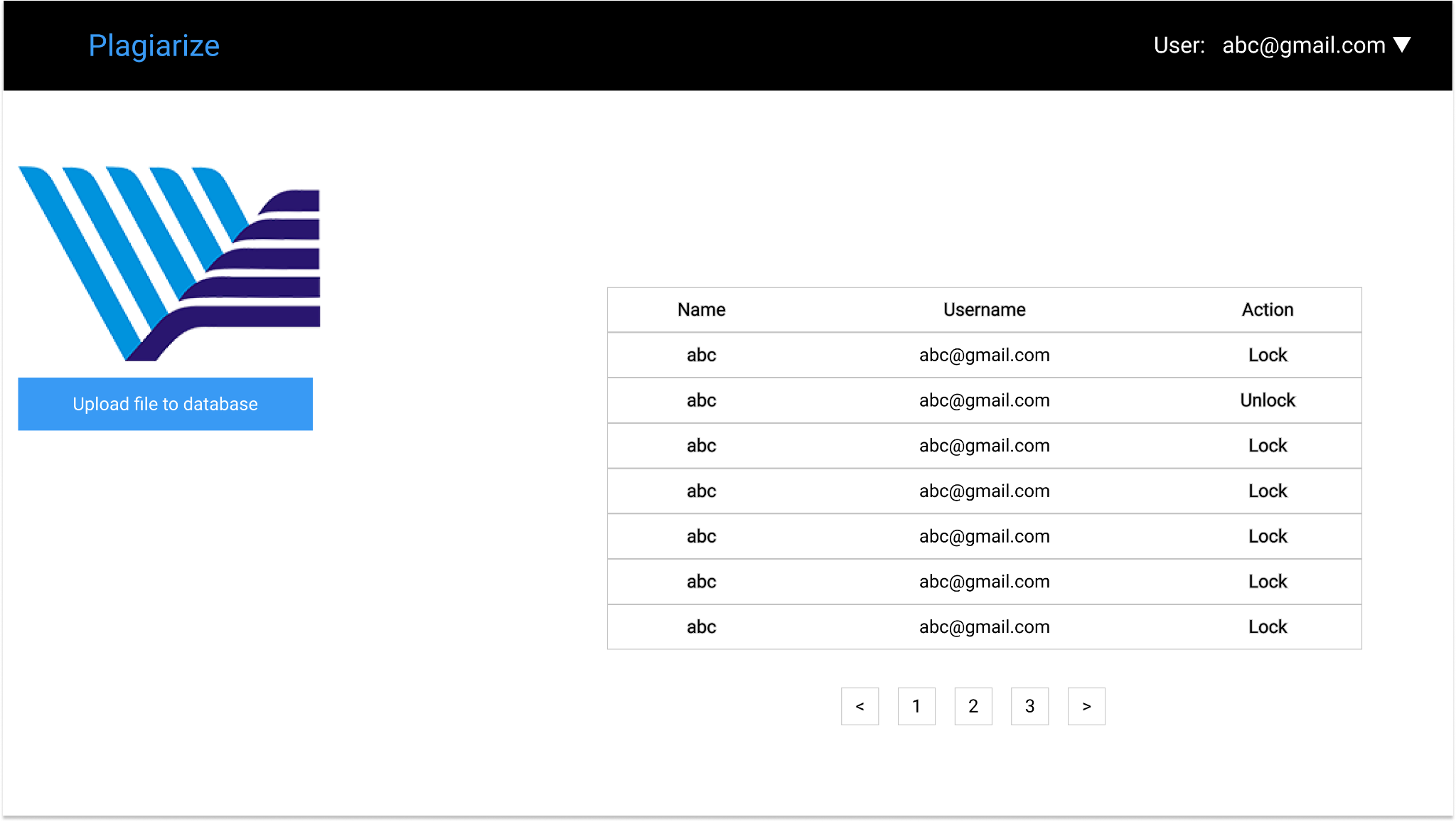
**Hình 3.8 Use case diagram cho chức năng của admin**

## Prototype

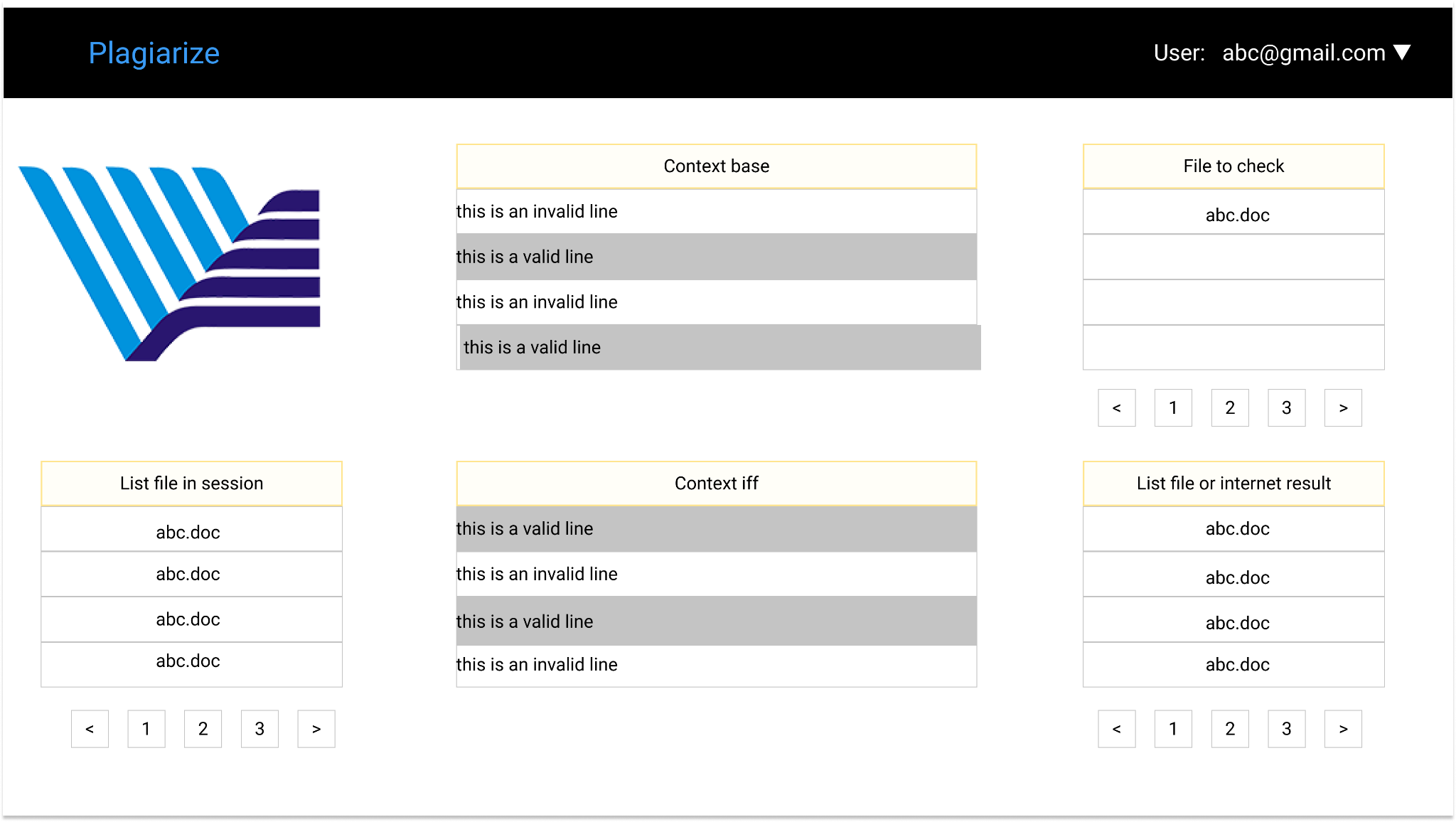
Sau đây nhóm sinh viên xin sử dụng những hình ảnh dưới đây làm prototype mẫu ban đầu cho những chức năng chính.



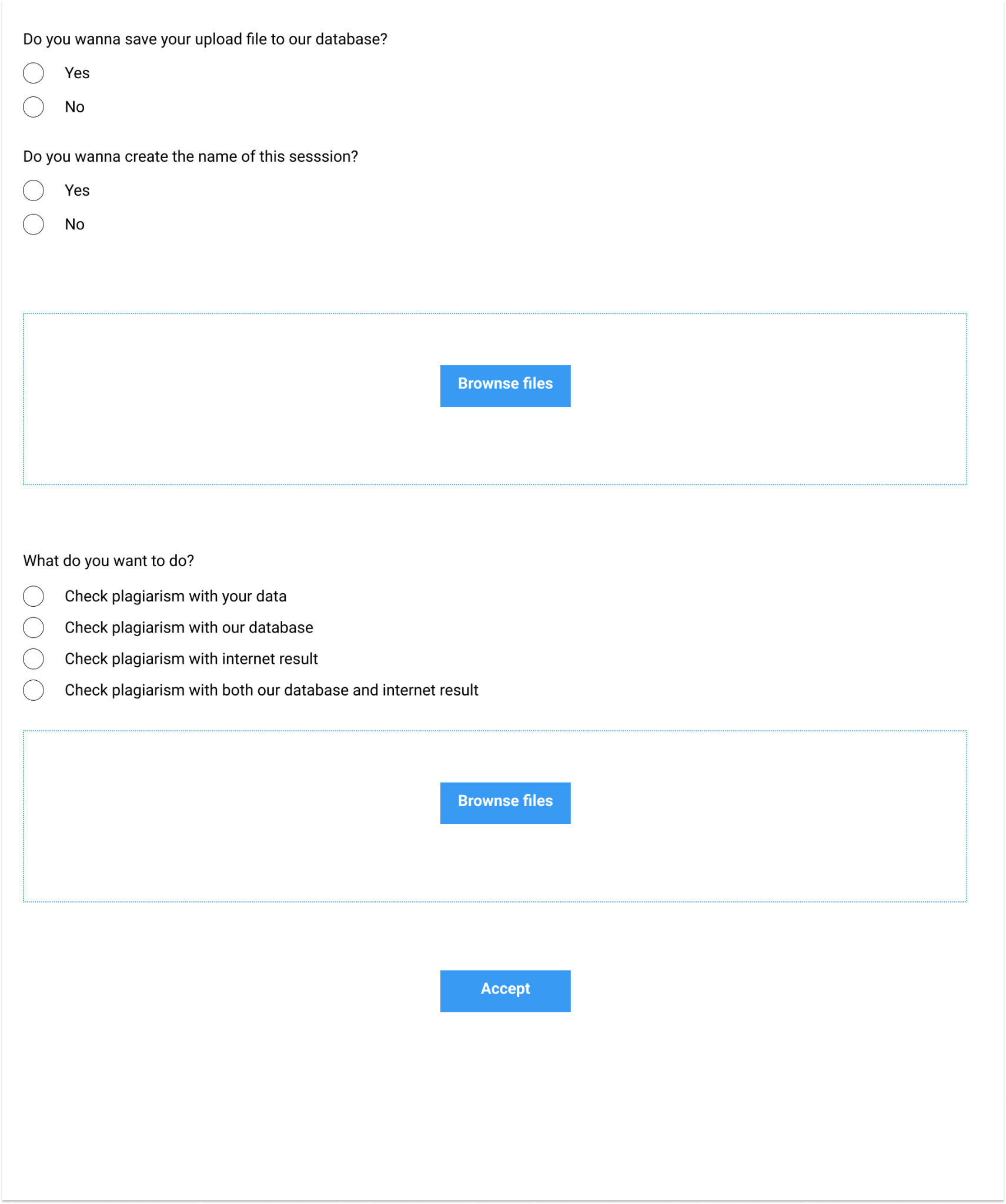
**Hình 3.9 Màn hình chính của người dùng**



**Hình 3.10 Màn hình chính của admin**



**Hình 3.11 Màn hình so sánh kết quả**



**Hình 3.12 Màn hình tải tệp lên để so sánh tương đồng**

## Tổng kết chương 3

Trong chương 3 nhóm sinh viên đã trình bày về các sơ đồ ở bước phân tích và thiết kế. Bên cạnh đó trình bày các bản mẫu prototype của trang web mà nhóm sẽ thực hiện.

# Phương pháp sử dụng

*Trong chương này nhóm sẽ trình bày phương pháp tiếp cận bài toán, các hướng giải quyết cụ thể cho từng chức năng của đồ án.*

Khi một tập tin văn bản mà người dùng tải lên hệ thống, sẽ trải qua các bước tiền xử lý chính như sau:

1. Rút trích văn bản
2. Tách câu, tách từ
3. Tính độ tương đồng (so với dữ liệu cục bộ hoặc do người dùng tải lên)

Trong trường hợp người dùng chọn chức năng tìm kiếm trên internet, các bước dưới đây sẽ tiếp tục được thực hiện:

1. Rút trích từ khóa chính
2. Tìm kiếm trên internet thông qua từ khóa chính
3. Download hoặc lấy dữ liệu từ website

Lặp lại bước 2 và bước 3, nhưng lúc này là so sánh với dữ liệu đã tìm được

## Tiền xử lý

Thư viện chính được sử dụng trong bước tiền xử lý là thư viện underthesea (GNU license).

Underthesea là một thư viện mã nguồn mở hỗ trợ xử lý ngôn ngữ tự nhiên tiếng Việt. Cung cấp cho người dùng các API tiện lợi và nhanh chóng trong các thao tác xử lý trên tiếng Việt như: tách từ, tách câu, gán nhãn cho từ, phân loại văn bản,...

Có rất nhiều thư viện hỗ trợ xử lý ngôn ngữ tự nhiên tiếng Việt như VnCoreNLP, pyvi,.... Nhưng nhóm chọn underthesea vì các lý do sau:

* Tốc độ: VnCoreNLP tốn nhiều thời gian khởi động và kết nối với server hơn các thư viện khác.
* Độ chính xác: pyvi không thể phân biệt tên riêng và các loại từ khác một cách chính xác hoàn toàn, có một số lỗi trong việc gán nhãn các loại từ.

### Rút trích văn bản

Các loại văn bản hệ thống có thể xử lý bao gồm: doc, docx, xlsx, csv, ppt, pptx, rtf, pdf, txt.

**Docx**

Có 4 bước chính để rút trích toàn bộ text:

**Bước 1:** Chuyển đổi từ file docx sang file XML

Sử dụng function opc\_to\_flat\_opc của thư viện docx-utils để chuyển file định dạng .docx sang file .xml và được đặt tên mặc định là “mydocx.xml”.

File XML sẽ chứa các tag đánh dấu như: <w:r>, <w:tbl>, <w:tr>, <w:t>,...

**Bước 2:** Parse file XML bằng cách sử dụng minidom

**Bước 3:** Lọc tag chứa text

Tag duy nhất chứa text là tag <w:t>. Trước tiên, phải lọc toàn bộ các tag <w:t> bằng phương thức getElementsByTagName(‘w:t’)

Các tag đã lọc khi in ra màn hình sẽ được chương trình thông báo như sau:

<DOM Element: w:t at 0x1914b9a4d30>

**Bước 4:** Rút trích text từ các tag <w:t>

Text sẽ được chứa trong tag <w:t> với ví dụ như sau: <w:t>Elephant</w:t>. “Elephant” chính text cần được được rút trích.

Với mỗi tag, rút trích text bằng lệnh sau: <tag>.firstChild.data

Tương tự cho toàn bộ các tag <w:t> khác, kết quả cuối cùng là toàn bộ text có trong file docx.

**Doc**

Nếu là định dạng doc, file doc sẽ được chuyển thành file docx và tiếp tục được xử lý theo định dạng docx như đã nêu.

**Các định dạng file khác được nhóm sử dụng các thư viện tương ứng để rút trích text.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | xlsx | csv | ppt/pptx | rtf | pdf | txt |
| Thư viện | pandas/csv | csv | pptx | x | pdfminer3 | x |

Văn bản sau khi rút trích sẽ là văn bản thô không bao gồm các định dạng như in đậm, in nghiêng,...

### Tách câu, tách từ

Thư viện sử dụng trong bước này là underthesea.

Từ một văn bản thô ban đầu sẽ trải qua các bước thực hiện như sau:

1. Tách đoạn: Tách văn bản thành các đoạn dựa vào các ký tự xuống dòng (‘\n’).
2. Tách câu: Sau khi đã tách đoạn văn bản, tiếp theo là tách thành các câu bằng underthesea.
3. Tách từ: Tiếp theo là tách từng câu trên thành thành các từ, giữa các chữ trong 1 từ được nối bằng dấu ‘\_’.

### Rút trích từ khóa

Phương pháp dùng cho việc rút trích từ khóa chính của văn bản là sử dụng thuật toán TF-IDF.

#### **Định nghĩa**

TF-IDF là một thống kê số học nhằm phản ánh tầm quan trọng của một từ đối với một văn bản trong một tập hợp hay một ngữ liệu văn bản. TF-IDF thường dùng dưới dạng là một trọng số trong tìm kiếm truy xuất thông tin, khai thác văn bản, và mô hình hóa người dùng.

#### **Công thức**

TF – Term frequency

TF là tần số xuất hiện của 1 từ trong 1 văn bản

|  |  |  |
| --- | --- | --- |
|  |  | **4.1** |

* Số lần xuất hiện của từ trong văn bản
* Số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản

IDF - Inverse document frequency

IDF là tần số nghịch của 1 từ trong tập văn bản

|  |  |  |
| --- | --- | --- |
|  |  | **4.2** |

* Tổng số văn bản trong tập
* Số văn bản chứa từ .

Trong trường hợp từ không xuất hiện ở bất kỳ văn bản nào thì mẫu số sẽ bằng 0. Vì thế trong trường hợp này, ta sẽ sửa đổi công thức từ công thức (4.2) thành công thức (4.3) mà không ảnh hưởng nhiều đến giá trị của nó như sau:

|  |  |  |
| --- | --- | --- |
|  |  | **4.3** |

#### **Tiền xử lý tập các từ**

Với các từ đã được tách từ văn bản, sẽ được tiền xử lý như sau:

Loại bỏ stop words: các từ xuất hiện nhiều mà không mang lại ý nghĩa hoặc không đóng góp giá trị nào cho văn bản sẽ được loại bỏ, bao gồm những từ như: và, những, như, các,…

Loại bỏ các ký tự đặc biệt, các dấu câu và các con số như: Ω, ∞, β, 1972, 5.65, ?, !,…

Các từ hợp lệ khi đưa vào thuật toán TF-IDF là những từ rõ ràng như: tự\_tin, chia\_sẻ, hài\_hước, bàn, ghế, thuyền,….

#### **Kết quả TFIDF**

Giá trị TF-IDF được tính bằng công thức sau:

|  |  |  |
| --- | --- | --- |
|  |  | **4.4** |

Kết quả xuất ra là tập các từ được đưa vào thuật toán và mỗi từ mang một giá trị TF-IDF. Từ có giá trị TF-IDF cao nhất là từ khóa chính của văn bản với độ chính xác cao nhất, các từ có giá trị thấp hơn thì độ chính xác càng giảm dần.

#### **Rút trích cụm từ**

Với việc chỉ sử dụng duy nhất 1 từ khóa để tìm kiếm trên internet thì rất khó mà có thể tập trung vào một nội dung cụ thể mà văn bản cần kiểm tra đang đề cập tới.

Ví dụ: Một bài báo cáo về thực tập tại ngân hàng có từ khóa chính là từ “tín dụng”. Nhưng khi tìm kiếm trên google search với từ “tín dụng” thì sẽ không thể nào tìm kiếm được các nội dung mà bài báo cáo có thể đạo.

Để tăng độ chính xác và tập trung vào nội dung chi tiết hơn thì cụm từ chính là giải pháp thay cho 1 từ. Cụ thể thuật toán sẽ tìm những câu có giá trị TFIDF cao nhất, giá trị TFIDF của câu được tính bằng cách lấy tổng các giá trị TFIDF của từng từ trong câu. Tiếp theo cụm từ được xác định bằng cách lấy cụm N từ liên tiếp sao cho cụm N từ có tổng giá trị TFIDF cao nhất trong tất cả cụm N từ có trong câu (với N là số lượng các từ tồn tại giá trị TFIDF).

## Tìm kiếm internet thông qua từ khóa

Sau khi rút trích được từ khoá của văn bản ở phần (4.1.3) thì sẽ tìm các nguồn tài liệu liên quan đến từ khoá của văn bản trên internet. Nhóm sử dụng thư viện ScrapeSearchEngine (MIT license) để thực hiện tìm kiếm các nguồn dữ liệu tiềm năng.

Các nguồn tìm được chia làm 2 loại chính:

* Văn bản: các tệp văn bản định dạng pdf, doc, docx…
* Các trang web HTML.

Đối với các nguồn là các văn bản thì sẽ tải về và thực hiện các bước tiền xử lý rồi tiến hành so sánh với văn bản của người dùng.

Đối với các trang web HTML thì sử dụng thư viện Beautiful Soup để rút trích các dữ liệu có trong trang web và sau đó tiến hành so sánh với văn bản của người dùng.

## Kiểm tra tương đồng

Nhóm đã sử dụng 2 phương pháp tiếp cận cho việc kiểm tra tương đồng của văn bản gồm:

* Thuật toán tính khoảng cách Levenshtein
* Mô hình học sâu Siamese

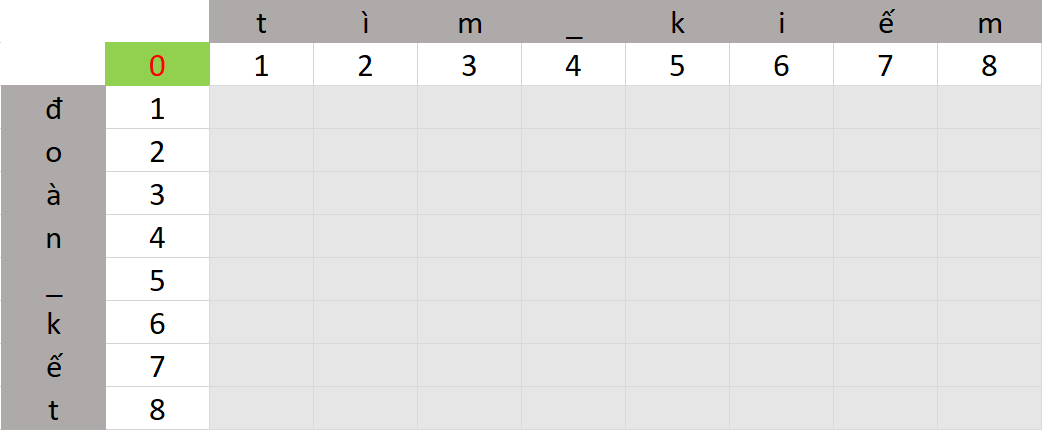
### Thuật toán tính khoảng cách Levenshtein

Khoảng cách Levenshtein giữa 2 chuỗi A và B là số lần sử dụng các các thao tác thêm, xóa, thay thế ký tự để biến chuỗi A thành chuỗi B.

Thuật toán được tính dựa trên ma trận khoảng cách của từng ký tự, có công thức xác định như sau:

Để rõ ràng hơn, ta áp dụng thuật toán vào một bài toán cụ thể như sau: tính khoảng cách Levenshtein giữa 2 từ “tìm\_kiếm” và “đoàn\_kết”.

**Bước 1**: Gán từng ký tự của các từ vào ma trận và đánh dấu vị trí như sau:



**Hình 4.1 Ma trận tách các từ thành các kí tự**

**Bước 2**: Tính giá trị tại từng ô trong ma trận

Giá trị tại vị trí , gọi là phụ thuộc vào 2 ký tự của hàng và cột có giống nhau hay không. Sẽ xảy ra 2 trường hợp:

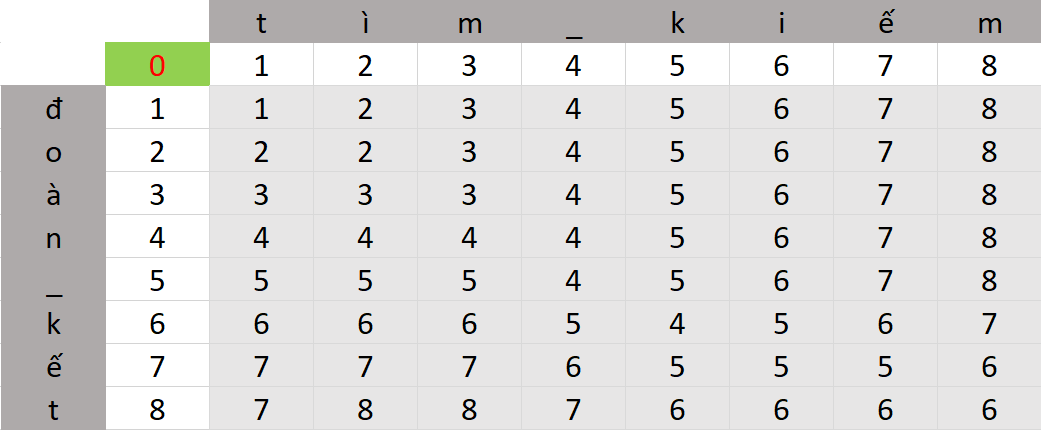
Giống nhau: . Tổng quát ta có công thức sau:

|  |  |  |
| --- | --- | --- |
|  |  | **4.5** |

|  |  |  |
| --- | --- | --- |
|  |  | **4.6** |

Khác nhau: . Tổng quát ta cũng có công thức sau:

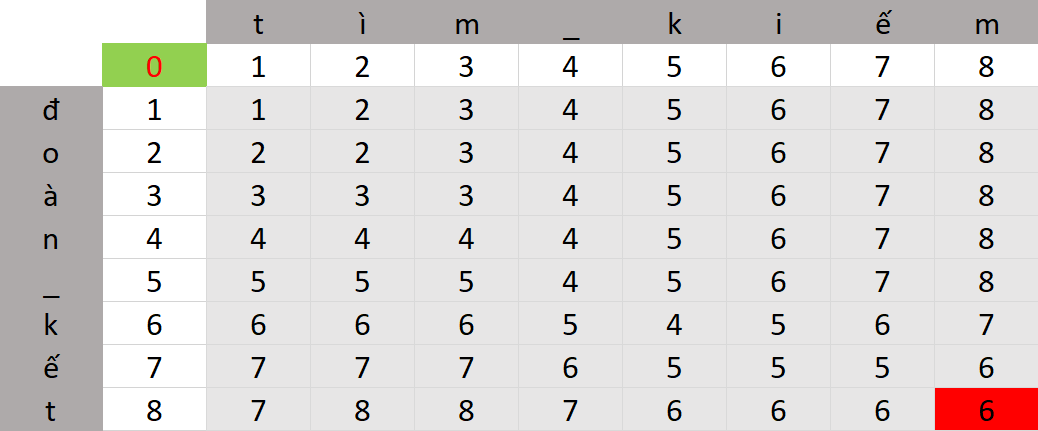
Vì “t” và “đ” là 2 ký tự khác nhau nên . Tương tự với các vị trí còn lại, ta sẽ xác định được toàn bộ ma trận như hình ảnh bên dưới.



**Hình 4.2 Ma trận giá trị D(i,j)**

**Bước 3**: Xác định giá trị khoảng cách

Giá trị khoảng cách là giá trị tại ô ( lần lượt là số hàng và số cột của ma trận)



**Hình 4.3 Ma trận kết quả**

Như vậy, khoảng cách Levenshtein của 2 từ “tìm\_kiếm” và “đoàn\_kết” là 6. Khoảng cách đó có ý nghĩa là cần ít nhất 6 thao tác: thêm hoặc xóa hoặc thay thế ký tự để biến đổi từ “tìm\_kiếm” sang từ “đoàn\_kết”.

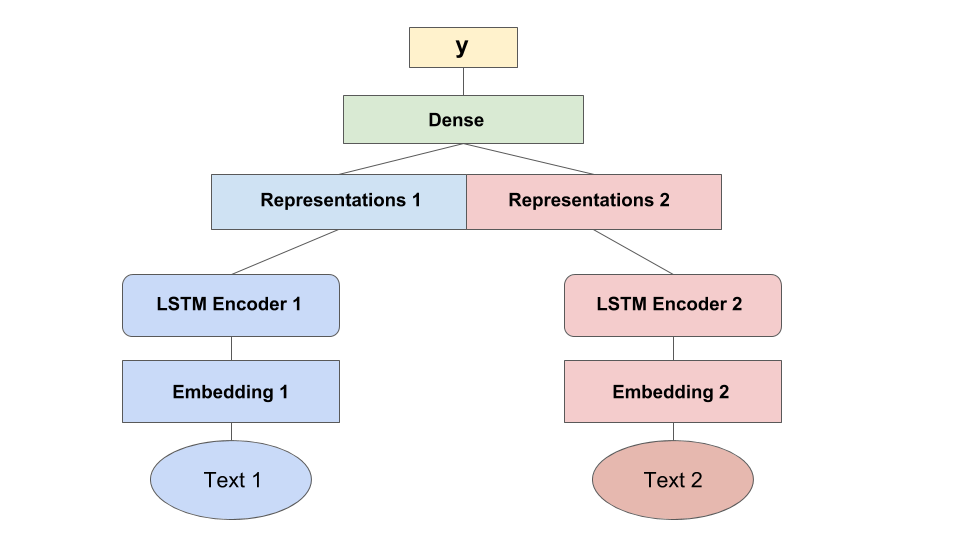
### Mô hình học sâu Siamese

Sieamese là một mạng nơ-ron Long Short-Term Memory dùng để xác đinh độ tương đồng về mặt ngữ nghĩa của các cặp câu với nhau với kích thước dài ngắn tuỳ ý. Mạng Sieamese được huấn luyện thông qua một ma trận Manhattan đơn giản.

Khác với các công trình nghiên cứu về việc xác định ngữ nghĩa của văn bản trước đây dùng nhiều thuật toán phức tạp được tính từ các đặc điểm của cặp câu input thì mô hình Sieamese đạt được một bước tiến xa hơn. Mô hình này được huấn luyện từ một tập ngữ liệu lớn và sau đó dự đoán cho các input chưa từng gặp qua trong tập huấn luyện trước đây.

#### **Sơ lược ý tưởng của bài toán**

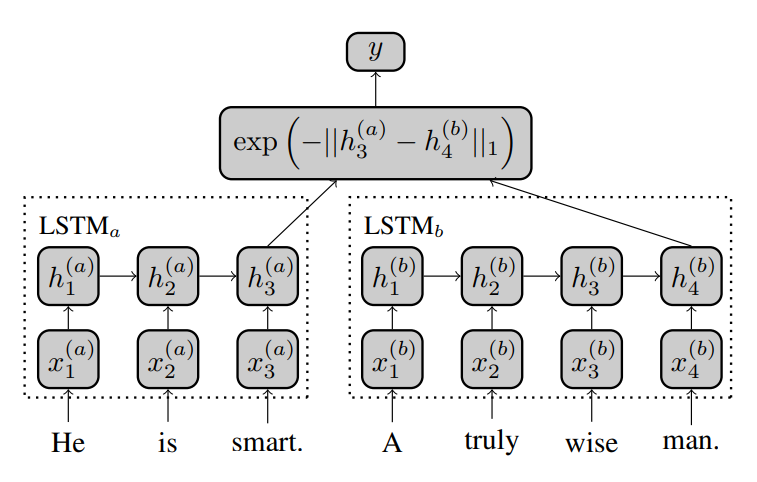
Mô hình bao gồm 2 hay nhiều mạng nơ-ron con giống nhau có cùng cấu trúc, cùng số lượng tham số và cùng số lượng trọng số với nhau.



**Hình 4.4 Mô tả cấu trúc mạng Sieamese**

Mỗi cặp mạng con được giao nhiệm vụ huân luyện một cặp câu đầu vào đã được gán nhãn về độ tương đồng.

Mỗi cặp câu input được chuyển thành vector theo một mô hình word2vec của Mikolov công bố năm 2013 với số chiều đầu vào là . Sau khi được đưa qua ma trận Manhattan thì số chiều tầng ẩn cuối cùng chỉ còn

Và mỗi mạng con sẽ học được từ mỗi câu input thành một vector đầu ra, khoảng cách giữa 2 vector đó là độ tương đồng về ngữ nghĩa của 2 câu đó.

**Hình 4.5 Ví dụ của một ma trận Manhattan**

Phương pháp này chỉ dừng ở bước thử nghiệm chưa được cải tiến và tối ưu vì dữ liệu huấn luyện còn hạn chế nên kết quả thu được là chưa đáng kể.

## Tổng kết chương 4

Qua chương này nhóm đã xác định các chức năng chính, các cách giải quyết từng chức năng một cách chi tiết và cụ thể. Tuy nhiên vẫn còn hạn chế ở một số chức năng chưa được tốt như mong đợi của nhóm.

# Giới thiệu các công nghệ sử dụng

*Nội dung chương này sẽ giới thiệu các công nghệ, kĩ thuật và các thư viện mà nhóm sử dụng để thực hiện đồ án.*

## Database và cơ sở dữ liệu MySQL

### Giới thiệu MySQL

Hệ quản trị cơ sở dữ liệu (Database Management System - DBMS) là phần mềm tương tác với người dùng, ứng dụng và chính cơ sở dữ liệu để thu thập và phân tích dữ liệu. Phần mềm DBMS bao gồm các tiện ích cốt lõi được cung cấp để quản trị cơ sở dữ liệu. Cùng với cơ sở dữ liệu, DBMS và các ứng dụng liên quan có thể được gọi là " Hệ thống cơ sở dữ liệu ". Thông thường thuật ngữ "cơ sở dữ liệu" cũng được sử dụng để nói đến bất kỳ DBMS, hệ thống cơ sở dữ liệu hoặc ứng dụng nào được liên kết với cơ sở dữ liệu.

MySQL là hệ quản trị cơ sở dữ liệu tự do nguồn mở phổ biến nhất thế giới và được các nhà phát triển rất ưa chuộng trong quá trình phát triển ứng dụng. Vì MySQL là hệ quản trị cơ sở dữ liệu tốc độ cao, ổn định và dễ sử dụng, có tính khả chuyển, hoạt động trên nhiều hệ điều hành cung cấp một hệ thống lớn các hàm tiện ích rất mạnh. Với tốc độ và tính bảo mật cao, MySQL rất thích hợp cho các ứng dụng có truy cập CSDL trên internet. Người dùng có thể tải về MySQL miễn phí từ trang chủ. MySQL có nhiều phiên bản cho các hệ điều hành khác nhau: phiên bản Win32 cho các hệ điều hành dòng Windows, Linux, Mac OS X, Unix, …

MySQL là một trong những ví dụ rất cơ bản về Hệ Quản trị Cơ sở dữ liệu quan hệ sử dụng Ngôn ngữ truy vấn có cấu trúc (SQL). MySQL được sử dụng cho việc bổ trợ NodeJs, PHP, Perl, và nhiều ngôn ngữ khác, làm nơi lưu trữ những thông tin trên các trang web viết bằng NodeJs, PHP hay Perl,...

|  |  |
| --- | --- |
| Kiểu dữ liệu | Kích thước |
| **TINYINT** | -128 → 127 |
| **SMALLINT** | -32768 → 32767 |
| **MEDIUMINT** | -8388608 → 8388607 |
| **INT** | - 2147483648 → 2147483647 |
| **BIGINT** |  |
| **FLOAT** | 4 bytes |
| **DOUBLE** | 8 bytes |
| **DATETIME** | 8 bytes |
| **DATE** | 3 bytes |
| **TIMESTAMP** | 4 bytes |
| **CHAR** | M × w bytes |
| **VARCHAR** | L + 2 bytes |
| **BINARY** | 0 - 255 |
| **VARBINARY** | 0 - 65635 |
| **ENUM** | An enumeration |

**Bảng 5.1 Kiểu dữ liệu và kích thước MySQL hỗ trợ**

### Ưu điểm của MySQL

**Tốc độ xử lý:** Khả năng chèn dữ liệu cực nhanh và hỗ trợ mạnh mẽ các chức năng chuyên dụng cho trang web. Các tính năng này cũng được sử dụng cho môi trường lưu trữ dữ liệu mà hệ quản trị này tăng cường đến hàng terabyte cho các server đơn. Nó cho phép lập chỉ mục (index) giảm thiểu thời gian tìm kiếm cho các câu truy vấn một cách tốt nhất.

**Tính bảo mật:** hệ quản trị cơ sở dữ liệu MySQL tích hợp các tính năng bảo mật an toàn tuyệt đối. Chỉ có những người sử dụng đã được xác nhận mới truy cập được vào cơ sở dữ liệu.

### Nhược điểm MySQL

**Giới hạn:** với nhiều tiện ích và hỗ trợ, đồng nghĩa với việc chi phí phải bỏ ra một cách tương xứng. Chi phí cho kích thước của hệ thống càng lớn sẽ càng tốn kém.

**Khả năng lưu trữ:** Mặc dù MySQL là sản phẩm công nghệ mã nguồn mở nhưng vẫn gặp rất nhiều vấn đề và phàn nàn kể từ khi được mua lại quyền sở hữu. Dung lượng có thể lưu trữ bị hạn chế, khi số lượng record tăng lên sẽ gây khó khăn khi truy suất cơ sở dữ liệu.

## Framework xây dựng backend - Django

### Giới thiệu về Django

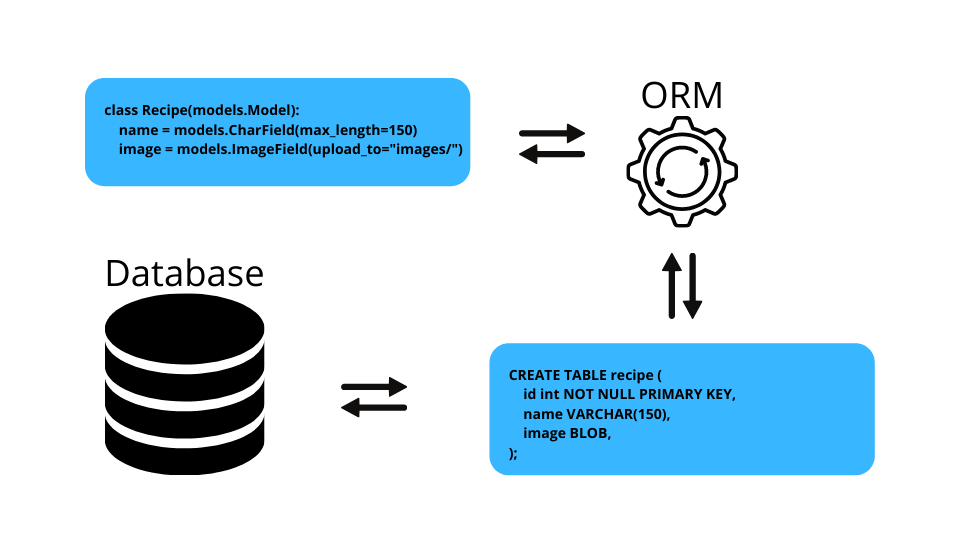
Django là một web framework khá nổi tiếng và được nhiều lập trình viên sử dụng hiện nay. Nó được viết hoàn toàn bằng ngôn ngữ Python, hỗ trợ đa dạng các thư viện và tính năng tích hợp sẵn cho người lập trình.

### Kiến trúc sử dụng của Django

Django sử dụng mô hình MVC để phát triển ứng dụng website, đơn giản hóa việc xây dựng các website phức tạp mà phải sử dụng đến cơ sở dữ liệu như MySQL, PostgreSQL. Nó giúp lập trình viên xây dựng các component riêng biệt có thể tái sử dụng và có thể tự chạy, đẩy nhanh quá trình phát triển, không phải xây dựng lại từ đầu.

### Lợi ích của Django

Với mô hình tách biệt rõ ràng từng thành phần, các component có thể được xây dựng riêng biệt tạo nên thuận lợi trong việc phát triển hệ thống nhiều nội dung cùng một lúc. Đẩy nhanh được quá trình phát triển phần mềm. Áp dụng ánh xạ quan hệ-đối tượng ORM(Object Relational Mapping) với ngôn ngữ lập trình hướng đối tượng, các Class sẽ được map thành 1 bảng trong database, 1 instance sẽ được map thành 1 record.



**Hình 5.1 Mô tả mối quan hệ của ORM**

## Ngôn ngữ lập trình Python

### Giới thiệu về Python

Python là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ, phổ biến và thông dụng đối với lập trình viên hiện nay. Được đánh giá là một trong những ngôn ngữ tốt nhất cho người mới tiếp xúc với lập trình. Không chỉ thế, mà đối với người lập trình lâu năm, nó cũng được sử dụng rộng rãi trong nhiều lĩnh vực và nhiều nền tảng khác nhau.

### Điểm nổi bật của Python

**Cấu trúc:** Với cú pháp đơn giản, được rút gọn để thuận tiện cho công việc lập trình, Python được đánh giá là dễ tiếp cận hơn nhiều khi so sánh với các ngôn ngữ khác như C#, java, C++. Nó giúp người dùng giảm thiểu các lỗi syntax cũng như typo không đáng có.

**Thư viện hỗ trợ**: Vì là mã nguồn mở, Python có một cộng đồng to lớn và phát triển. Đem lại một khối lượng thư viện đa dạng mà mọi người có thể sử dụng, giảm thiểu khối lượng công việc chỉ với các lệnh setup và import.

**Đa nền tảng**: với cấu trúc xây dựng đặc biệt, các chương trình được xây dựng bằng Python có thể di chuyển từ nền tảng này sang nền tảng khác mà không phải thay đổi câu lệnh nào. Nó chạy được trên hầu hết các nền tảng thông dụng hiện nay như Windows, MacOS, Linux.

## Framework xây dựng frontend - Angular

### Giới thiệu về angular

Angular là một javascript framework được google phát triển để dùng trong các single page application và được viết bằng typescript.Angular cung cấp rất nhiều tính năng tích hợp và thư viện rất tiện lợi để xây dựng các ứng dụng web.

### Lợi ích của angular

**Đa nền tảng**: Angular cho phép ứng dụng chạy tương thích tốt trên rất nhiều nền tảng di động khác nhau.

**Tiện lợi**: Angular được tích hợp rất nhiều tính năng, công cụ và các thư viện rất tiện lợi để hỗ trợ người dùng trong việc xây dựng trang web. Tiêu biểu trong số đó là angular cli và thư viện angular material.

**Dễ tổ chức và quản lý code**: angular rất phức tạp và khó sử dụng, nhưng điều đó khiến phần mềm sẽ được sắp xếp theo hướng module hóa khiến phần mềm gọn gàng, logic và giúp người dùng dễ tìm lỗi và chỉnh sửa hơn. Điều này rất tiện lợi khi người dùng muốn xây dựng một trang web với nhiều kích cỡ khác nhau.

**Dễ bảo trì, nâng cấp**: Angular hỗ trợ rất nhiều cách để bảo trì code rất hiệu quả.Khi thay đổi hoặc nâng cấp phiên bản, tất cả những package liên quan sẽ được tự động thay đổi hoặc nâng cấp - bao gồm cả HTTP, routing và thư viện angular material. Việc nâng cấp phiên bản cũng rất đơn giản với một câu lệnh: ”ng-update”. Ngoài ra angular cũng có sẵn hướng dẫn chi tiết những điều cần lưu ý khi nâng cấp từ phiên bản.

**Khả năng routing và binding tốt**:Angular có cơ chế routing cho phép tải trang một cách bất đồng bộ trên cùng một trang cho phép chúng ta tạo single page application hiệu quả, ngoài ra angular còn hoạt động theo cơ chế two way data-binding data hoàn toàn tự động.Cơ chế two way data-binding khiến bất cứ thay đổi nào trên view đều được tự động cập nhật vào component class và ngược lại.

**Thư viện thiết kế hiệu quả:** Angular có thư viện angular material cho phép người dùng sử dụng một lượng lớn các material component và module UI một cách dễ dàng .Hệ thống material design là một hệ thống thiết kế google tạo ra nhằm giúp người dùng có thể xây dựng một giao diện hiệu quả và có khả năng responsive cao.

## Xác thực bằng json web token - JWT

### Giới thiệu về JWT

JWT là một tiêu chuẩn để tạo ra các object theo định dạng json có tác dụng tăng tính bảo mật khi giao tiếp giữa hai bên client - server bằng json. Object trên có dạng chuỗi chứa những thông tin cần thiết để giao tiếp và đã được mã hóa và định nghĩa một cách nhỏ gọn và khép kín để truyền một cách an toàn thông tin giữa các dạng đối tượng json.

JWTs có thể được ký bằng một thuật toán bí mật (với thuật toán HMAC) hoặc một cặp public key và private key sử dụng mã hoá RSA, hiện nay JWT rất hay được sử dụng bởi những lợi ích và sự bảo mật nó mang đến.

### Giải thích về JWT

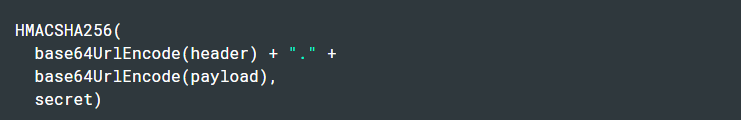
JWT là sự kết hợp của ba phần với nhau, mỗi phần được ngăn cách với nhau qua một dấu chấm. Nói một cách đơn giản thì cấu trúc của JWT có dạng như sau:

**xxxxxx.yyyyyy.zzzzzz**

Mỗi phần trong chuỗi JWT trên đều được encode base 64.

Ba phần trong cấu trúc trên bao gồm:

* Header: Chứa hai thông tin chính là loại token (mặc định là JWT - Thông tin này cho biết đây là một Token JWT) và thuật toán đã dùng để mã hóa (HMAC SHA256 - HS256 hoặc RSA).
* Payload: Chứa các claims. Claims ở đây có nghĩa là là những thông tin về thực thể ( người dùng,..) và các thông tin liên quan. Claims được chia làm ba loại là register claims, private claims và public claims.
* Signature: Signature trong JWT là một chuỗi được mã hóa bởi header, payload cùng với một chuỗi bí mật theo nguyên tắc sau:



**Hình 5.2 Nguyên tắc mã hoá JWT**

### Trường hợp cần sử dụng:

Ta có thể sử dụng JWT cho những trường hợp sau:

* Xác thực và phân quyền
* Trao đổi thông tin giữa client - server

### Cơ chế hoạt động:

Khi 1 người dùng đăng nhập thành công, server sẽ trả về 1 chuỗi JWT chứa những thông tin cần thiết để định danh người dùng đó.

Bất cứ khi nào người dùng muốn truy cập vào những vùng dữ liệu được bảo vệ cần xác thực, front end phải gửi một JWT ở trong header có cấu trúc như sau:



**Hình 5.3 Cấu trúc JWT**

Sau khi server nhận được request sẽ kiểm tra xem token đó có lỗi hay không, nếu được xác nhận là hợp lệ, user sẽ được truy cập vào những vùng được bảo vệ đó.Đối với hệ thông Plagiarize thì Bearer sẽ được thay thế bằng “JWT”.

### Lợi ích của JWT

**Tính cô đọng**: Toàn bộ yêu cầu xác thực đều được gói gọn trong header của 1 url request, điều này khiến ứng dụng web nhanh khi chuyển đổi và dễ sử dụng.

**Tự chứa thông tin cần thiết**: Chuỗi token trên chứa những thông tin quan trọng và cần thiết của người dùng, vậy nên không cần phải liên tục truy vấn đến database để lấy thông tin đó nữa.

**Tính bảo mật**: Thông tin trong chuỗi sẽ an toàn và mang độ tin cậy cao. Nếu bên thứ ba muốn giải mã chuỗi token được mã hóa đó thì sẽ tốn rất nhiều thời gian. Đối với ứng dụng thì việc làm mới token thường rất nhanh chóng và phụ thuộc hoàn toàn vào nhu cầu của người phát triển. Ngoài ra việc chèn thêm hay chỉnh sửa chuỗi token cũng là bất khả thi do chỉ cần 1 sai lệch nhỏ đã khiến token thành 1 chuỗi hoàn toàn khác.

**Đa nền tảng**: JWT có thể được sử dụng trên hầu hết các ngôn ngữ lập trình và có thể liên kết trực tiếp vào đối tượng.

**Tính phi trạng thái**: Server không cần phải lưu thông tin về các phiên của người dùng ở trong database hoặc memory, vì những thông tin cần tiết sẽ nằm ở trong JWT.

## Tổng kết chương 5

Trong chương này nhóm đã trình bày các kiến thức nền tảng của các công nghệ được sử dụng trong đồ án. Bên cạnh đó còn nêu lên ưu khuyết điểm của từng công nghệ và lý do nhóm lựa chọn công nghệ đó cho đồ án này.

# Triển khai và đánh giá sản phẩm

*Chương này sẽ trình bày kết quả của đồ án, chức năng và giao diện của đồ án.*

## Kết quả đạt được

Nhóm đã hoàn thành các chức năng của sản phẩm trong bảng sau:

|  |  |
| --- | --- |
| **Tên chức năng** | **Trạng thái chức năng** |
| Đăng ký tài khoản người dùng | Đã hoàn thành |
| Xác thực tài khoản bằng email | Đã hoàn thành |
| Quên mật khẩu | Đã hoàn thành |
| Đổi mật khẩu | Đã hoàn thành |
| Cập nhật các thông tin thay đổi | Đã hoàn thành |
| Thay đổi ngôn ngữ | Đã hoàn thành |
| Đăng nhập | Đã hoàn thành |
| Upload file cần kiểm tra | Đã hoàn thành |
| Upload file để kiểm tra | Đã hoàn thành |
| Kiểm tra với dữ liệu của người dùng | Đã hoàn thành |
| Kiểm tra với dữ liệu của hệ thống | Đã hoàn thành |
| Kiểm tra với dữ liệu tìm được từ internet | Đã hoàn thành |
| Lưu vào database của hệ thống | Đã hoàn thành |
| Gửi email kết quả so sánh | Đã hoàn thành |
| Xem kết quả từng phiên | Đã hoàn thành |
| Đặt tên phiên | Đã hoàn thành |
| Khoá tài khoản user | Đã hoàn thành |
| Cập nhật database hệ thống | Đã hoàn thành |

**Bảng 6.1 Danh sách chức năng**

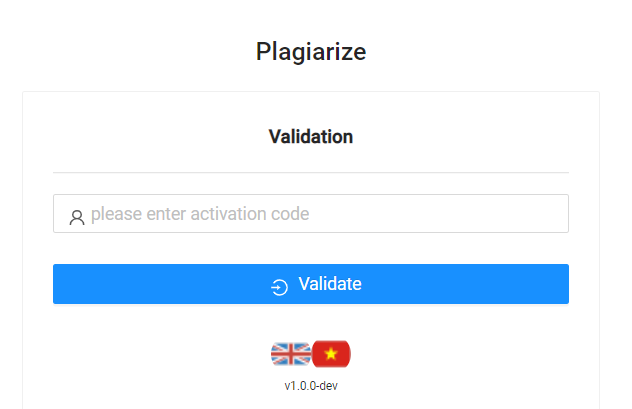
## Mô tả chức năng

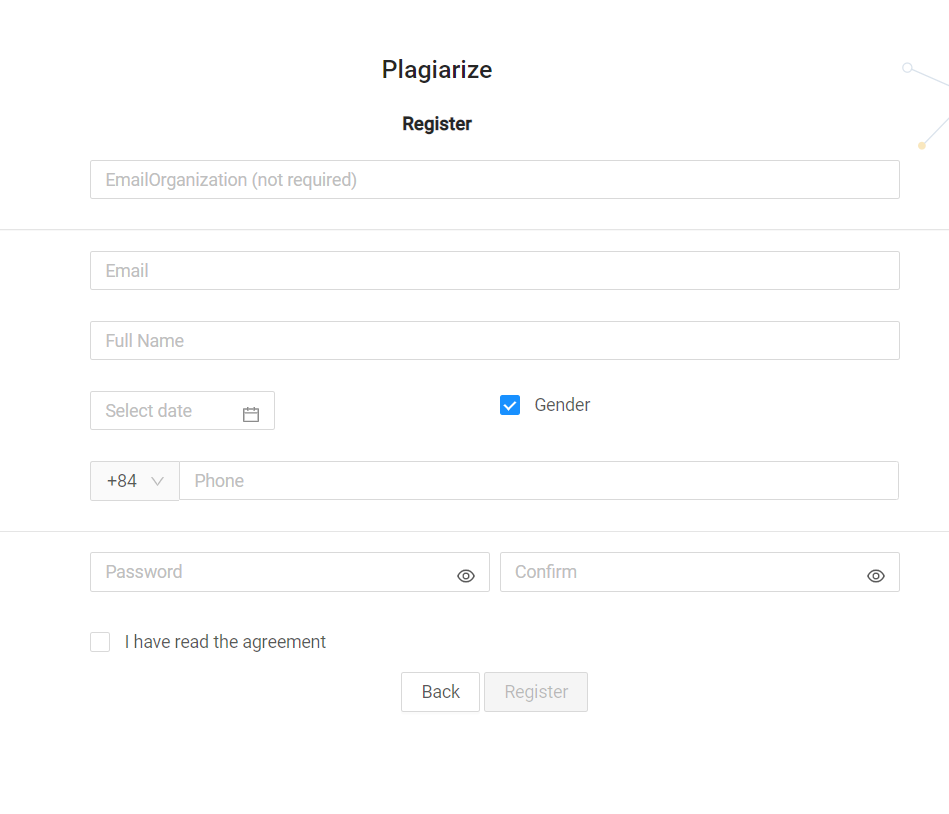
Chương này sẽ trình bày kết quả của các chức năng chính của website.

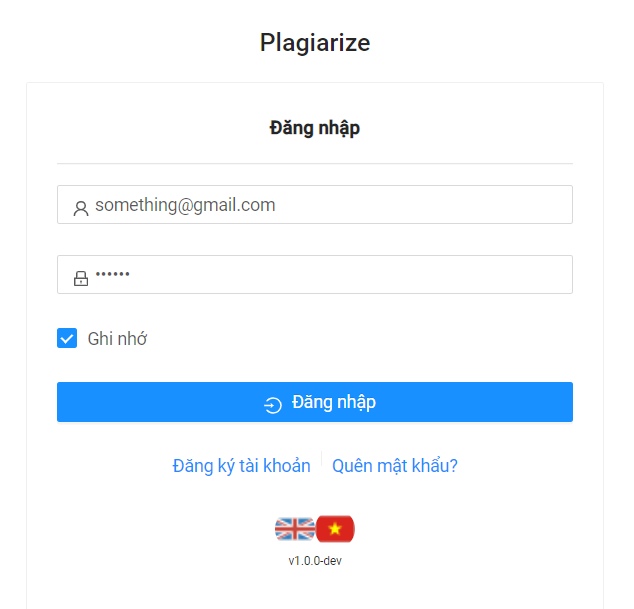
### Chức năng đăng ký và đăng nhập

Người dùng nếu chưa có tài khoản trước đó cần phải thực hiện đăng ký tài khoản mới (hình 6.1) bằng email. Khi đăng ký người dùng cung cấp một số thông tin cơ bản về cá nhân cũng như tổ chức mà người dùng đang làm việc (nếu có).

Khi đăng ký tài khoản thì sẽ nhận được một mã xác thực tài khoản qua email mà người dùng đăng ký. Sau khi nhập mã xác thực (hình 6.2) thì tài khoản mới có thể sử dụng được.

Tiếp theo người dùng có thể truy cập giao diện đăng nhập (hình 6.3) để tiến hành đăng nhập vào để sử dụng. Mặt khác nếu người dùng quên đi mật khẩu thì có thể thực hiện yêu cầu cấp lại mật khẩu thông qua email đã đăng ký. 





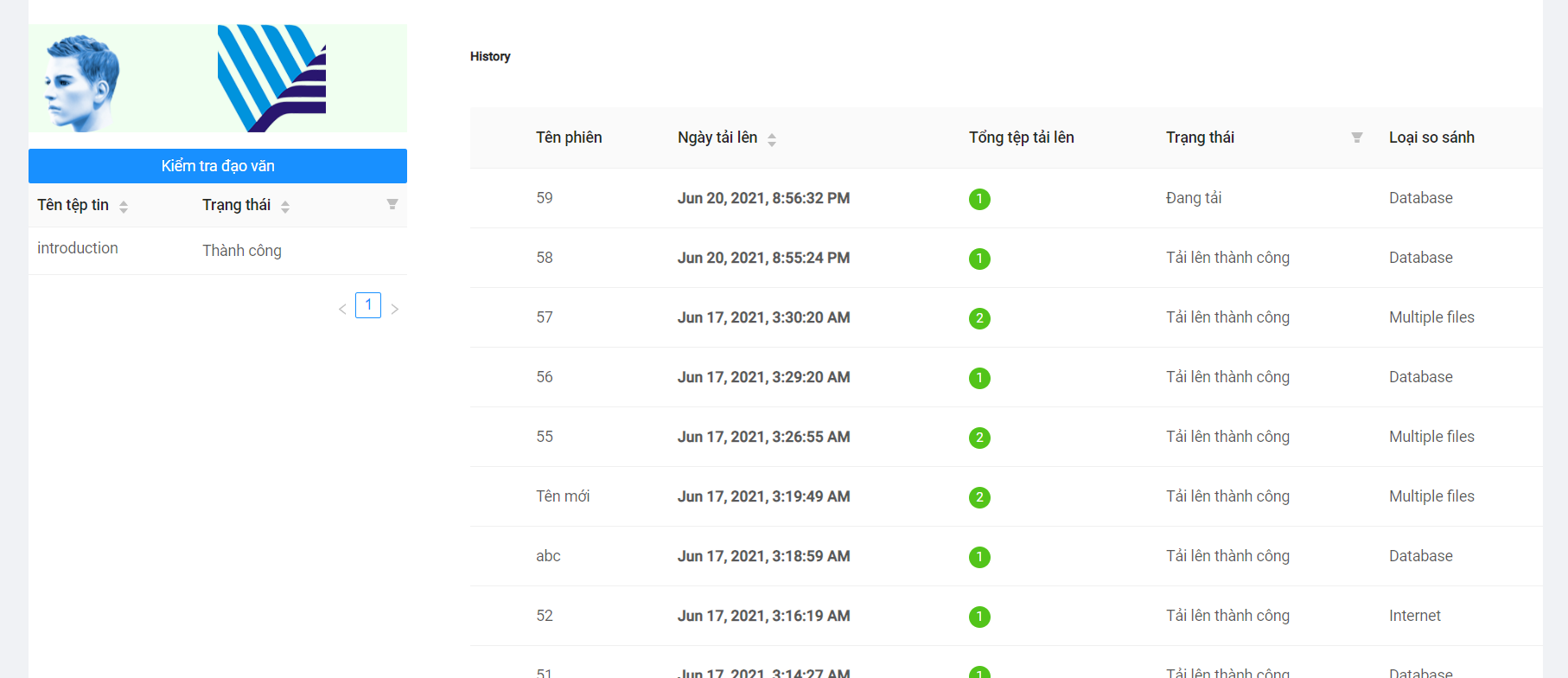
**Hình 6.3 Giao diện đăng nhập**

### Nhóm chức năng kiểm tra tương đồng văn bản

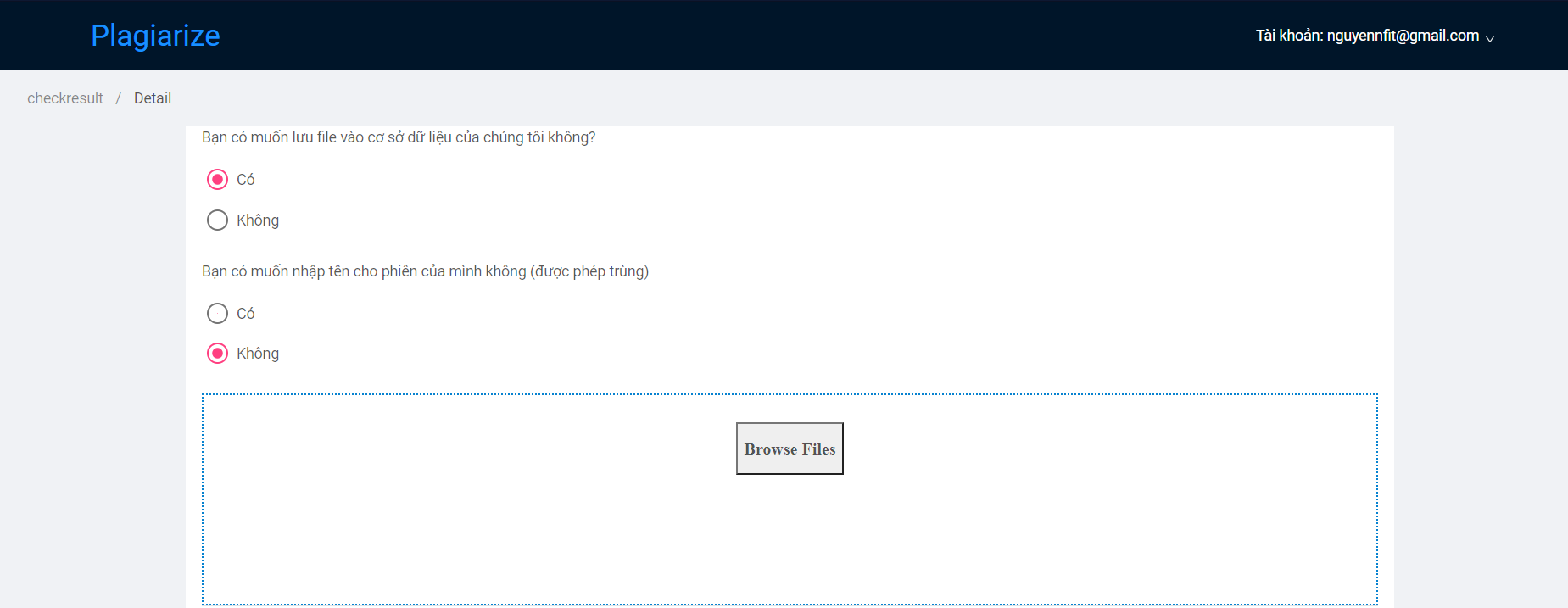
Người dùng tích vào “Kiểm tra đạo văn” trên màn hình chính (hình 6.4) và bắt đầu kiểm tra.

Sau đó tích chọn các lựa chọn của hệ thống:

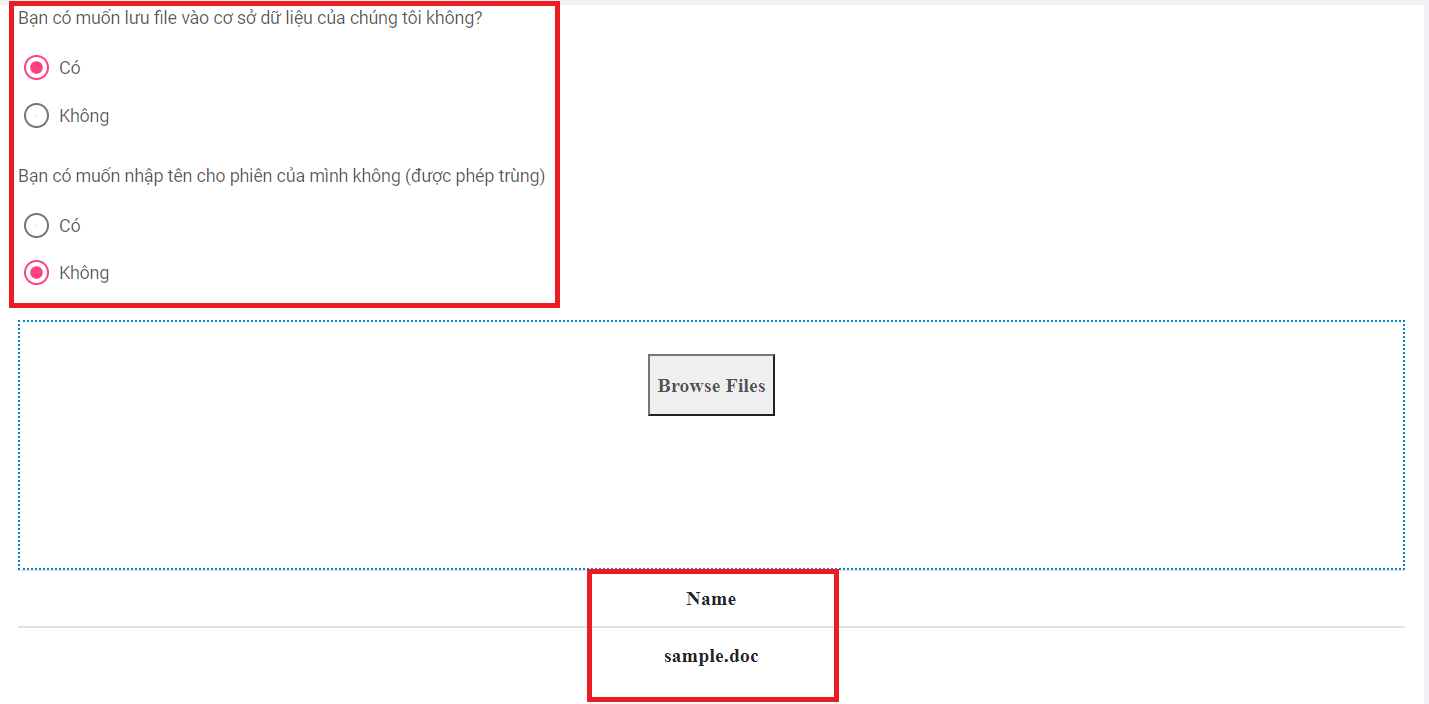
* Có cho phép hệ thống sử dụng dữ liệu của người dùng up lên làm kho ngữ liệu cục bộ hay không. Nếu có thì tất cả các file của người dùng up lên sẽ được lưu vào database của hệ thống để làm phong phú kho ngữ liệu của hệ thống giúp những lần kiểm tra sau đạt được độ chính xác cao hơn.
* Có muốn đặt tên cho phiên kiểm tra của mình hay không. Việc đặt tên giúp cho người dùng dễ dàng quản lý lịch sử kiểm tra của mình và dễ dàng truy xuất kết quả hơn.



**Hình 6.4 Giao diện chính của hệ thống**



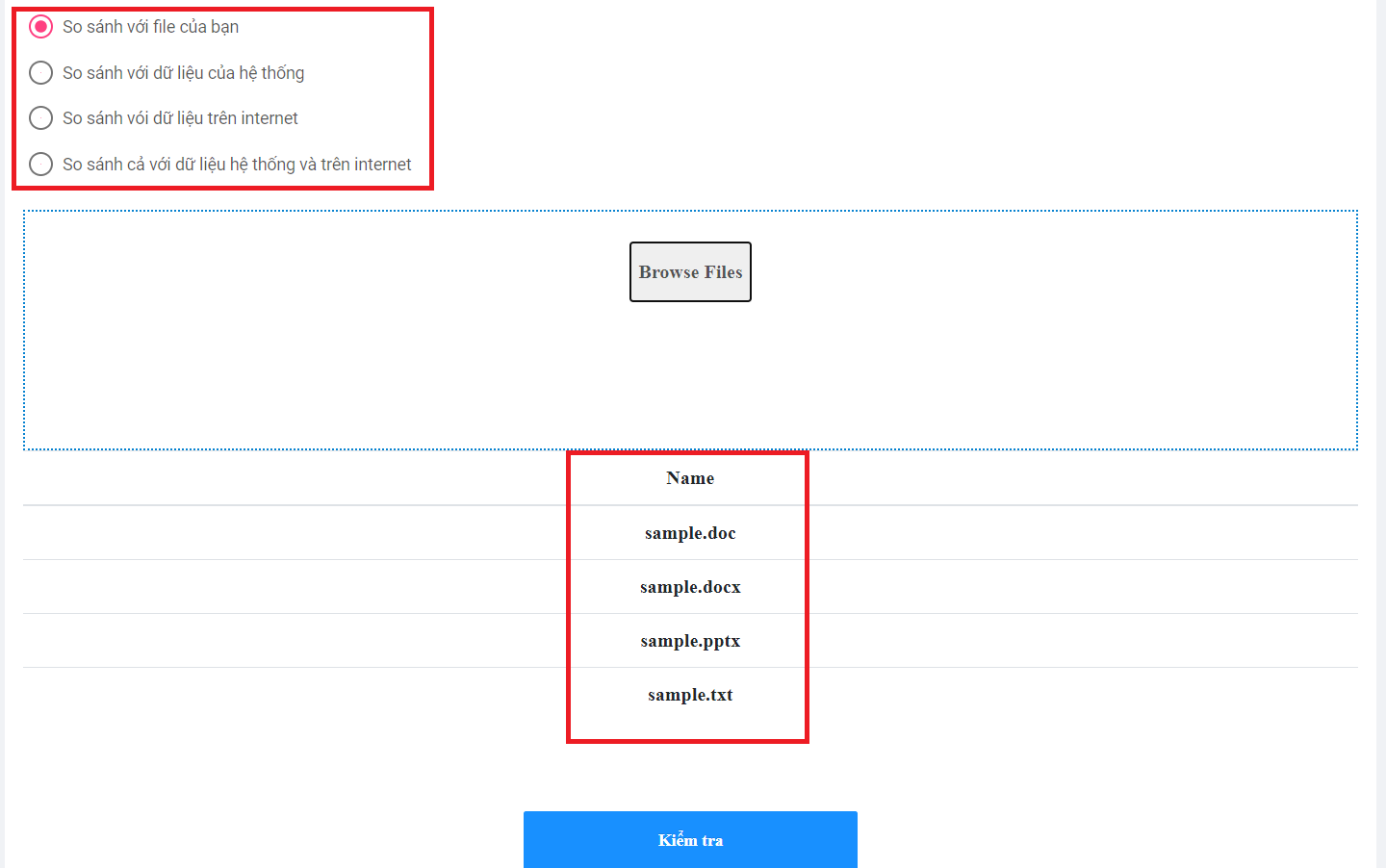
**Hình 6.5 Các tùy chọn của người dùng**



**Hình 6.6 Người dùng tải file cần kiểm tra lên hệ thống**

#### **Kiểm tra với dữ liệu của người dùng**

Người dùng up tệp tin cần kiểm tra lên theo định dạng cho phép và tích chọn vào lựa chọn “Kiểm tra với dữ liệu của bạn”. Sau đó người dùng up tiếp tục 1 loạt các tệp để làm kho ngữ liệu kiểm tra và cuối cùng ấn kiểm tra. Bạn sẽ được một thống báo quay về trang chủ và đợi kết quả kiểm tra.



**Hình 6.7 Người dùng up các file làm database để so sánh**

**Lưu ý**: Nếu phía trên người dùng tích vào có lưu vào dữ liệu hệ thống thì cả tệp để kiểm tra và các tệp để làm kho ngữ liệu được lưu vào database của hệ thống.

#### **Kiểm tra với dữ liệu của hệ thống chương trình**

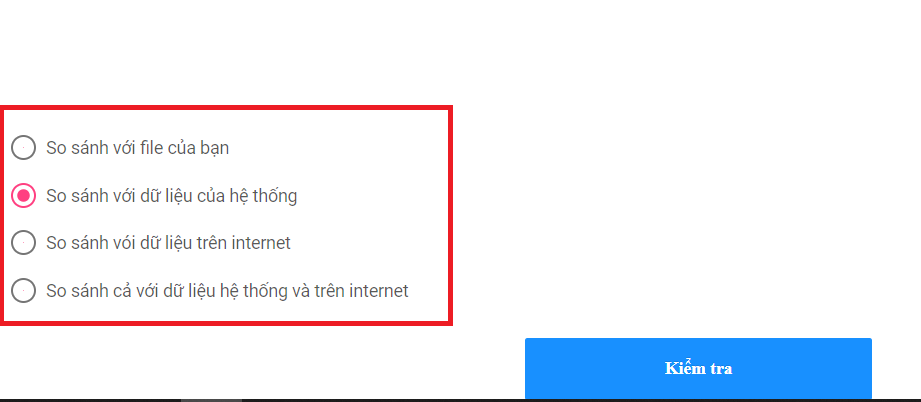
Người dùng up tệp tin cần kiểm tra lên theo định dạng cho phép và tích chọn vào lựa chọn “Kiểm tra với dữ liệu của hệ thống” và ấn kiểm tra. Bạn sẽ được một thống báo quay về trang chủ và đợi kết quả kiểm tra.

#### **Kiêm tra với dữ liệu tìm được từ internet**

Người dùng up tệp tin cần kiểm tra lên theo định dạng cho phép và tích chọn vào lựa chọn “Kiểm tra với dữ liệu trên internet” và ấn kiểm tra. Bạn sẽ được một thống báo quay về trang chủ và đợi kết quả kiểm tra.

#### **Kiểm tra với dữ liệu của cả hệ thống và internet**

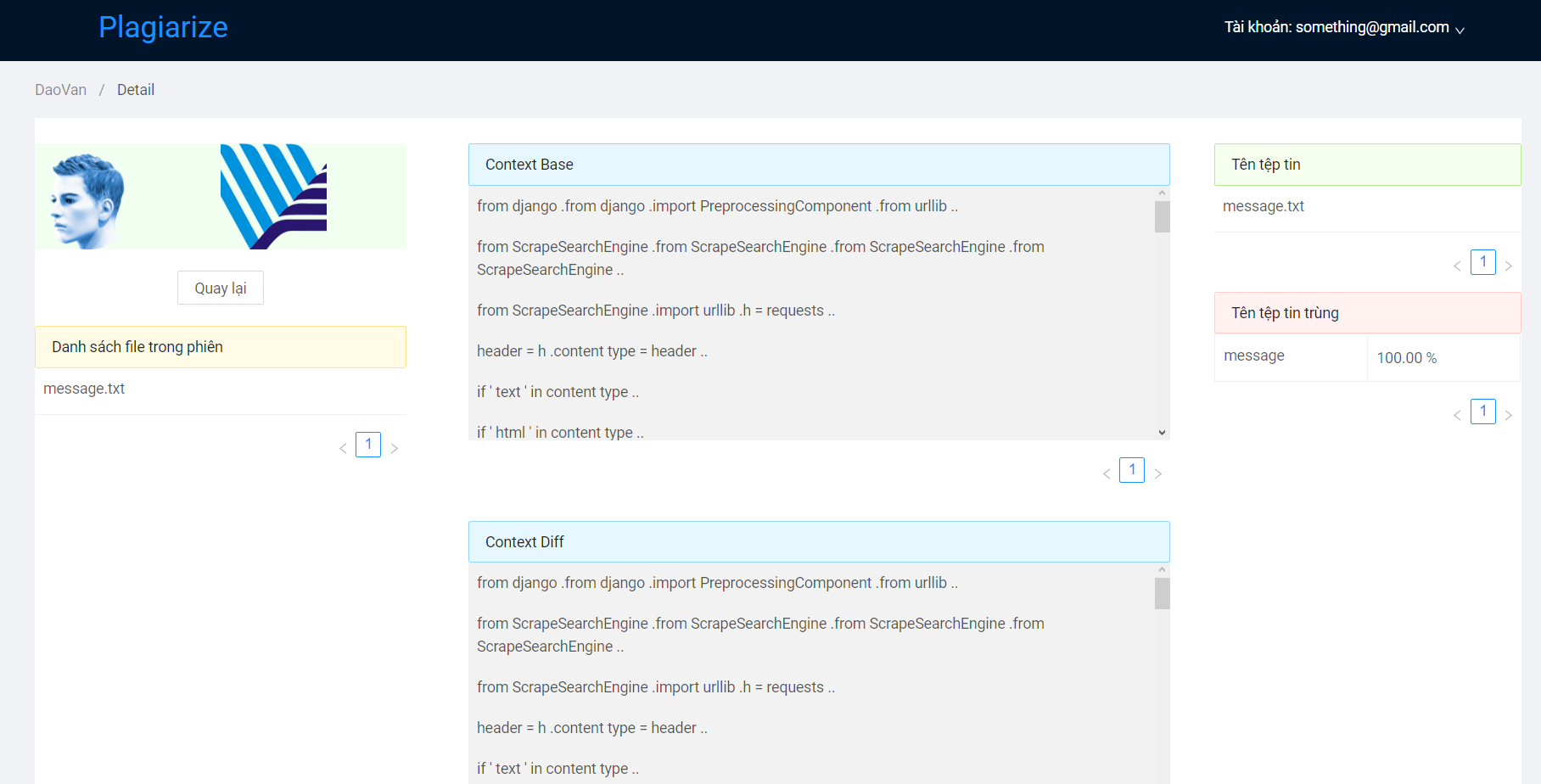
Người dùng up tệp tin cần kiểm tra lên theo định dạng cho phép và tích chọn vào lựa chọn “Kiểm tra với dữ liệu của hệ thống và internet” và ấn kiểm tra. Bạn sẽ được một thống báo quay về trang chủ và đợi kết quả kiểm tra.



**Hình 6.8 Tuỳ chọn cho các chức năng kiểm tra**

#### **Kiểm tra kết quả**

Để kiểm tra kết quả chung của các bước thì ở màn hình chính người dùng chọn vào phiên của mình và kiểm tra kết quả. Bên cạnh đó thì kết quả của mỗi lần thao tác đều được gửi vào email của người dùng tiện kiểm tra hơn.

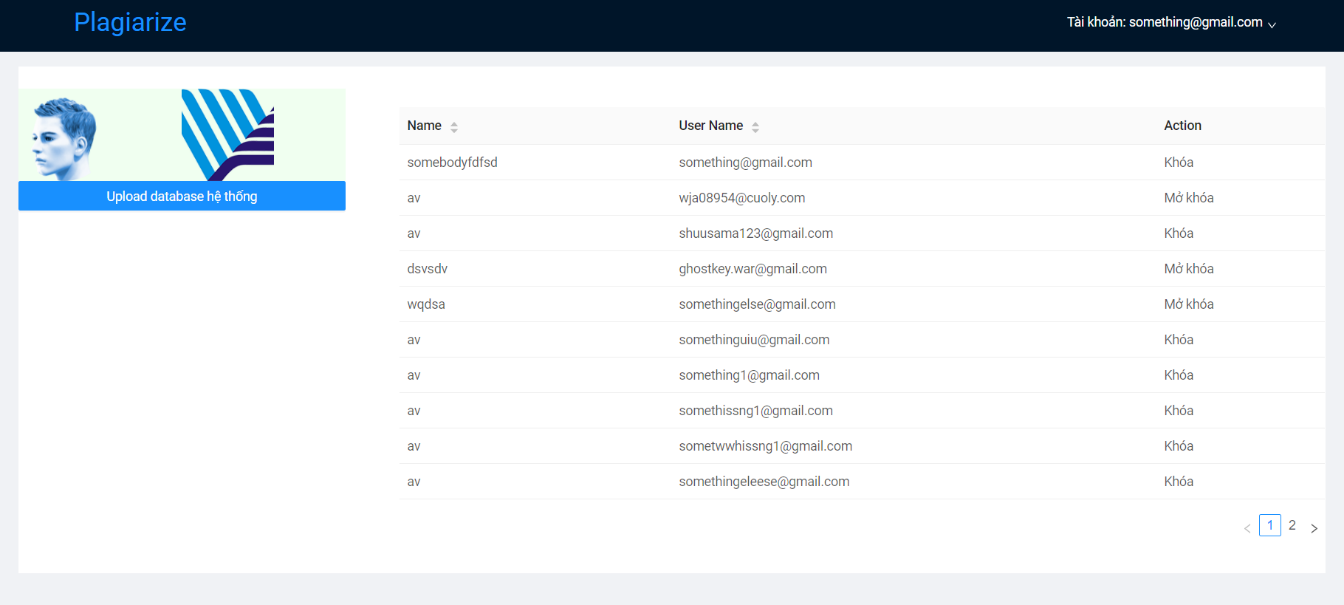


**Hình 6.9 Màn hình kết quả**

## Nhóm chức năng của admin

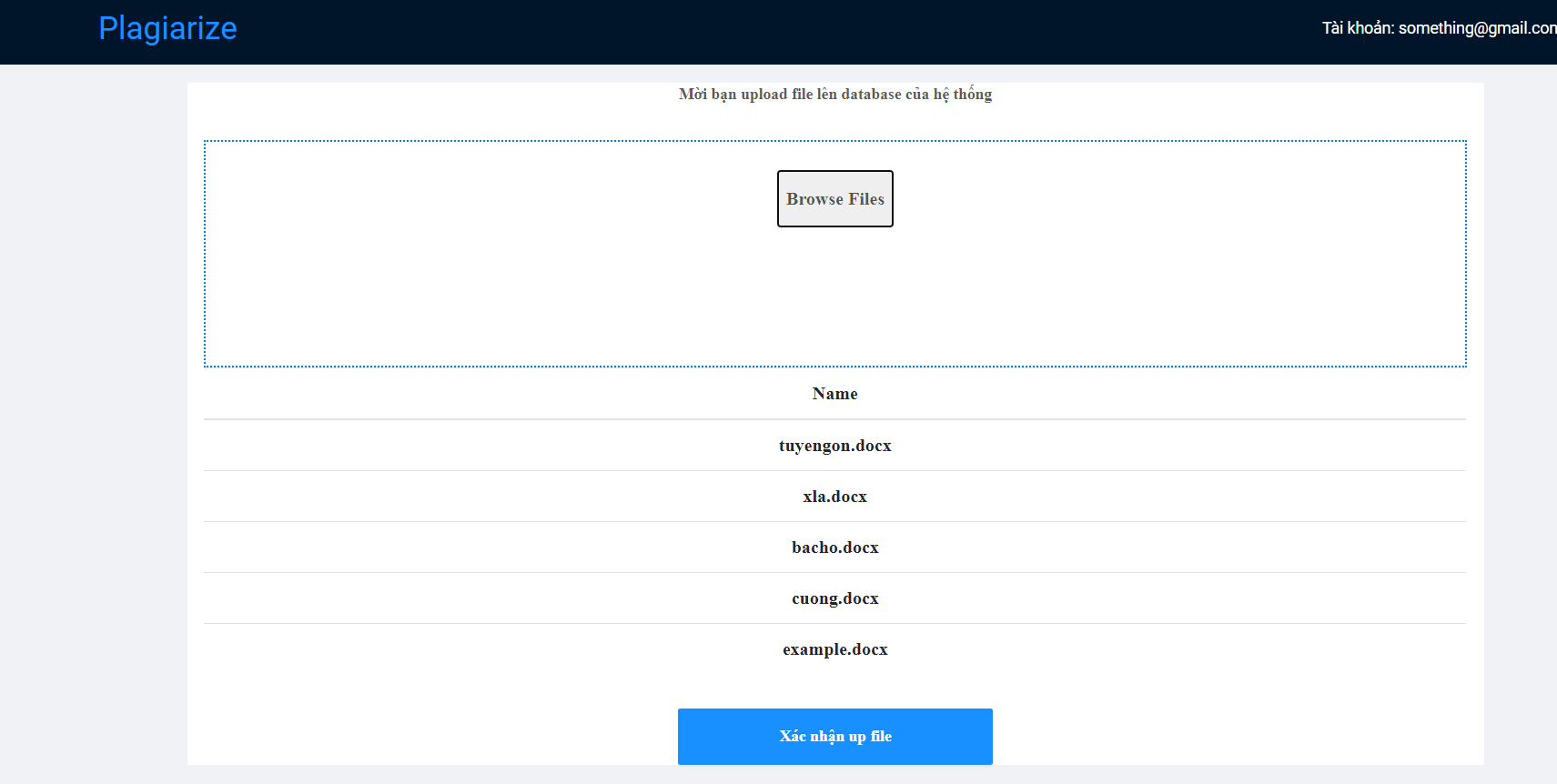
### Chức năng khóa tài khoản người dùng

Tại giao diện admin chọn khóa/ mở khóa một người dùng bất kỳ, ứng dụng sẽ thông báo thành công nêu tài khoản đó bị khóa/ được mở khóa hoặc thông báo thất bại nếu có lỗi xảy ra. (hình 6.10)



### Chức năng file lên database của hệ thống.

Chức năng này giống với việc người dùng up file lên để kiểm tra dữ liệu, nhưng đơn giản hơn do admin chỉ upload trực tiếp file vào database của mình chứ không cần so sánh. (Hnh 11)



**Hình 6.11 Giao diện cập nhật database của admin**

## Tổng kết chương 6

Qua chương này nhóm đã trình bày các kết quả mà nhóm đạt được sau quá trình thực hiện đồ án. Tổng hợp các chức năng và giới thiệu giao diện cũng như cách sử dụng của sản phẩm.

# Tổng kết và hướng phát triển

*Chương này gồm các thuận lợi, khó khăn mà nhóm đã gặp phải trong suốt quá trình thực hiện đồ án, cách làm việc của nhóm và các thiếu sót chủ quan cũng như khách quan của nhóm. Nêu ra các vấn đề còn tồn đọng của đồ án và đề xuất các hướng phát triển của sản phẩm*.

## Những thuận lợi, khó khăn gặp phải trong quá trình thực hiện và hướng khắc phục

### Thuận lợi

Có các cuộc họp hằng tuần để trao đổi, thảo luận, phân chia công việc hàng tuần giữa giáo viên hướng dẫn và các thành viên thực hiện.

Các thành viên nhóm trao đổi, thực hiện công việc thông qua các nền tảng ứng dụng trực tuyến hỗ trợ giao tiếp và xem source code nhanh chóng đối phó được với tình hình mùa dịch.

Thành viên nhóm có kiến thức về Python và các kỹ thuật liên quan giúp khắc phục các sự cố lỗi nhanh chóng.

### Khó khăn

Ngoài những thuận lợi thì cũng có các khó khăn mà nhóm gặp phải như:

Tiếp cận với nhiều nội dung mới, chưa có nhiều kiến thức liên quan nên việc nghiên cứu cũng như học hỏi tốn nhiều thời gian tìm hiểu, cũng như khắc phục sự cố.

Một số thư viện mới chưa biết sử dụng nên xuất hiện nhiều lỗi và cũng mất nhiều thời gian sửa lỗi.

### Hướng khắc phục

Để giải quyết các khó khăn gặp phải nhóm đã chọn các hướng giải quyết như sau:

* Dành thêm nhiều thời gian nghiên cứu và thử nghiệm các nội dung, thư viện cũng như tính năng để đảm bảo được kết quả tốt nhất.
* Trao đổi và thảo luận để tối ưu, cải thiện hệ thống tốt hơn. Ngoài ra cũng tham khảo từ phía giáo viên hướng dẫn để có ý kiến góp ý cho nội dung, tính năng trong đồ án.

## Hạn chế

Với những hạn chế về thời gian cũng như kiến thức chuyên môn, hệ thống cũng có các mặt hạn chế:

* Chưa hỗ trợ được người dùng quản lý các tập tin tải lên theo cây thư mục.
* Giới hạn người dùng so sánh một văn bản trong một phiên tải lên.
* Chưa cho phép người dùng lựa chọn độ tương đồng tùy ý.

## Hướng phát triển

Một ứng dụng hay hệ thống thì không thể hoàn thiện và đáp ứng được mọi yêu cầu của người dùng ngay từ bản đầu tiên. Từ những hạn chế đã phân tích, nhóm đã xác định nội dung phát triển và nâng cấp hệ thống trong tương lai như sau:

Hỗ trợ loại tài khoản tính phí tương ứng với các tính năng hỗ trợ tốt hơn, cho phép người dùng có thể kiểm tra nhiều văn bản cùng lúc.

Thêm các tính năng của admin có thể tải dữ liệu lên hệ thống trực tiếp không cần thông qua kiểm tra tương đồng.

Cho phép người dùng lựa chọn nhiều mức đánh giá kiểm tra tương đồng của văn bản.

Chức năng lựa chọn vùng văn bảng đa dạng để làm đối tượng so sánh.

## Quản lý, phân chia công việc

### Phương pháp quản lý công việc

Nhóm chúng phân chia nội dung công việc thực hiện theo các component riêng biệt của hệ thống. Sau đó sẽ tiến hành phân rã các component thành các công việc nhỏ hơn để xử lý và dễ theo dõi hơn.

Nhóm quản lý, trao đổi công việc của các thành viên với giáo viên hướng dẫn hằng tuần vào thứ 3 hoặc thứ 5 qua Zoom để nắm tiến độ như sau:

* Báo cáo công việc đã làm trong tuần qua.
* Các khó khăn gặp phải và hướng giải quyết.
* Các công việc cần làm trong tuần kế tiếp.

Ngoài ra nhóm có thảo luận, hỗ trợ các công việc hàng tuần ngay khi gặp vấn đề cũng như thắc mắc về nội dung mình làm thông qua Discord, nhanh chóng cải thiện được các sự cố.

Nhóm dùng công cụ Trello để quản lý công việc hàng tuần của các thành viên. Giúp nắm bắt được các nội dung tốt hơn.

Tại giai đoạn hoàn thành đồ án, nhóm đẩy nhanh tiến độ thực hiện việc trao đổi qua voice chat thường xuyên hơn thông qua Discord. Giúp khắc phục được nhiều sự cố, cũng như tìm thấy các lỗi ảnh hưởng đến hệ thống.

**7.4.2 Giới thiệu về Github**

Github là một dịch vụ lưu trữ trên web dành cho các dự án phát triển phần mềm. Nó cung cấp cả phiên bản trả tiền lẫn miễn phí cho người dùng. Ngoài ra còn có các chức năng social networking như feeds, followers như một mạng xã hội phổ biến. Các commit được lưu lại như hành động mà người dùng thực hiện trong repository của mình, đem lại sự tiện lợi cho việc quản lý source code cũng như tiến độ công việc của các thành viên khác trong cùng project.

Đây là công cụ hàng đầu không thể thiếu hiện nay đối với mọi developer trong việc quản lý dự án.

Github còn có tính năng thống kê được tích hợp sẵn, giúp developer dễ dàng theo dõi, thống kê được quá trình thực hiện project. Với mỗi một repository người dùng có thể tạo nhiều branch khác nhau cho từng thành viên. Từ các branch này có thể đối chiếu sự thay đổi trong mỗi commit riêng biệt trước khi merge vào branch chính.

Github hỗ trợ hầu hết mọi loại project được viết bởi các ngôn ngữ lập trình như C++, C#, java, Python, TypeScript, Ruby. Đối với các project mã nguồn mở sẽ được hỗ trợ lưu trữ miễn phí.

Vào mỗi tuần, sau khi thực hiện công việc được giao, từng thành viên trong nhóm sẽ tiến hành gửi commit những thay đổi của mình lên trên branch được chia ra sẵn. Các nội dung này sẽ được lưu trữ theo mốc thời gian trên Github giúp cho các thành viên theo dõi và kiểm tra dễ dàng hơn. Khi đã hoàn thành bản cuối của hệ thống, nhóm tiến hành merge các branch, chỉ để lại branch chính là bản hoàn thiện nhất.

## Tổng kết chương 7

Trong chương này, nhóm đã trình bày các vấn đề nảy sinh trong quá trình thực hiện đề tài. Các thuận lợi, khó khăn cũng như những trờ ngại vì sự thiếu sót về mặt chuyên môn lẫn kinh nghiệm của nhóm. Bên cạnh đó thì nhóm có đề xuất các hướing phát triển và hoàn thiện sẳn phẩm trong tương lai.

# Danh mục công trình của tác giả

# Phụ lục

8.1 Kết quả thực hiện các test case

8.1.1 Nhóm chức năng tạo tài khoản

8.1.2 Nhóm chức năng đăng nhập

8.1.3 Nhóm chức năng thay đổi thông tin tài khoản

8.1.4 Nhóm chức năng kiểm tra tương đồng

8.1.5 Nhóm chức năng xem nội dung báo cáo

8.2 Hướng dẫn sử dụng hệ thống

Các tính năng được hướng dẫn cụ thể kèm hình minh họa trong mục 6.2 Mô tả chức năng.

|  |  |
| --- | --- |
| link bỏ link vô đây | mục đánh số mục của mình tham khảo,chi tiết câu nào càng tốt 2light dùm t. |
| [tf–idf - Wikipedia](https://en.wikipedia.org/wiki/Tf%E2%80%93idf#:~:text=In%20information%20retrieval%2C%20tf%E2%80%93idf,in%20a%20collection%20or%20corpus.&text=tf%E2%80%93idf%20is%20one%20of,popular%20term%2Dweighting%20schemes%20today.) | 4.1.3.1 cuối đoạn đầu |
| [Conference about Indexes](https://laracon.eu/) | 5.1.2 hightlight màu vàng |
| [Django documentation.](https://docs.djangoproject.com) | 5.2.1 hightlight màu vàng |
| [MySQL Wikipedia](https://vi.wikipedia.org/wiki/MySQL) | 5.1.1 hightlight màu vàng |
| [Django MVC Structure](https://www.geeksforgeeks.org/django-project-mvt-structure/) | 5.2.2 hightlight màu vàng |
| [Github Wikipedia](https://vi.wikipedia.org/wiki/GitHub) | 7.4.2 hightlight màu vàng |

|  |  |
| --- | --- |
| Từ viết tắt | định nghĩa |
| jwt | json web token |
| DBMS | Database Management System |
| CSDL | cơ sở dữ liệu |
| DoIT | Document Improvement Tool |
| MVC | Model view control |
| ORM | Object Relational Mapping |
| HTTP | Hypertext Transfer Protocol |
| UI | User interface |
| HMAC | hash-basedmessage authentication code |
| RSA | Rivest–Shamir–Adleman |
| url | Uniform Resource Locator |
| TF-IDF | Term Frequency – Inverse Document Frequency |
| NLP | Natural Language Processing |