**LỜI CẢM ƠN**

Trước hết, nhóm em xin bày tỏ tình cảm và lòng biết ơn của mình tới giảng viên Th.S Huỳnh Thái Học. Người đã hướng dẫn và giúp đỡ em trong quá trình thực hiện đồ án môn Phát triển ứng dụng của mình. Trong quá trình thực hiện đồ án của mình gặp không ít những khó khăn và thầy đã giúp em giải quyết những trở ngại đó.

Tuy có nhiều cố gắng trong quá trình học tập, và cũng như trong quá trình làm đồ án môn không thể tránh khỏi những thiết sót, nhóm em rất mong được sự góp ý quý báu của thầy cũng như tất cả các bạn để kết quả của em được hoàn thiện hơn.

Một lần nữa em xin chân thành cảm ơn thầy và các bạn.

TP.Hồ Chí Minh, Ngày 17 Tháng 12 Năm 2017

**MỤC LỤC**

[**1. Tổng quan:** 2](#_Toc501202376)

[1.1.Tên đề tài và nhiệm vụ: 2](#_Toc501202377)

[1.2. Tiến độ thực hiện và chi tiết: 3](#_Toc501202378)

[**2. Nội dung chi tiết đồ án phát triển ứng dụng** 4](#_Toc501202379)

[2.1. Kiến trúc của một website: 4](#_Toc501202380)

[2.2. Mô hình Client-Server: 4](#_Toc501202381)

[2.3. Chi tiết công việc thực hiện: 5](#_Toc501202382)

[a) Trích xuất information (comment): 5](#_Toc501202383)

[b) Trích xuất information (ảnh): 6](#_Toc501202384)

[c) Trích xuất information (link - liên kết): 7](#_Toc501202385)

[d) Giao diện đồ họa - GUI(Graphical user interface): 7](#_Toc501202386)

[e) Trích xuất information (email): 8](#_Toc501202387)

[**3. Kết luận đồ án:** 8](#_Toc501202388)

**BÁO CÁO ĐỒ ÁN PHÁT TRIỂN ỨNG DỤNG**

Họ và tên: Đặng Võ Đức Trọng MSSV: 15055721

Họ và tên: Nguyễn Chí Thắng MSSV: 15055101

Giáo viên HD: Huỳnh Thái Học

Link github: https://github.com/AjackTi/DAHP1

# **1. Tổng quan:**

## 1.1. Tên đề tài và nhiệm vụ:

\* Đề tài: Extract information from website by Regular Expression (Regex)

\* Nhiệm vụ:

* Tìm hiểu về mô hình Client-Server.
* Tìm hiểu về cách trích xuất nội dung và các tag trong 1 website bất kỳ nào đó (get HTML).
* Match các tag của một website để lấy các thông tin cần thiết.
* Tạo một giao diện đơn giản chứa các phần từ như label, textbox, button, …
* Thiết kế giao diện, kết nối các class chi tiết và lớp GUI.

## 1.2. Tiến độ thực hiện và chi tiết:

|  |  |  |
| --- | --- | --- |
| **Thời gian** | **Thực hiện** | **Tiến độ công việc thực hiện** |
| Buổi 1 - Thứ 3 (22/08/2017) | Đặng Võ Đức Trọng (Tìm hiểu về Java cơ bản).  Nguyễn Chí Thắng (tìm hiểu lại kiến thức Java cơ bản). | Mục tiêu: Tìm hiểu về Java cơ bản  Hoàn thành: Đã hoàn thành mục tiêu. |
| Buổi 2 - Thứ 3 (29/08/2017) | Đặng Võ Đức Trọng  Nguyễn Chí Thắng | Mục tiêu: Tìm hiểu về Java cơ bản  Hoàn thành: Đã hoàn thành mục tiêu. |
| Buổi 3 - Thứ 3 (5/09/2017) | Đặng Võ Đức Trọng (Tìm hiểu về Regular Expression.).  Nguyễn Chí Thắng (tìm và đọc một số tài liệu về Regex  Tài liệu: <http://o7planning.org/vi/10175/huong-dan-su-dung-bieu-thuc-chinh-quy-trong-java>). | Mục tiêu: Tìm hiểu về Regular Expression.  Match thử một vài bài tập(VD: email, số điện thoại, …)  Hoàn thành: Đã hoàn thành.  Chưa hoàn thành: |
| Buổi 4 – Thứ 3 (12/09/2017) | Đặng Võ Đức Trọng (Thiết lập một server IIS, Lấy thông tin từ URL).  Nguyễn Chí Thắng (trích xuất hình ảnh, link và e-mail) | Mục tiêu:  + Lấy thông tin từ URL sử dụng class openconnection(), get data từ class này.  + (phụ)Thiết lập một server IIS trên máy tính(<http://localhost/>).  Hoàn thành: Đã thiết lập thành công localhost trên máy tính sử dụng source code html của trang muốn trích xuất (<http://localhost/index/>).  Chưa hoàn thành: Thực hiện thành công sử dụng class openconnection() nhưng không thể lấy được tất cả các thẻ có chứa comment. |
| Buổi 5 – Thứ 3 (19/09/2017) | Đặng Võ Đức Trọng(Tìm hiểu về giao diện trong Java, Thiết kế 1 giao diện).  Nguyễn Chí Thắng (tìm hiểu cách tạo giao diện trong Java). | Mục tiêu:  + Tìm hiểu về giao diện trong Java  + Thiết kế 1 giao diện đơn giản bao gồm 2 text box(1 chứa input và 1 chứa result), 1 button, …  Hoàn thành: Đã hoàn thành  Chưa hoàn thành: (Nguyễn Chí Thắng) chưa tạo được giao diện. |
| Buổi 6 – Thứ 3 (26/09/2017) | Đặng Võ Đức Trọng(Thiết kế giao diện + xử lý các phần tử trên giao diện).  Nguyễn Chí Thắng (tiếp tục tìm hiểu cách tạo giao diện). | Mục tiêu: Thiết kế giao diện + xử lý các phần tử trên giao diện(click button, chỉnh sửa vị trí, chỉnh sửa kích thước, …)  Hoàn thành:  Hoàn thành phần thiết kế.  Chưa hoàn thành:  Lỗi ngay lúc xử lý sự kiện của các phẩn tử trên GUI.  (Nguyễn Chí Thắng) vẫn chưa tạo được giao diện. |

# **2. Nội dung chi tiết đồ án phát triển ứng dụng**

## 2.1. Kiến trúc của một website:

Bao gồm tất cả các thẻ <head></head>, <body></body>, <title></title>, …

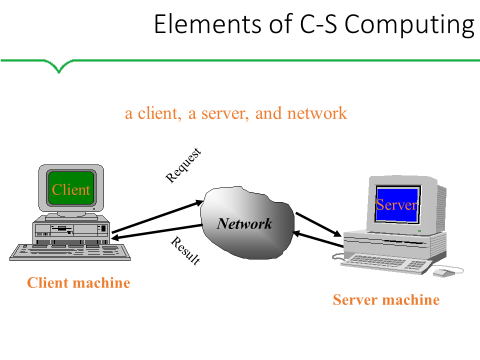
Comment (cần lấy) ở trong thẻ <div class="comment-list" style="display:block">(thẻ cha chứa tất cả các thẻ khác chứa comment).

## 2.2. Mô hình Client-Server:

A simple definition of CS is “server software accepts requests for data from client software and returns the results to the client” – (*Rajkumar Buyya -* School of Computer Science & Software Engineering, Monash University Melbourne, Australia)

Trong mô hình này, một hoặc vài máy sẽ được chọn để đảm nhận việc quản lý và cung cấp tài nguyên (chương trình, dữ liệu, thiết bị,…) được gọi là máy chủ (Server), các máy khác sử dụng tài nguyên này được gọi là máy khách (Clients).

Máy chủ là máy tính đảm bảo việc phục vụ các máy khách bằng cách điều khiển việc phân phối tài nguyên nằm trong mạng với mục đích sử dụng chung. Máy khách là máy sử dụng tài nguyên do máy chủ cung cấp.

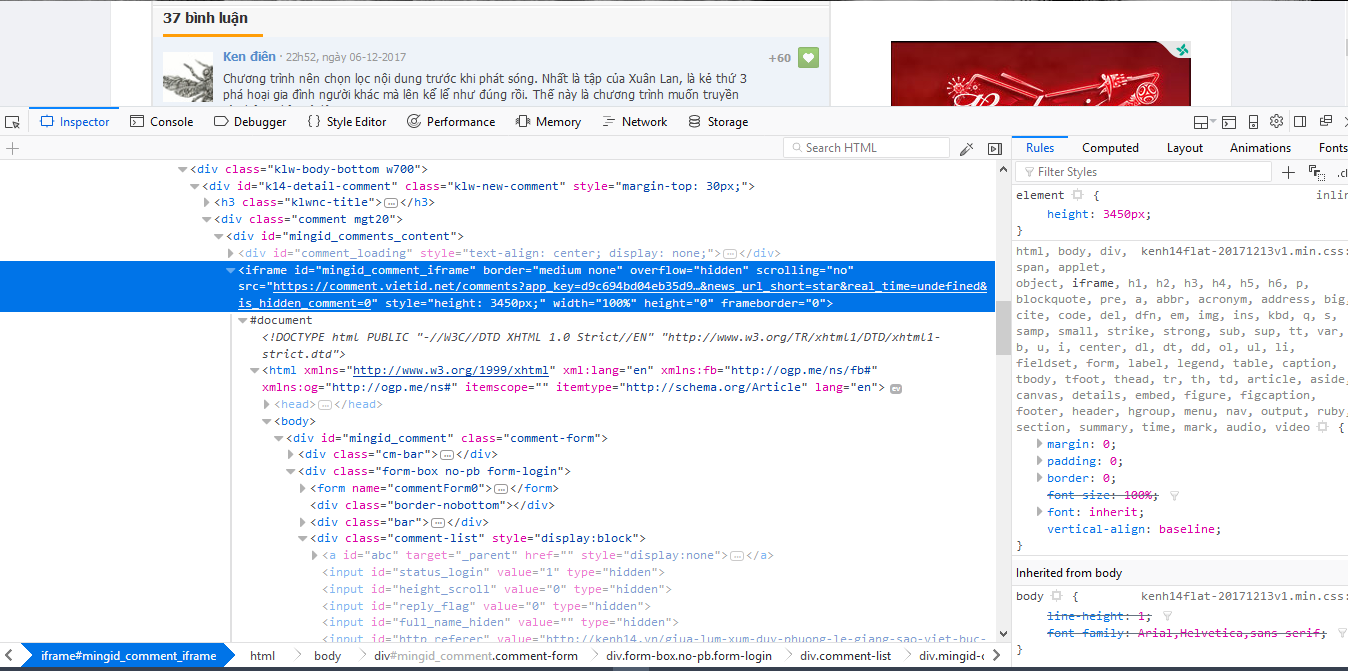
Mô hình khách – chủ có ưu điểm là dữ liệu được quản lý tập trung, bảo mật tốt, thích hợp với các mạng trung bình và lớn.

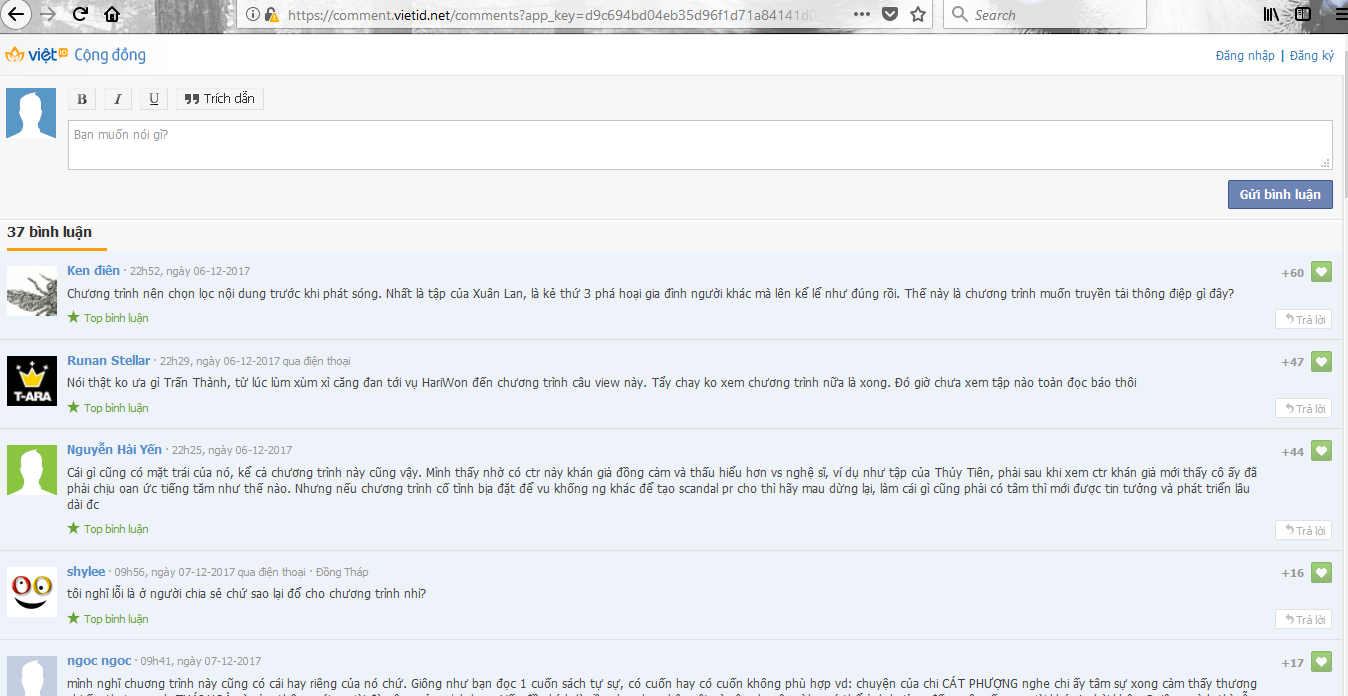
Trích từ: chuong 2\_client\_server.ppt

## 2.3. Chi tiết công việc thực hiện:

### a) Trích xuất information (comment):

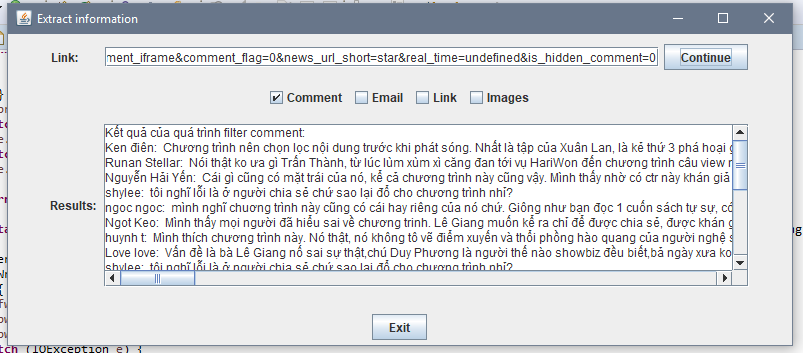
Ta thấy rằng website (kenh14.vn) mà ta muốn lấy comment ra thực sự là một trang khác chứa các comment này, chỉ là tương tác trên kenh14.vn chứ không phải là do kenh14.vn quản lý.





Như thế ta sẽ lấy comment từ trang phụ này, cũng chứa các comment tương tự như trong kenh14.vn.

Kết quả ta sẽ được:



Ở đây ta có thể lấy được top bình luận cũng như các bình luận thông thường cùng với là tên id của người bình luận.

*Khó khăn gặp phải:*

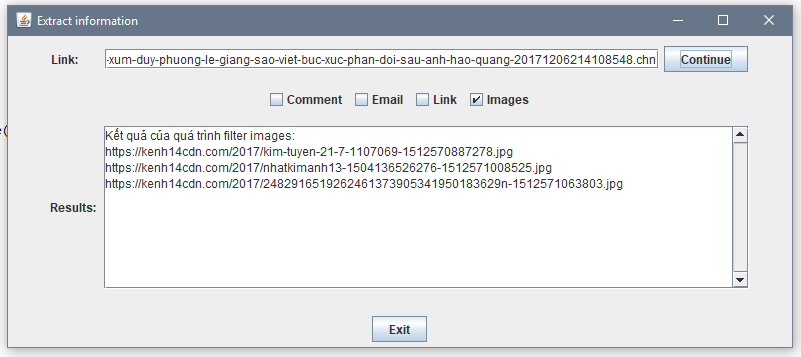
Chưa lấy trực tiếp từ trang chính được, mà phải thông qua 1 trang phụ.

### b) Trích xuất information (ảnh):

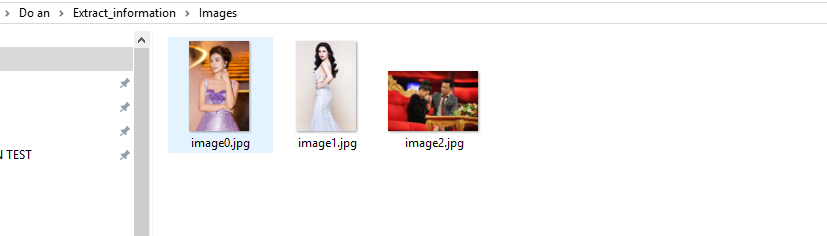
Khác với lấy comment, ta có thể lấy trực tiếp từ trang chính hoặc thông qua trang phụ.

Ở đây, nhóm lấy email thông qua Regular Expression, còn lấy ảnh thông qua thẻ <img src=””>

Kết quả ta sẽ được:



Và sau đó chương trình sẽ tự động download những ảnh lấy được vào trong thư mục Images.



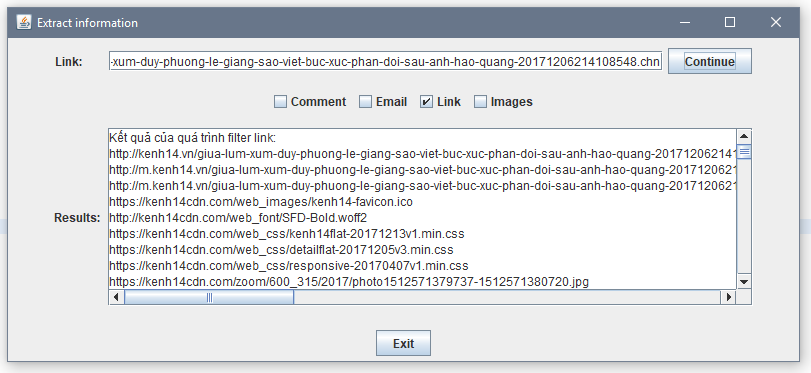
*Khó khăn gặp phải:*

+ Chỉ lấy được một số hình,… vì nó cũng như lấy comment – không lấy được tất cả các tag của trang (ví dụ như có các tag <img > ở tag con cuối cùng thì ta cũng không thể lấy được).

### c) Trích xuất information (link - liên kết):

Nhóm dựa vào thẻ chứa href=”” để lấy ra các link, có thể những link không đầy đủ nhưng khi ta click vào nó thì nó vẫn có thể hoạt động được.

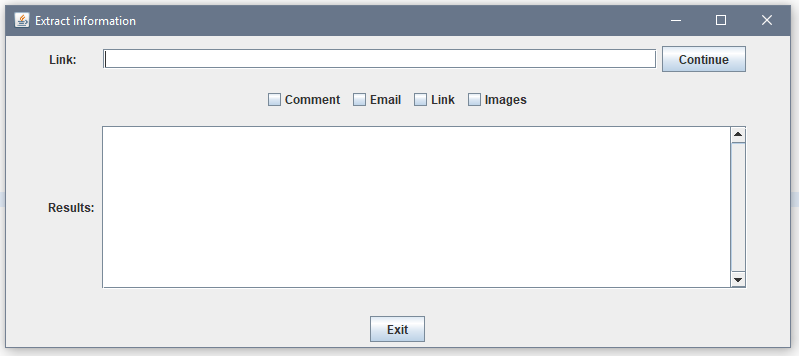
Kết quả đạt được:



Có thể thấy rằng hầu như các link được lọc ra.

### d) Giao diện đồ họa – GUI ([Graphical user interface](https://en.wikipedia.org/wiki/Graphical_user_interface)):

Kết quả đạt được:



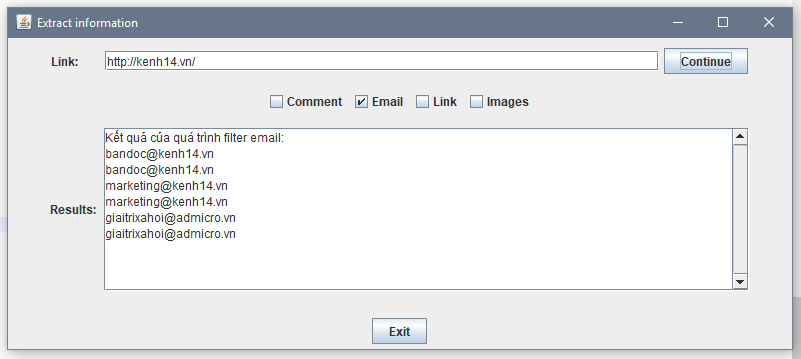
*Khó khăn gặp phải:*

Một số thành phần của giao diện chưa được linh hoạt.

### e) Trích xuất information (email):

Kết quả đạt được:

Dựa vào Regular Expression của email, nhóm đã trích xuất tất cả các email của trang.



# **3. Kết luận đồ án:**

Đạt được:

+ Trích xuất được information từ website (nhưng vẫn chưa lấy được comment trong trang đó).

+ Nếu lấy một thẻ có chứa comment (kenh14.vn) thì vẫn lấy ra được comment trong thẻ đó. (sử dụng Regex).

+ Chưa đúng tiến độ (vì theo như tiến độ thì buổi 6 nhóm phải hoàn thành 1 giao diện và các sự kiện liên quan với các element trên nó, nhưng nhóm vẫn chưa hoàn thành, cụ thể là bị lỗi lúc xử lý sự kiện trên các element này)

KL: Công việc được khoảng 90%.

*Link tài liệu:*

<https://www.regular-expressions.info/>

<https://stackoverflow.com/questions/11087163/how-to-get-url-html-contents-to-string-in-java>

<https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxodXluaGhvY3xneDo0NmRhMGZiMjhmNjQ4MjM1>

<https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxodXluaGhvY3xneDo3MmUyYjQxMmYwYmZlYTM>

https://stackoverflow.com/questions/12618245/error-catching-exception-while-system-out-print