**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN - ĐHQG TPHCM**

**KHOA CÔNG NGHỆ PHẦN MỀM**

**ĐỒ ÁN MÔN HỌC CHUYÊN NGÀNH**

ĐỀ TÀI: HỆ THỐNG KHUYẾN NGHỊ VIỆC LÀM

*GVHD:*

*TS HUỲNH NGỌC TÍN*

*SINH VIÊN THỰC HIỆN:*

*12520248 - Trần Minh Luận*

*12520492 – Nguyễn Thanh Anh Tuyên*

MỤC LỤC

[**GIỚI THIỆU BÀI TOÁN** 1](#_Toc454631308)

[**Chương 1** **XÁC ĐỊNH VÀ ĐẶC TẢ YÊU CẦU** 2](#_Toc454631309)

[**1.** **DANH SÁCH CÁC YÊU CẦU** 2](#_Toc454631310)

[**1.1.** **Upload dataset** 2](#_Toc454631311)

[**1.2.** **Xem thống kê dữ liệu đầu vào:** 2](#_Toc454631312)

[**1.3.** **Cho phép chạy thực nghiệm các thuật toán khuyến nghị.** 4](#_Toc454631313)

[**1.4.** **So sánh các thuật toán khuyến nghị** 4](#_Toc454631314)

[**2.** **PHÂN TÍCH YÊU CẦU** 5](#_Toc454631315)

[2.1 Biểu đồ phân rã chức năng hệ thống 5](#_Toc454631316)

[2.2 Sơ đồ DFD mức 0 5](#_Toc454631317)

[2.3 Sơ đồ DFD mức 1 6](#_Toc454631318)

[**Chương 2** **THIẾT KẾ** 7](#_Toc454631319)

[**2.1** **KIẾN TRÚC HỆ THỐNG** 7](#_Toc454631320)

[2.1.1. Kiến trúc 7](#_Toc454631321)

[2.1.2. Mô tả kiến trúc 8](#_Toc454631322)

[**2.2** **THIẾT KẾ DỮ LIỆU** 8](#_Toc454631323)

[2.2.1. System DB 8](#_Toc454631324)

[2.2.2. User dataset 10](#_Toc454631325)

[**2.3** **THIẾT KẾ GIAO DIỆN** 11](#_Toc454631326)

[2.3.1. Trang đăng nhập 11](#_Toc454631327)

[2.3.2. Trang upload dataset 12](#_Toc454631328)

[2.3.3. Trang đăng ký tài khoản 15](#_Toc454631329)

[2.3.4. Trang thống kê dữ liệu 16](#_Toc454631330)

[2.3.5. Trang đăng nhập 16](#_Toc454631331)

[2.3.6. Trang đăng ký 16](#_Toc454631332)

[2.3.7. Trang quên mật khẩu 16](#_Toc454631333)

[**Chương 3** **LẬP TRÌNH VÀ KIỂM THỬ** 17](#_Toc454631334)

[**KẾT LUẬN** 18](#_Toc454631335)

# **GIỚI THIỆU HỆ THỐNG**

## Vấn đề

Có rất nhiều hướng giải bài toán khuyến nghị việc làm như sử dụng phương pháp collaborative filtering, sử dụng phương pháp content-based và rất nhiều hướng giải khác. Tuy nhiên làm sao để tìm được một hướng giải tốt là một vấn đề khó khăn. Vì vậy, hệ thống này được xây dựng nhằm cung cấp các đánh giá, so sánh kết quả thực nghiệm của những thuật toán khuyến nghị khác nhau. Từ đó chúng ta có thể có những quyết định phù hợp.

## Phát biểu bài toán

Hệ thống “Đánh giá các thuật toán khuyến nghị” (Recommender algorithm evaluation) là hệ thống cho phép người dùng đưa vào tập dữ liệu thực nghiệm. Sau đó tiến hành phân tích, thống kê dữ liệu đầu vào và chạy thực nghiệm các thuật toán khuyến nghị.

Hệ thống này phục vụ cho việc giải quyết bài toán khuyến nghị việc làm bằng cách cung cấp các đánh giá, so sánh các kết quả thực nghiệm. Từ đó tìm ra giải pháp hiệu quả để giải bài toán khuyến nghị việc làm.

* Đầu vào của hệ thống: Tập dữ liệu dataset.
* Đầu ra của hệ thống:
  + Biểu đồ thống kê dữ liệu.
  + Kết quả chạy của các thuật toán khuyến nghị dựa trên tập dataset đầu vào.
  + Các đánh giá so sánh kết quả thực nghiệm.

# **XÁC ĐỊNH VÀ ĐẶC TẢ YÊU CẦU**

## **Danh sách các yêu cầu**

Bảng danh sách các yêu cầu:

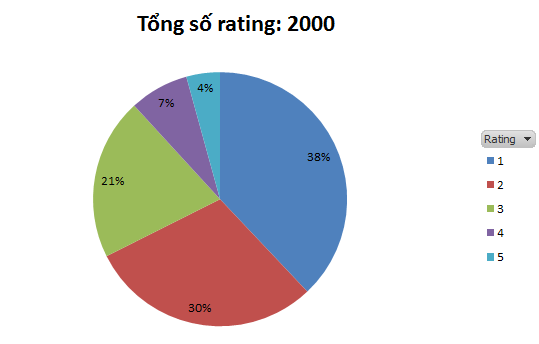
|  |  |  |
| --- | --- | --- |
| Mã Yêu cầu | Tên yêu cầu | Ghi chú |
| 1.1 | Upload dataset |  |
| 1.2 | Xem thống kê dữ liệu |  |
| 1.3 | Chạy thực nghiệm các thuật toán khuyến nghị |  |
| 1.4 | So sánh các thuật toán khuyến nghị |  |

### **Upload dataset**

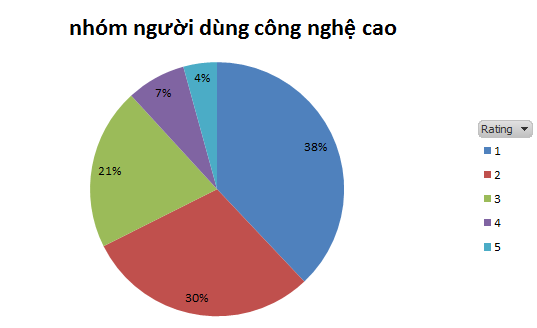
* Để bắt đầu sử dụng hệ thống. Người dùng phải cung cấp tên dataset cho hệ thống. Dataset được upload bao gồm 03 file.
  + File Cv.txt: Chứa dữ liệu về người dùng bao gồm các thông tin về cv người dùng.
  + File Job.txt: Chứa dữ liệu về việc làm sẽ được khuyến nghị.
  + Score.txt: Chứa dữ liệu về sở thích của người dùng trong file Cv.txt đối với những việc làm trong file Job.txt
* Sau khi chọn file người dùng chọn chức năng upload để bắt đầu upload dữ liệu. Hệ thống sẽ lưu trữ dataset vào thư mục có dạng: mã người dùng/tên dataset và thông báo kết quả thực hiện cho người dùng.

### **Xem thống kê dữ liệu đầu vào:**

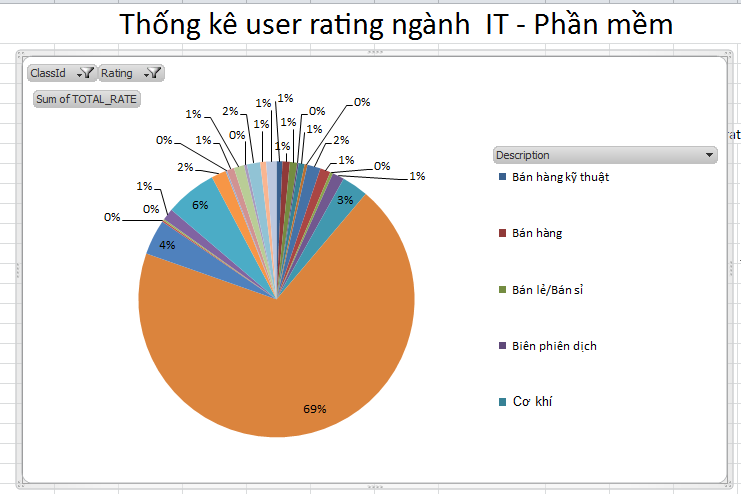
* Người dùng sẽ chọn tập dữ liệu cần phân tích để đưa vào module phân tích dữ liệu. Nếu chưa có thì người dùng có thể upload lên hệ thống. Sau khi người dùng chọn dữ liệu đầu vào. Người dùng chọn chức năng phân tích để bắt đầu phân tích dữ liệu. Sau khi hệ thống phân tích xong sẽ hiển thị ra màn hình cho người dùng.
* Các thống kê cần thực hiện.
  + Dựa vào file score.txt tiến hành thống kê tỉ lệ rating theo từng mức từ 1 – 5. Biểu đồ được vẽ là biểu đồ hình quạt có dạng bên dưới.



* + Thống kê người dùng theo từng ngành nghề(NND). Với mỗi NND thống kê tổng số rating mà thuộc NND đó. Bao gồm tỉ lệ % rating 1 điểm, tỉ lệ % rating 2 điểm. Tỉ lệ % rating 3 điểm. Tỉ lệ % rating 4 điểm và tỉ lệ % rating 5 điểm. Biểu đồ được vẽ như bên dưới.



Với mỗi nhóm rating như ở trên. Khi người dùng chọn một nhóm rating có giá trị x thì sẽ xuất thống kê tỉ lệ % các ngành mà NND đã cho x điểm. Biểu đồ được vẽ như hình bên dưới.



### **Cho phép chạy thực nghiệm các thuật toán khuyến nghị.**

* Người dùng tạo chọn dataset cần thực nghiệm, chọn thuật toán cần chạy. Nếu tập dữ liệu chưa có thì người dùng có thể upload lên hệ thống. Sau đó người dùng bấm chạy. Hệ thống sẽ thực hiện chạy thuật toán khuyến nghị. Sau khi chạy xong. Hệ thống sẽ thông báo cho người dùng kết quả thực nghiệm và lưu trữ lại vào file output.txt.
* Cấu trúc file output gồm 3 cột: UserId, JobId, Score. Các cột ngăn cách nhau bằng dấu kí tự tab.

### **So sánh các thuật toán khuyến nghị**

* Người dùng sẽ nhập những file output.txt sinh ra khi chạy thực nghiệm vào hệ thống. Hệ thống sẽ thực hiện tính toán và trả về kết quả kết quả so sánh các thuật toán khuyến nghị cho người dùng.

## **PHÂN TÍCH YÊU CẦU**

### Biểu đồ phân rã chức năng hệ thống

Recomender evaluation system

Upload dataset

View dataset Statistic

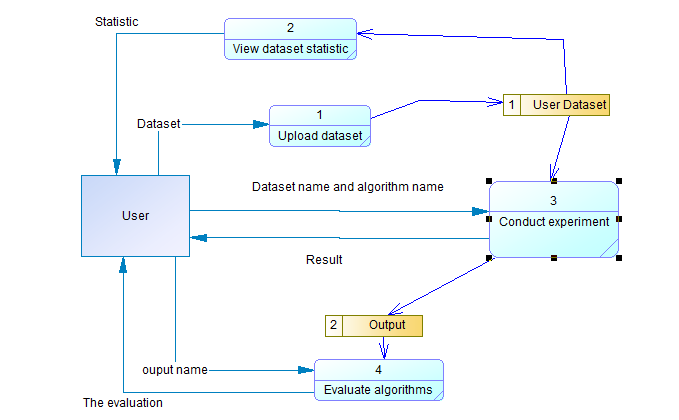
Conduct experiment

Evaluate algorithms

### Sơ đồ tổng thể hệ thống



### Sơ đồ DFD cấp 1 của hệ thống

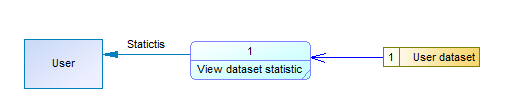


1. Mô tả process upload dataset.



* Mô tả: người dùng đưa dataset gồm 3 file job.txt, Cv.txt và Score.txt vào hệ thống. Hệ thống kiểm tra và lưu dataset lại.
* Các biểu mẫu liên quan:
  + File Job.txt: Định dạng file text, gồm các cột
  + File Cv.txt:
  + File Score.txt:

1. Mô tả process view dataset statistic.



* Mô tả: từ dataset mà người dùng đã upload. Hệ thống sẽ phân tích. Kết quả phân tích sẽ được lưu lại vào kho Data Statistic và trực quan hóa dữ liệu cho người dùng.

1. Mô tả process conduct experiment.



* Mô tả: Người dùng nhập tên của thuật toán cần chạy thực nghiệm và tên của dataset. Hệ thống sẽ lấy dataset tương ứng từ kho User Dataset và chạy thực nghiệm dựa thuật toán mà người dùng chọn.

1. Mô tả process evaluate algorithms



* Mô tả: người dùng đưa vào tên các kết quả thực nghiệm cần so sánh. Hệ thống sẽ lấy từ kho result set các kết quả thực nghiệm tương ứng. Sau đó hệ thống sẽ so sánh và trực quan hóa kết quả cho người dùng.

# **THIẾT KẾ HỆ THỐNG**

## **KIẾN TRÚC HỆ THỐNG**

### Kiến trúc tổng quát

**SYSTEM DB**

**User DataSet**

**Spring MVC**

**Apache Mahout**

**Apache Hadoop**

**RecAlg****EvaluationSystem**

**Dataset Analyzation**

**Experimental conduction & evaluation**

**System Interfaces**

**System core modules**

**System DB**

**Libraries**

### Mô tả các thành phần trong kiến trúc

* + System interfaces
* RecAlgComparisonSystem: Mô-đun này cung cấp cho người dùng chức năng quản lý dataset, xem thống kê dataset, quản lý các tiến trình thực thi các thuật toán khuyến nghị, xem kết quả và các đánh giá kết quả khuyến nghị.
  + System core modules
* Dataset analyzation: Đây là mô-đun phân tích dataset của người dùng.
* Experimental conduction & evaluation: chạy thực nghiệm các thuật toán khuyến nghị và đánh giá kết quả thực hiện.
  + Database
* System DB: database dùng để tổ chức lưu trữ thông tin người dùng, các tác vụ thực hiện của người dùng và các thông tin thiết lập chạy thuật toán.
* User dataset: hệ thống lưu trữ files dataset do người dùng upload cùng với đó là các file kết quả khuyến nghị, kết quả so sánh và kết quả thống kê.
  + Libraries
* Spring MVC: dùng hỗ trợ xây dựng website tương tác người dùng.
* Mahout: sử dụng các thư viện recommend có sẵn trong Mahout để thực hiện khuyến nghị.
* Hadoop: sử dụng để lưu trữ dữ liệu lớn, hỗ trợ cho việc phát triển trong tương lai.

### Kiến trúc triển khai

Hệ thống được thiết kế dựa vào mô hình Client-Server. Bởi vì tiến trình thực hiện các thuật toán khuyến nghị, tiến trình phân tích dữ liệu là các tiến trình tốn nhiều thời gian nên không phù hợp cho các ứng dụng web. Chính vì lí do này hệ thống được chia thành nhiều project con. Và các project con này sẽ gọi lẫn nhau và trao đổi thông tin cho nhau. Gồm các sub-project như sau.

Client computers

RecAlg

Comparison

System

Analyzation

Experimental

Conduction &

evaluation

Database

## **THIẾT KẾ DỮ LIỆU**

## System DB

Database này gồm có 2 bảng sau:

* USER (UserId, UserName, Email, Password)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Tên cột | Mô tả | Kiểu dữ liệu | Miền giá trị | Ghi chú |
| 1 | UserId | Mã người dùng | Integer | 11 | Not null, primary key |
| 2 | UserName | Tên người dùng | Nvarchar | 50 | Not null |
| 3 | Email | Email đăng nhập người dùng | Varchar | 50 | Not null |
| 4 | password | Mật khẩu đăng nhập | Varchar | 50 | Not null |

* Task (TaskId, UserId, TaskName, TimeCreate, Status, Algorithm, InputFolder, OutputFile)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Tên cột | Mô tả | Kiểu dữ liệu | Miền giá trị | Ghi chú |
| 1 | TaskId | Mã của task | Integer | 11 | Not null, primary key |
| 2 | UserId | Mã người dùng | Integer | 11 | Not null, foreign key đến User |
| 3 | TaskName | Tên task | Nvarchar | 100 | Not null |
| 4 | TimeCreate | Thời gian tạo task | Datetime |  | Not null |
| 5 | Status | Trạng thái task | Nvarchar | 10 | Not null |
| 6 | Algorithm | Tên thuật toán | Varchar | 20 | Not null |
| 7 | InputFolder | Đường dẫn đến thư mục chứa file input | Varchar | 100 | Not null |
| 8 | OutputFile | Đường dẫn file output | Varchar | 100 | Not null |

## User dataset

* Dữ liệu đầu vào để chạy các thuật toán gồm 3 file Job.txt, CV.txt và Score.txt. Quy ước trong các file .txt mỗi cột cách nhau bằng một ký tự tab.
* File Job.txt gồm các cột sau: JobId, JobName, Location, Salary, Category, Requirement, Tag, Description.
* File CV.txt gồm các cột sau: UserId, CVId, UserName, CVName, UserAddress, ExpectedSalary, Category, Language, Education, Skill, CareerObjective.
* File Score.txt gồm các cột sau: UserId, JobId, Score.
* Dữ liệu đầu ra

File output.txt có cấu trúc gồm 3 cột: UserId, JobId, Score.

## **THIẾT KẾ GIAO DIỆN**

### Giao diện đăng nhập

* Giao diện



* Mô tả giao diện:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu | Chức năng |
| 1 | Email | HTML Form Input(text) | Chứa địa chỉ email đăng nhập. |
| 2 | Mật khẩu | HTML form Input(password) | Chứa password của người dùng. |
| 3 | Đăng nhập | Html form Input(Đăng nhập) | Gửi dữ liệu lên trang xử lý |
| 4 | Lấy lại mật khẩu | HTML link | Chuyển sang trang lấy lại mật khẩu. |
| 5 | Đăng ký | HTML link | Chuyển sang trang đăng ký |

* Mô tả biến cố:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Biến cố | Xử lý | Mã số xử lý |
| 1 | Người dùng Bấm chọn nút đăng nhập | Kiểm tra email và mật khẩu đã được đăng ký hay chưa. Nếu chưa tồn tại trong hệ thống thì thông báo và yêu cầu nhập lại. Nếu đã tồn tại thì chuyển sang màn hình trang chủ. | XL1 |
| 2 | Người dùng bấm link lấy lại mật khẩu | Chuyển sang trang lấy lại mật khẩu | XL2 |
| 3 | Người dùng bấm link đăng ký | Chuyển sang trang đăng ký | XL3 |
| 4 | Load trang | Kiểm tra trạng thái đăng nhập của user. Nếu user đã đăng nhập thì chuyển về trang chủ. Ngược lại thì hiển thị trang đăng nhập | XL4 |

### Trang upload dataset

* Giao diện mẫu



* Mô tả giao diện

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu | Chức năng |
| 1 | Danh sách dataset | HTML TABLE | Thông tin danh sách các dataset đã upload |
| 2 | Xem thống kê | HTML link | Hiển thị đường dẫn tới trang thống kê dataset |
| 3 | Tên dataset | HTML form input(text) | Chứa tên dataset do người dùng nhập |
| 4 | Chọn File score.txt | HTML form upload field | Chứa đường dẫn tới file score.txt do người dùng chọn |
| 5 | Chọn file Cv.txt | HTML form upload field | Chứa đường dẫn đến file cv.txt do người dùng chọn |
| 6 | Chọn file Job.txt | HTML form upload | Chứa đường dẫn đến file job.txt do người dùng chọn |
| 7 | Upload | HTML form input(submit) | Nút dùng để kích hoạt chức năng upload. |

* Mô tả biến cố

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Biến cố | Xử lý | Mã xử lý |
| 1 | Load trang | * Lấy danh sách dataset mà người dùng đã upload hiển thị lên bảng 1. * Kiểm tra email và mật khẩu đã được đăng ký hay chưa. Nếu chưa tồn tại trong hệ thống thì thông báo và yêu cầu nhập lại. Nếu đã tồn tại thì chuyển sang màn hình trang chủ. | XL5 |
| 2 | Người dùng bấm xem thống kê | Chuyển đến trang thống kê dataset của dataset tương ứng | XL6 |
| 3 | Người dùng bấm upload | Kiểm tra thông tin người dùng đã nhập. Lưu trữ lại dataset và thông báo kết quả thực hiện cho người dùng. | XL7 |

### Giao diện đăng ký tài khoản

* Giao diện mẫu



* Mô tả giao diện

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu | Mô tả |
| 1 | Tên tài khoản | HTML form input(text) | Chứa tên tài khoản do người dùng nhập |
| 2 | Email | HTML form input(text) | Chứa email do người dùng nhập |
| 3 | Mật khẩu | HTML form input(password) | Chứa mật khẩu do người dùng nhập |
| 4 | Nhập lại mật khẩu | HTML form input(password) | Chứa mật khẩu do người dùng nhập |
| 5 | Đăng ký | HTML form input(submit) | Kích hoạt chức năng đăng ký tài khoản |
| 6 | Đăng nhập | HTML link | Quay về trang đăng nhập |

* Mô tả biến cố

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Biến cố | Xử lý | Mã xử lý |
| 1 | Load trang | Kiểm tra cookie trên máy người dùng. Nếu người dùng đã đăng nhập thì yêu cầu đăng xuất trước khi tiếp tục thực hiện | XL8 |
| 2 | Người dùng chọn đăng ký tài khoản | Kiểm tra các thông tin mà người dùng nhập vào hệ thống. Nếu còn thiếu hoặc sai thì yêu cầu người dùng nhập lại. Nếu thông tin hợp lệ thì chuyển thông tin về trang xử lý. | XL9 |
| 3 | Người dùng chọn đăng nhập | Quay trở về trang đăng nhập hệ thống | XL10 |

### Giao diện thống kê dữ liệu

### Giao diện chạy thực nghiệm các thuật thuật toán

* Giao diện mẫu:



* Mô tả giao diện:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu | Mô tả |
| 1 | Danh sách task | HTML table | Hiển thị danh sách các tác vụ đã được tạo và trạng thái. |
| 2 | Task | HTML link | Hiển thị liên kết đến trang kết quả thực nghiệm. |
| 3 | Tên task | HTML form input(text) | Chứa tên task người dùng nhập |
| 4 | Chọn thuật toán | HTML select | Dùng để lựa chọn thuật toán cần thực nghiệm |
| 5 | Chọn dataset | HTML select | Dùng để lựa chọn dataset cần thực nghiệm. |

* Các biến cố và xử lý:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Biến cố | Xử lý | Mã xử lý |
| 1 | Load trang | * Kiểm tra trạng thái đăng nhập của user. Nếu chưa đăng nhập thì yêu cầu đăng nhập. * Load danh sách các tác vụ mà người dùng đang đăng nhập hệ thống đã tạo. Tô màu xanh với các tác vụ đã chạy thành công. Tô màu vàng với các tác vụ đang chạy. Tô màu đỏ với các tác vụ bị lỗi. | XL11 |
| 2 | Người dùng bấm vào tên tác vụ. | Nếu tác vụ thực hiện thành công. Chuyển người dùng đến trang xem kết quả. Ngược lại không làm gì cả. | XL12 |
| 3 | Người dùng chọn tạo tác vụ | Kiểm tra tính hợp lệ dữ liệu mà người dùng nhập. Nếu có sai sót hoặc thiếu thì yêu cầu người dùng nhập lại. Nếu hợp lệ thì lưu trữ thông tin và bảng task và tiến hành thực nghiệm. | XL13 |

### Giao diện xem kết quả thực nghiệm

* Giao diện mẫu



* Mô tả giao diện

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu | Mô tả |
| 1 | Resulte | HTML link | Hiển thị link download file result |
| 2 | Bảng kết quả | HTML table | Hiển thị kết quả thực nghiệm |

* Mô tả các biến cố và xử lý

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Biến cố | Xử lý | Mã xử lý |
| 1 | Load trang | * Kiểm tra trạng thái đăng nhập của người dùng. Nếu chưa đăng nhập thì về trang đăng nhập * Lấy mã tác vụ cần xem và load kết quả thực hiện vào bảng 1 | XL14 |
| 2 | Người dùng chọn result | * Chuyển đến trang download dữ liệu | XL15 |

### Giao diện quên mật khẩu

# **LẬP TRÌNH VÀ KIỂM THỬ**

# **KẾT LUẬN**