TRƯỜNG ĐẠI HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN THỰC TẬP 1**

NGÀNH: CÔNG NGHỆ THÔNG TIN

**THU THẬP VÀ LÀM SẠCH DỮ LIỆU**

Data Gathering and Cleansing

**GVHD:**

**Trần Kim Mỹ Vân – Phan Hồ Viết Trường**

**Mentor: Đặng Như Khánh Hoàng**

TP. Hồ Chí Minh – năm 2020

Nhận xét của giảng viên và mentor của công ty DMSpro:

Danh sách thành viên

|  |  |
| --- | --- |
| MSSV | Họ tên |
| 187IT23616 | Cao Khánh Việt |
| 187IT14070 | Nguyễn Nhật Huy |
| 187IT14067 | Hồ Quốc Hùng |
| 187IT14104 | Trần Mộng Nghi |

Mục lục

[I. Giới thiệu và nội dung tổng quát của báo cáo: 4](#_Toc59476875)

[Mở đầu 4](#_Toc59476876)

[Tổng quát về nội dung đồ án: 4](#_Toc59476877)

[II. Về công ty DMSpro: 5](#_Toc59476878)

[Giới thiệu: 5](#_Toc59476879)

[Các sản phẩm, giải pháp tiêu biểu: 5](#_Toc59476880)

[Danh hiệu , thành tựu: 6](#_Toc59476881)

[III. Thu thập và làm sạch dữ liệu (Data Gathering and Cleansing): 7](#_Toc59476882)

[Phân tích yêu cầu người dùng: 8](#_Toc59476883)

[Lấy dữ liệu (chủ yếu là các trang HTML, php): 8](#_Toc59476884)

[Trích lọc thông tin cần thiết và lưu trữ dữ liệu đã cấu trúc: 8](#_Toc59476885)

[Lưu trữ dữ liệu trên các nền tảng lưu trữ cho công việc nhóm sau: 8](#_Toc59476886)

[IV. Các công cụ được sử dụng: 9](#_Toc59476887)

[1. Sprint 0: Tìm tài liệu về các công cụ hỗ trợ trong đồ án: 12](#_Toc59476888)

[2. Sprint 1: Tìm hiểu yêu cầu người dùng: 13](#_Toc59476889)

[3. Sprint 2: Dùng các phần mềm bên thứ 3 cung cấp để tìm hiểu về việc thu thập dữ liệu: 15](#_Toc59476890)

[4. Sprint 3: Thiết kế đường dẫn dữ liệu và tiến hành cào các trang web: 16](#_Toc59476891)

[5.Sprint 4: Dùng Python để cào dữ liệu từ các trang web tuyển dụng: 17](#_Toc59476892)

[6. Sprint 5: Hoàn thành, tối ưu hóa các code để cào và thu thập dữ liệu, viết báo cáo cho đồ án. 18](#_Toc59476893)

[7. Sprint 6: Chuẩn bị báo cáo cho đồ án: 19](#_Toc59476894)

[V. Kết quả thực tập: 19](#_Toc59476895)

[Toàn bộ tài liệu đề cập trong báo cáo 21](#_Toc59476896)

# Giới thiệu và nội dung tổng quát của báo cáo:

# Mở đầu

* Big data đang là một cuộc cạnh tranh lớn trong thời đại 4.0 hiện nay. Nó là một nguồn thông tin vô cùng đa dạng về tất cả các lĩnh vực trong đời sống. Tuy có giá trị vô cùng lớn nhưng không phải ai cũng biết khai thác nó và tận dụng để mang lại hiệu quả. Big data chủ yếu là những thông tin thô và ở dạng phi cấu trúc nên khó có thể nhìn thấy quan hệ giữa giữa các đối tượng từ đó không tận dụng được nguồn lợi chúng mang lại.
* Nhắm khai thác có hiệu quả, chúng ta cần những công cụ, phương thức để phân tích lượng thông tin khổng lồ nói trên, đó là mục đích của Đồ án thực tập 1 này.

# **Tổng quát về nội dung đồ án:**

* **Nội dung đồ án:** Thu thập các dữ liệu tuyển dụng của doanh nghiệp, tổ chức lưu trữ chúng và dùng API để chia sẻ dữ liệu đến các trang web và thiết bị di động đồng thời trực quan hóa để người dùng có cái nhìn rõ ràng nhất, từ đó đưa ra các nhận định để giới thiệu công nghệ DMS tới các doanh nghiệp.
* Đồ án thực tập 1 này sẽ gồm các bước chính sau:
* Thu thập và làm sạch dữ liệu **(Data Gathering and Cleansing).**
* Tổ chức và lưu trữ dữ liệu phi cấu trúc **(Unstructure data Organization and Storage).**
* Chia sẻ thông tin thông qua API **(Sharing Data).**
* Trực quan hóa dữ liệu **(Data Visualization).**
* Đồng bộ dữ liệu trên thiết bị di động **(Data synchronization with mobile).**
* Để giới hạn phạm vi của đồ án, chúng ta chỉ tập trung vào các doanh nghiệp hoạt động trong *3 ngành thực phẩm, dược phẩm và mĩ phẩm*.
* Theo phân công của đồ án, nhóm 1 thực hiện công việc thu thập và làm sạch dữ liệu tuyển dụng từ các trang web của các doanh nghiệp. Đây là công việc khởi đầu cho cả đồ án và cũng là công việc quan trọng cần sự chính xác, tỉ mỉ và sự phối hợp tốt với các nhóm khác.

# See the source imageVề công ty DMSpro:

# **Giới thiệu:**

* DMSpro jsc là công ty chuyên cung cấp giải pháp DMS tiên phong ở Việt Nam sử dụng điện toán đám mây trong. Được thành lập vào năm 2011, DMSpro nhanh chóng phát triển và ngày càng có tầm ảnh hưởng lớn đối với các doanh nghiệp kinh doanh các sản phẩm tiêu dùng.
* Các giải pháp của công ty hoạt động trên rất nhiều lĩnh vực như đồ uống và nước giải khát, sữa, dược phẩm cho đến mỹ phẩm làm đẹp và điện tử, …
* Không chỉ hướng đến việc giúp doanh nghiệp quản lý tốt khâu kinh doanh của mình, DMSpro còn hướng đến những người bán lẻ với mục đích giúp họ tiếp cận các sản phẩm, thông tin khuyến mãi của doanh nghiệp tốt hơn.

# **Các sản phẩm, giải pháp tiêu biểu:**

* **retailPro**: Là ứng dụng công nghệ được xây dựng trên nền tảng điện toán đám mây Azure của Microsoft. Ứng dụng giúp chuỗi cửa hàng, nhà bán lẻ & các tiệm tạp hóa đặt hàng trực tiếp từ nhà sản xuất/nhà phân phối, đồng thời dễ dàng đăng ký tham gia các chương trình khuyến mãi, trưng bày trực tuyến, theo dõi lịch sử mua hàng.
* **pgPro**: Là ứng dụng giúp tối ưu hóa đội ngũ tiếp thị bằng các chức năng như hiển thị rõ ràng các nhiệm vụ cần làm, cho biết hiệu quả làm việc, nhận các hướng dẫn kịp thời, và kiểm soát lượng hàng bán ra
* **calendarPro**: Giúp cho người quản lý bán hàng các cấp có thể lập lịch biểu cho mình hoặc nhân viên cấp dưới từ đó gia tăng tính tương tác và phối hợp công việc với nhau.
* **visibilityPRO**: Giải pháp quản lý trực quan bằng hình ảnh sẽ cung cấp hình ảnh trực quan theo thời gian thực về tình hình thị trường, kết quả kinh doanh của nhà phân phối, hoạt động bán hàng của nhân viên và trưng bày tại từng cửa hiệu. Từ đó, các doanh nghiệp, nhà sản xuất phân phối, bán hàng hòa toàn có thể phân tích được độ phủ thị trường.
* **SFA – Sales Force Automation** : Hỗ trợ nhân viên bán hàng thao tác bán hàng theo quy trình chuẩn của doanh nghiệp, giúp nhân viên bán hàng tiết kiệm thời gian, tăng chất lượng chăm sóc khách hàng, mở rộng độ phủ thị trường và tối đa doanh thu.
* **eroutePro**: Quản lý tuyến bán hàng trên bản đồ số. Ứng dụng hiển thị phân bổ vị trí điểm bán hàng trên bản đồ, tối ưu hóa tuyến bán hàng, hỗ trợ nhân viên và hình ảnh tại cửa hàng.

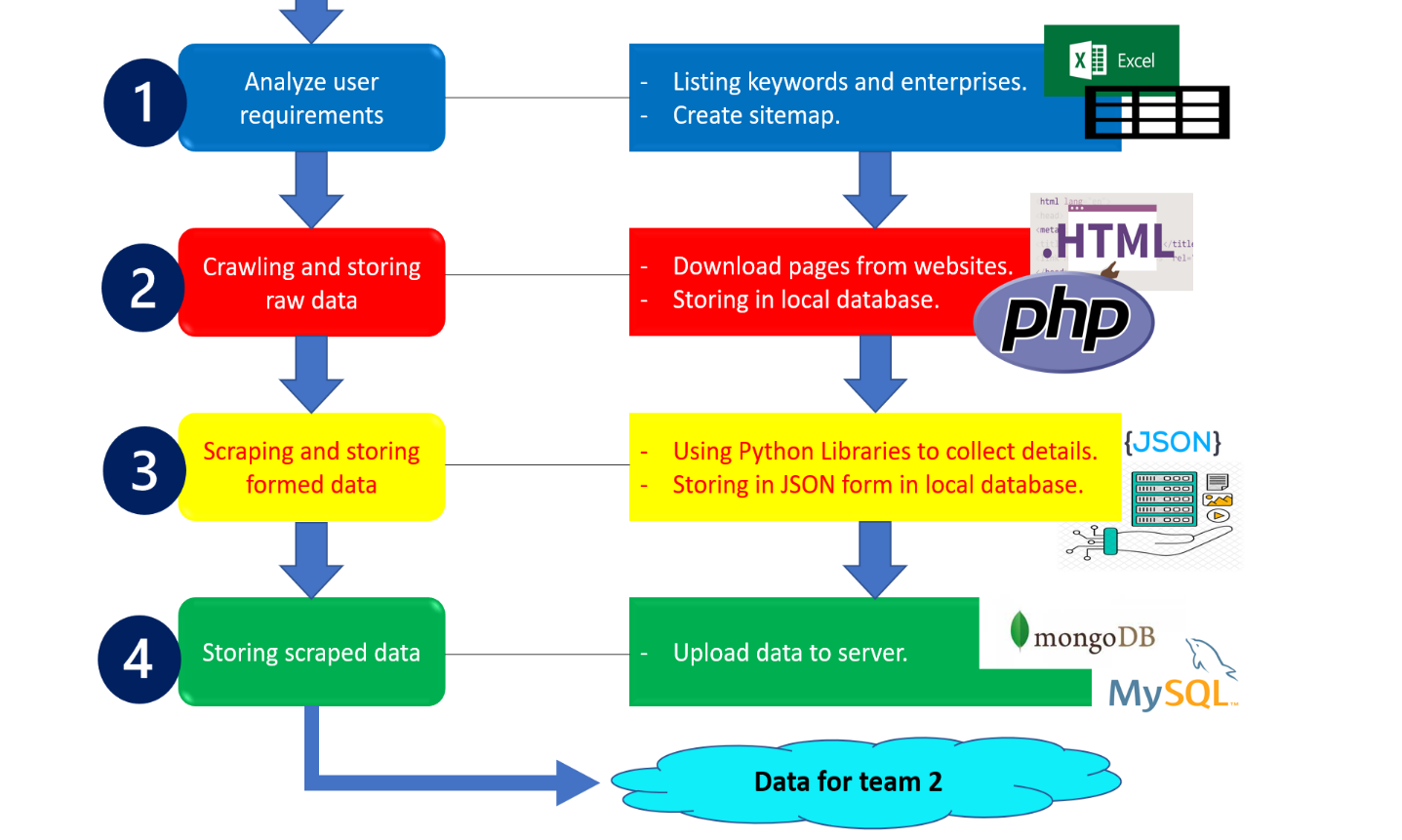
# **Danh hiệu , thành tựu:**

* Danh hiệu 4 sao tại Sao Khuê 2015.
* Top 50 doanh nghiệp CNTT Việt Nam 2017.
* Sản phẩm tiêu biểu của ngành phần mềm, dịch vụ CNTT Việt Nam tại Sao Khuê 2018 với Giải pháp phân phối và bán hàng trên nền tảng điện toán đám mây S.DMS.
* Top 50 doanh nghiệp CNTT hàng đầu và TOP 10 doanh nghiệp có năng lực công nghệ 4.0 tiêu biểu của Việt Nam 2019.
* Danh hiệu Sao Khuê 2020 cho sản phẩm, giải pháp ứng dụng công nghệ mới trong cuộc cách mạng công nghệ 4.0 với retailPro và pgPro.

Trang web công ty: <https://www.dmspro.vn>

# Thu thập và làm sạch dữ liệu (Data Gathering and Cleansing):

Công việc chính của quá trình này là *tập hợp dữ liệu* từ các nguồn đã xác định trước thỏa mãn những yêu cầu của người dùng đã đề ra sau đó *trích lọc những thông tin cần thiết* tạo thành các đoạn chữ và cuối cùng là *đưa chúng vào cấu trúc dữ liệu JSON* theo các trường rồi lưu trữ trên MongoDB.



# **Phân tích yêu cầu người dùng:**

* Đưa ra các hướng đi phù hợp: mục đích, yêu cầu của người dùng, các thông tin cần có, thành quả mong muốn đạt được,… Từ các yêu cầu trên ta sẽ xác định nguồn dữ liệu.
* *Liệt kê các từ khóa* hỗ trợ trong việc cào và trích lọc dữ liệu nhằm tăng khả năng xử lý, thu hẹp phạm vi tìm kiếm đồng thời tối ưu lượng thông tin thu thập được.
* *Liệt kê danh sách doanh nghiệp* trong 3 ngành thực phẩm F&B, dược phẫm Pharma và mỹ phẩm Cosmetic.
* *Tạo sơ đồ đường đi* cho hoạt động trích lọc dữ liệu: thể hiện nơi tập trung các thông tin của trang web của doanh nghiệp từ đó tập trung khai thác vào phần này, tránh các đoạn mã HTML thực hiện các chức năng của trang.

# Lấy dữ liệu (chủ yếu là các trang HTML, php):

* Để tránh việc bị phát hiện đang khai thác thông tin từ trang đó đồng thời phục vụ cho mục đích thực hành, ta sẽ lưu trữ các vào cơ sở dữ liệu. Sau khi hoàn thành đồ án, ta có thể sử dụng các nền tảng bên thứ 3 cung cấp hỗ trợ việc chuyển đổi IP, thiết lập lịch trình cho công việc thu thập dữ liệu.

# Trích lọc thông tin cần thiết và lưu trữ dữ liệu đã cấu trúc:

* Dùng các thư viện Python hỗ trợ trích lọc dữ liệu từ trang web, ta sẽ thu thập được các thông tin cần thiết dưới dạng chuỗi và từ các từ khóa đã liệt kê để đưa các thông tin đó vào các trường dưới dạng JSON.

# Lưu trữ dữ liệu trên các nền tảng lưu trữ cho công việc nhóm sau:

* Lưu trữ dữ liệu tuyển dụng vào MongoDB theo các trường dữ liệu đã thống nhất với các nhóm company\_code, time\_post, time\_deadline, vacancy, quantity, salary, workplace, requirements.

# Các công cụ được sử dụng:

* Python: Python là một ngôn ngữ lập trình mạnh và ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực của ngành công nghệ thông tin ngày nay vì tính đơn giản, dễ hiểu, hiệu quả và có một cộng đồng hoạt động tích cực, các thư viện được thường xuyên cập nhật và sửa lỗi. Cụ thể trong quá trình thực hiện, nhóm đã sử dụng các thư viện sau:
* BeautifulSoup4: BeautifulSoup là một thư viện Python dùng để lấy dữ liệu ra khỏi các file HTML và XML. Nó hoạt động cùng với các parser (trình phân tích cú pháp) cung cấp cho bạn các cách để điều hướng, tìm kiếm và chỉnh sửa trong parse tree (cây phân tích được tạo từ parser). Nhờ các parser này nó đã giúp các lập trình viên tiết kiệm được nhiều giờ làm việc. BeautifulSoup4 sử dụng module lxml có sẵn khi cài đặt Python cung cấp quyền truy cập an toàn và thuận tiện bằng API ElementTree.
* requests: Sử dụng để gửi tất cả các loại yêu cầu HTTP, truyền các tham số trong URL, gửi các header tùy biến và xác minh SSL.
* subprocess, os: Đây là thư viện có sẵn khi cài đặt Python, giúp chạy các lệnh tác động lên hệ thống.
* json: Thư viện hỗ trợ tương tác với kiểu dữ liệu JSON.
* MongoDB: Là nền tảng lưu trữ dữ liệu phi cấu trúc hỗ trợ đa nền tảng được thiết kế theo hướng đối tượng, có cấu trúc linh hoạt.Vì thế MongoDB dùng để lưu trữ dữ liệu phức tạp và đa dạng theo kiểu JSON.
* MySQL: Là một hệ thống quản trị cơ sở dữ liệu mã nguồn mở hoạt động theo mô hình client-server.
* Jira: Là phần mềm quản lý công việc, theo dõi lỗi vấn đề trong dự án, cách hoạt động dựa vào trọng tâm là kết quả công việc. Các tính năng đáng chú ý của Jira như khả năng quản lý, theo dõi tiến độ của dự án; quản lý các công việc, lỗi, cải tiến, tính năng mới; tạo và lưu những bộ lọc có cấu hình cao xuyên suốt mọi vấn đề trong hệ thống; xây dựng quy trình làm việc tương thích với yêu cầu của từng dự án; cung cấp nhiều loại báo cáo thống kê với nhiều loại biểu đồ khác nhau.
* GitHub: là một hệ thống quản lý dự án và phiên bản code, hoạt động giống như một mạng xã hội cho lập trình viên. Các lập trình viên có thể clone lại mã nguồn từ một repository và Github chính là một dịch vụ máy chủ repository công cộng, mỗi người có thể tạo tài khoản trên đó để tạo ra các kho chứa của riêng mình để có thể làm việc. Git Hub có đầy đủ những tính năng của Git, ngoài ra còn bổ sung những tính năng để tương tác với nhau.

Chi tiết quá trình thực hiện:

* Quá trình làm việc được chia thành các Sprint từ 0 đến 5 ứng với các bước thực hiện như trong data pipeline. Trong mỗi sprint, các công việc vể mặt kĩ thuật yêu cầu tất cả các thành viên thực hiện, các công việc về viết tài liệu và làm báo cáo cuối sprint sẽ chia đều cho các thành viên.
* Quá trình làm việc hằng tuần:
* Mỗi ngày, tất cả thành viên sẽ gặp mặt với mentor bên công ty DMSpro để báo cáo việc mình làm trong ngày, dự tính công việc hôm nay và các vấn đề gặp phải trong quá trình làm việc (tất cả được lưu trên file Daily meeting.xlsx <https://bitly.com.vn/7j1qhd> )
* Cuối mỗi tuần, các thành viên sẽ tập trung và thực hiện trình bày kết quả mình làm được trong tuần, chuẩn bị báo cáo với giáo viên bộ môn trong khoa và với các nhóm khác.
* Bắt đầu mỗi sprint, đại diện nhóm sẽ họp với các nhóm khác để thống nhất các việc cần thiết cho đồ án và thông báo lại cho team.
* Các hoạt động sẽ diễn ra trên nền tảng MS Team do trường cung cấp hoặc Skype. Code sẽ được lưu trên GitHub <https://github.com/DMSPRO-JSC/Data-Gathering-Cleansing> . Các tài liệu sẽ được lưu trên MS Team. Các công việc sẽ được quản lý trên Jira( <https://jira.dmspro.vn:8443/projects/DGC/summary>)

23/11

21/12

14/12

7/12

30/11

16/11

9/11

# 1. Sprint 0: Tìm tài liệu về các công cụ hỗ trợ trong đồ án:

* Mục tiêu: Các thành viên tìm hiểu về các công cụ, thư việc Python, cách chúng hoạt động trong việc thu thập và làm sạch dữ liệu.
* Công việc cần hoàn thành:
* Mỗi thành viên tìm hiểu về các công cụ, phần mềm có sẵn hỗ trợ trong việc thu thập và làm sạch thông tin.
* Tìm hiểu về các thư viện Python để xử lý HTML, php, txt.
* Kết quả: Các công cụ tìm hiểu được chia thành 3 loại chính:
* Các công cụ, phần mềm có sẵn: Các phần mềm này cung cấp giao diện đồ họa thân thiện với người dùng, giúp người dùng có thể lấy được thông tin từ các trang web mà không cần quá nhiều hiểu biết về cách chúng hoạt động như ***ScrapeStorm, ScreamingFog, Cyotek WebCopy, Octoparse, ….*** Các công cụ này nhanh, tiện lợi nhưng khả năng chính xác không cao và thường chỉ hoạt động trên các trang web có cấu trúc giống hệt nhau( các thông tin phải nằm trong cùng tag).
* Các thư viện Python: Có nhiều thư viện Python hỗ trợ xử lý thông tin từ website, chúng được cộng đồng liên tục cập nhật và hỗ trợ như ***DeliciousAPI, BeautifulSoup, selenium, scrapy, SQLAIchemy, …*** Trong đồ án này, nhóm sẽ quyết định sử dụng BeautifulSoup để lấy thông tin từ các trang web.
* Các nền tảng hỗ trợ cho việc thu thập dữ liệu: Các nền tảng này hỗ trợ các chức năng như chuyển IP ( IP rotating), thiết lập thời gian cho việc thu thập thông tin( scheduling), … Có thể kể đến scrapy.org, scrapinghub.com.

Ngoài ra, các thành viên còn tìm hiểu về các chủ đề khác như cách thức tìm kiếm các từ khóa, các yếu tố ảnh hưởng đến việc cạnh tranh trên các máy tìm kiếm, xử lý ngôn ngữ tự nhiên. Các tài liệu tìm được có trong thư mục tài liệu của đồ án (link ở cuối báo cáo).

# 2. Sprint 1: Tìm hiểu yêu cầu người dùng:

* Kết quả cần đạt được:

+ Danh sách các công ty hoạt động trong 3 lĩnh vực thực phẩm, dược phẩm và mĩ phẩm.

|  |  |  |
| --- | --- | --- |
| Name | string | Tên doanh nghiệp |
| Description bussiness model | string | Mô tả loại hình kinh doanh |
| Size | int | Quy mô (đánh giá qua số lượng nhân viên) |
| Address | string | Địa chỉ trụ sở chính |
| Industry | string | Lĩnh vực hoạt động |
| Method discover | string | Nguồn thông tin tìm thấy công ty |
| Source found | string | Đường dẫn tới tranh tuyển dụng |
| Sitemap | string | Sơ đồ của tranh web( file xml hoặc ghi thủ công) |

|  |  |  |
| --- | --- | --- |
| Keywords | string | Từ khóa |
| Type | string | Short/medium/long tail |
| Popularity | int | Số lần tìm kiếm trong 3 tháng gần nhất |
| Keywords Difficult | float | Độ khó để đưa bài đăng có từ khóa vào top đầu của các máy tìm kiếm |

+ Danh sách các từ khóa dùng cho việc thu thập thông tin.

+ Hướng giải quyết cho đồ án.

* Công việc cần hoàn thành:

+ Liệt kê danh sách các công ty <https://bitly.com.vn/bbcgk0>.

+ Liệt kê danh sách các từ khóa <https://bitly.com.vn/mc9k7t> .

+ Thiết kế hướng đi cho hoạt động của nhóm.

* Kết quả:

+ Danh sách các công ty và từ khóa được lưu theo định dạng các trường.

+ Hướng giải quyết cho đồ án (Draft Solution): Tài liệu này giúp định hướng, xác định các bước cần thực hiện để đạt kết quả cuối cùng, là mẫu cho data pipeline sau này đồng thời là khuôn mẫu cho các hoạt động trước mắt <https://bitly.com.vn/vgkbn8>

**1.Xác định yêu cầu người dùng:** Mục đích là tìm các doanh nghiệp có khả năng đầu tư vào DMS thông qua nghiên cứu dữ liệu tuyển dụng của họ.

**2.Xác định nguồn dữ liệu:** Có thể là trang chủ của các doanh nghiệp, mạng xã hội hoặc các tranh tuyển dụng.

**3.Xác định các từ khóa:** Các từ khóa hỗ trợ tìm kiếm thông tin như keyword của của các trường dữ liệu sử dụng trong json, các từ khóa liên quan đến DMS.

**4.Tạo sơ đồ cho các trang web:** Mỗi doanh nghiệp sẽ có cách thiết kế website khác nhau và nhằm giúp cho việc thu thập thông tin trở nên nhanh chóng và có thể áp dụng phổ biến, ta cần thiết lập sơ đồ nơi tập trung thông tin và tiến hành khai thác.

**5.Thu thập và xử lý thông tin:** Sử dụng Python viết các chương trình thu thập và lưu các thông tin đó dưới dạng json trên MongoDB.

# 3. Sprint 2: Dùng các phần mềm bên thứ 3 cung cấp để tìm hiểu về việc thu thập dữ **liệu:**

* Mục tiêu:

+ Sơ đồ sitemap của các trang web tuyển dụng.

+ Dùng các phần mềm để thu thập thông tin từ các trang web trên.

* Công việc cần hoàn thành:

+ Sitemap của các công ty được lưu trong file ListEnterprise.xlsx trên MS team.

+ Dữ liệu mẫu từ các trang tuyển dụng.

+ Dữ liệu mẫu cho nhóm 2.

* Kết quả:

+ Sơ đồ của các trang web trong sprint này được tạo bằng các công cụ tìm được trong sprint 0, gồm các link của các trang web.

+ Sitemap của các trang web tuyển dụng giúp cho việc thu thập thôn tin trở nên nhanh hơn, thuận tiện hơn, tiết kiệm thời gian khi không phải quét toàn trang web để tìm dữ liệu. Các sitemap được lưu trên GitHub <https://github.com/DMSPRO-JSC/Data-Gathering-Cleansing/tree/NewsCrawler> .

+ Dữ liệu mẫu thu được qua việc sử sụng các phần mềm là đầy đủ nhưng chưa đạt yêu cầu do độ chính xác không cao và việc chỉ quét những tag HTML tương tự khi người dùng chọn ban đầu làm cho dữ liệu bị thiếu, không đúng với dịnh dạng thống nhất ban đầu. Dữ liệu thu thập được bằng việc sử dụng các công cụ <https://bitly.com.vn/ti1m5w> .

+ Dữ liệu mẫu gửi cho team 2 là dữ liệu thu được từ việc sử dụng các công cụ nhưng cần qua chỉnh sửa thủ công nhiều và cũng chưa đạt yêu cầu về độ sạch.

# **4. Sprint 3: Thiết kế đường dẫn dữ liệu và tiến hành cào các trang web:**

* Mục tiêu:

+ Dùng các phần mềm có sẵn để thu thập dữ liệu.

+ Dữ liệu về doanh nghiệp trên MySQL.

+ Thiết kế đường đi của dữ liệu.

* Công việc cần hoàn thành:

+ Tạo các code làm việc cào dữ liệu từ các trang tuyển dụng sử sụng sitemap đã tạo trong sprint 2.

+Thiết kế đường đi của dữ liệu (data pipeline).

+ Lưu trữ dữ liệu về các doanh nghiệp trong MySQL theo các trường đã định sẵn.

* Kết quả:

+ Sử dụng Python làm nhiệm vụ lấy các trang dưới dạng HTML, php từ các trang web tuyển dụng của doanh nghiệp thành các file HTML, php, txt trong máy cá nhân phục vụ cho mục đích thực hành thu thập thông tin do đã trình bày ở trên về việc bị chặn IP khi gửi các yêu cầu liên tục tới các trang web <https://github.com/DMSPRO-JSC/Data-Gathering-Cleansing/blob/NewsCrawler/crawl.py> .

+ Dữ liệu về doanh nghiệp là dữ liệu có cấu trúc sẽ được lưu trong MySQL do dự liệu này không được truy vấn và biểu diễn nhiều.

|  |  |  |
| --- | --- | --- |
| Enterprise\_name | String | Tên công ty |
| Address | String | Đại chỉ trụ sở chính |
| Industry | String | Lĩnh vực kinh doanh |
| Scale | Int | Quy mô(số lượng nhân viên) |

+ Thiết kế đường đi của dữ liệu: Mô tả đường đi, các bước xử lý, các trạng thái của dữ liệu qua từng bước( đã trình bày ở phần III) nó mô phỏng kết quả cần đạt được của từng bước thực hiện, từ dữ liệu thô dưới dạng HTML, chuyển thành các đoạn chuỗi chứa thông tin tổng quát rồi được đưa vào trong các trường theo cấu trúc JSON <https://bitly.com.vn/1vhfz1> .

# **5.Sprint 4: Dùng Python để cào dữ liệu từ các trang web tuyển dụng:**

* Mục tiêu:

+ Code để cào dữ liệu từ các trang web mà không cần thông qua các file sitemap dạng xml.

+ Tạo cấu trúc, các trường dữ liệu cho file JSON.

+ Bắt đầu quá trình viết báo cáo cho đồ án.

* Công việc cần hoàn thành:

+ Mỗi thành viên phải có code thu thập dữ liệu từ 5 công ty trong danh sách đã xác định từ sprint 2.

+ Tập hợp tài liệu từ đầu để chuẩn bị viết báo cáo.

+ Lưu trữ dữ liệu về công việc trong MongoDB. Connection string : mongodb://10.86.94.2:27017.

* Kết quả:

+ Code thu thập dữ liệu từ các file HTML, php, txt đã tải về trên máy cá nhân, từ đó ghi dữ liệu vào file JSON theo các trường đã thống nhất trước.

+ Cấu trúc của file JSON sẽ có dạng:

|  |  |  |
| --- | --- | --- |
| Company\_code | Int | ID của doanh nghiệp |
| Time\_post | String | Ngày đăng/ ngày bắt đầu tuyển dụng |
| Time\_deadline | String | Ngày hết hạn tuyển dụng |
| Vacancy | String | Vị trí tuyển dụng |
| Quantity | Int | Số lượng tuyển dụng |
| Salary | String | Lương |
| Workplace | String | Nơi làm việc |
| Requirements | Array | Các yêu cầu |

+ Bắt đầu quá trình viết báo cáo dựa trên các báo cáo hàng tuần, tập hợp thành 1 thư mục riêng cho cả thầy cô và bên công ty DMSpro có thể xem được.

# 6. Sprint 5: Hoàn thành, tối ưu hóa các code để cào và thu thập dữ liệu, viết báo cáo cho đồ án.

* Mục tiêu:

+ Code cào và thu thập dữ liệu hoàn chỉnh.

+ Đưa code lên server, đặt lịch trình.

+ Hoàn thành báo cáo và chuẩn bị báo cáo cuối đồ án.

* Công việc cần hoàn thành:

+ Tối ưu hóa các code từ các sprint trước, cả code cào và code thu thập thông tin.

+ Tải các code cào và thu thập thông tin server, đặt thời gian họat động vào crontab.

+ Hoàn thành báo cáo cho đồ án.

* Kết quả:

+ Code cào thông tin mà chỉ gửi yêu cầu 1 lần tới website tuyển dụng từ đó lấy các link chứa thông tin tuyển dụng rồi dùng cmd để tải về. Code có thể áp dụng cho một nhóm các trang web có cấu trúc tương tự nhau.

+ Code thu thập thông tin từ các file đã cào về, code cần xử lý các glitch , thu thập đầy đủ và chuẩn hóa các dữ liệu thu thập được.

+ Đưa code lên server 10.86.94.2 do bên công ty DMSpro cung cấp, kiểm tra việc khởi chạy và lưu trữ. Ở đây nhóm sẽ đặt lịch trình 2 lần một tháng

A picture containing text, monitor, screenshot, screen

Description automatically generated(ngày 1 và 15) đồng thời các dữ liệu tìm được lên MongoDB tiếp theo các dữ liệu cũ.

# **7. Sprint 6: Chuẩn bị báo cáo cho đồ án:**

* Mục tiêu:

+ Báo cáo và trình chiếu cuối đồ án.

+ Các thành viên nắm rõ các bước, cách vận hành.

* Công việc cần hoàn thành:

+ Làm trình chiếu cuối đồ án.

+ Trao đổi về các vấn đề trong đồ án.

* Kết quả:

+ Hoàn thành bài báo cáo cho Đồ án thực tập 1.

- Dữ liệu về tuyển dụng của các doanh nghiệp dưới dạng JSON sẽ được nhóm 2 xử lý, tổ chức lưu trữ theo ObjectId trong MongoDB rồi mới được truy xuất qua API của nhóm 3 để nhóm 4 và 5 có thể biểu diễn trên web và thiết bị di động.

# **Kết quả thực tập:**

* Từ các dữ liệu tuyển dụng trên các trang của doanh nghiệp, bằng các đoạn mã Python đã trở thành dữ liệu dạng JSON trên MongoDB để có thể truy xuất và xử lý cách nhanh chóng, hiệu quả và mang lại ý nghĩa cho doanh nghiệp.
* Các thành viên còn được làm quen với Jira, ứng dụng quản lý công việc dùng cho các công ty trong các dự án, sử dụng GitHub làm nền tảng lưu trữ và trao đổi code cách hiệu quả mà trong quá trình học chưa bao giờ được thực hành.
* Do áp lực về thời gian đối với nhóm đầu và do khả năng hạn chế, nhóm đã không đạt kết quả như mong muốn cả về thời gian và chất lượng của đầu ra, nhưng kết quả cũng đáp ứng được một phần quá trình thực hành thu thập và làm sạch dữ liệu.
* Nhận thấy kết quả cuối cùng vẫn cần nhiều cài tiến, khả năng phát triển về các mặt:
* Các spider áp dụng được cho nhiều trang web nhất.
* Thu thập được nhiều và chính xác các thông tin.
* Phát triển các giao diện theo hướng thân thiện cho người sử dụng.

Toàn bộ tài liệu đề cập trong báo cáo: <https://bitly.com.vn/encj7g>

Jira : <https://jira.dmspro.vn:8443/projects/DGC/summary>

GitHub : <https://github.com/DMSPRO-JSC/Data-Gathering-Cleansing>

Các phần mềm, công cụ:

- ScrapeStorm : <https://www.scrapestorm.com>

- ScreamingFog : <https://www.screamingfrog.co.uk/seo-spider/>

- Cyotek WebCopy : <https://www.cyotek.com/cyotek-webcopy>

- Octoparse : <https://www.octoparse.com>

- scrapy : scrapy.org

- ScrapingHub : <https://app.scrapinghub.com>