TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DỰ ÁN CÔNG NGHỆ THÔNG TIN 2**

**Phát triển hệ thống xử lý song song thu thập dữ liệu trên nền tảng PySpark và ứng dụng**

*Người hướng dẫn*: **PhGS LÊ ANH CƯỜNG**

*Người thực hiện*: **TRẦN PHƯƠNG NGỌC ANH – 51702061**

**TRƯƠNG NGUYỄN NGỌC CHÂU – 51702068**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DỰ ÁN CÔNG NGHỆ THÔNG TIN 2**

**Phát triển hệ thống xử lý song song thu thập dữ liệu trên nền tảng PySpark và ứng dụng**

Người hướng dẫn: **PhGS LÊ ANH CƯỜNG**

Người thực hiện: **TRẦN PHƯƠNG NGỌC ANH**

**TRƯƠNG NGUYỄN NGỌC CHÂU**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Trong quá trình thực hiện dự án công nghệ thông tin 2, không thể không nhắc đến sự quan tâm và hướng dẫn tận tình của giảng viên hướng dẫn – thầy Lê Anh Cường. Chúng em xin gửi lời cảm ơn chân thành đến thầy, người truyền tải những kiến thức và kinh nghiệm dẫn dắt chúng em hoàn thành dự án một cách tốt nhất. Bên cạnh đó, những lời hướng dẫn và dạy bảo tận tình của thầy còn là nguồn tri thức quý báu cho chúng em tích lũy thêm nhiều kiến thức mới trên con đường hoàn thiện chính mình.

Do vốn hiểu biết vẫn còn hạn chế nên trong quá trình thực hiện không thể tránh khỏi có những thiếu sót. Chúng em rất mong nhận được những ý kiến đóng góp của thầy để bài làm được hoàn thiện hơn.

Một lần nữa chúng em xin chân thành cảm ơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của chúng tôi và được sự hướng dẫn của PhGS Lê Anh Cường;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 28 tháng 05 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trần Phương Ngọc Anh*

*Trương Nguyễn Ngọc Châu*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Đề tài phát triển hệ thống xử lý song song thu thập dữ liệu (tin tức và văn bản) các cơ quan bộ ngành trên nền tảng PySpark và ứng dụng bao gồm các phần như sau:

1. Thu thập dữ liệu của các cơ quan bộ ngành: crawl các loại tin tức, văn bản pháp quy và văn bản điều hành từ trang web chính thức của các cơ quan bộ ngành.
2. Xây dựng hệ thống xử lý song song trên nền tảng PySpark.
3. Phát triển ứng dụng bằng cách tích hợp tích hợp hệ thống xử lý song song thu thập dữ liệu vào trang web để hiển thị dữ liệu được crawl.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc25746)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc14210)

[TÓM TẮT iv](#_Toc15691)

[MỤC LỤC 1](#_Toc2191)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4](#_Toc18331)

[CHƯƠNG 1 – GIỚI THIỆU 5](#_Toc17380)

[1.1 Giới thiệu đề tài: 5](#_Toc3824)

[1.2 Kết quả mong đợi: 6](#_Toc27011)

[CHƯƠNG 2 – TỔNG QUAN 8](#_Toc29780)

[2.1. Tổng quan hệ thống: 8](#_Toc23350)

[2.2. Mô tả chức năng hệ thống: 8](#_Toc19392)

[2.3. Công nghệ sử dụng: 9](#_Toc6401)

[2.3.1. Python: 9](#_Toc26965)

[2.3.1.1. Tính năng chính của Python 10](#_Toc20994)

[2.3.1.2. Ứng dụng của Python: 11](#_Toc17712)

[2.3.2. Scrapy: 13](#_Toc30069)

[2.3.3. SQL: 15](#_Toc21224)

[2.3.3.1. Lịch sử phát triển: 15](#_Toc3257)

[2.3.3.2. Đặc điểm và các ấn bản của SQL: 16](#_Toc20988)

[ Đặc điểm: 16](#_Toc11904)

[ Các ấn bản của SQL Server: 17](#_Toc18109)

[2.3.3.3. Mục đích sử dụng SQL: 18](#_Toc7322)

[2.3.3.4. Tìm hiểu thêm về cấu trúc sử dụng SQL Server: 18](#_Toc1910)

[2.3.3.5. Các dịch vụ, công cụ và phiên bản SQL Server: 19](#_Toc26872)

[2.3.4. Web - Django: 20](#_Toc26591)

[2.3.4.1. Ưu điểm: 21](#_Toc9173)

[2.3.4.2. Nhược điểm: 22](#_Toc7454)

[2.3.4.3. Mô hình MTV (Model - Template - View): 22](#_Toc25648)

[2.3.4.4. Cấu trúc thư mục của một project Django: 23](#_Toc1767)

[2.3.4.5. Lí do nên chọn Django: 24](#_Toc17649)

[2.3.4.6. Các nền tảng sử dụng Django: 25](#_Toc11892)

[2.4. PySpark - kỹ thuật song song: 26](#_Toc1145)

[CHƯƠNG 3 – THỰC NGHIỆM 27](#_Toc4720)

[3.1. Các sơ đồ đặc tả hệ thống: 27](#_Toc8555)

[3.2. Phân tích cơ sở dữ liệu: 27](#_Toc20234)

[3.2.1. Bảng ministry\_info: 28](#_Toc26344)

[3.2.2. Bảng article\_category\_type: 28](#_Toc27594)

[3.2.3. Bảng legislation\_category\_type: 29](#_Toc11168)

[3.2.4. Bảng article\_category\_info: 29](#_Toc9437)

[3.2.5. Bảng legislation\_category\_info: 30](#_Toc25178)

[3.2.6. Bảng ministry\_article\_category\_configuration: 30](#_Toc29683)

[3.2.7. Bảng ministry\_article\_detail\_configuration: 31](#_Toc20460)

[3.2.8. Bảng ministry\_legislation\_category\_configuration: 32](#_Toc31973)

[3.2.9. Bảng ministry\_legislation\_detail\_configuration: 32](#_Toc6582)

[3.2.10. Bảng article\_info: 33](#_Toc25794)

[3.2.11. Bảng legislation\_info: 34](#_Toc14612)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

**CÁC CHỮ VIẾT TẮT**

CSTD Công suất tác dụng

MF Máy phát điện

BER Tỷ lệ bít lỗi

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 2. 1. Hình ảnh thư viện Scrapy trong Python 13](#_Toc2106)

[Hình 2. 2. Hình ảnh Microsoft SQL Server 15](#_Toc5970)

[Hình 2. 3. Minh họa cấu trúc sử dụng SQL Server 19](#_Toc6577)

[Hình 2. 4. Minh họa tính năng doanh nghiệp SQL Server 20](#_Toc11088)

[Hình 2. 5. Hình ảnh Django Python 21](#_Toc29996)

[Hình 2. 6. Hình ảnh mô hình MTV (Model - Template - View) Django 23](#_Toc160)

[Hình 2. 7. Ví dụ cấu trúc thư mục của một project Django 24](#_Toc10594)

[Hình 2. 8. Minh họa project sử dụng Django 25](#_Toc12252)

**DANH MỤC BẢNG**

[Bảng 1. 1. Danh sách thông tin các bộ ngành tại Việt Nam 6](#_Toc5732)

[Bảng 1. 2. Mô tả mô hình MTV (Model - Templat - View) Django 23](#_Toc13209)

[Hình 3. 1. Sơ đồ usecase 27](#_Toc26915)

[Hình 3. 2. Tổng quan cơ sở dữ liệu 28](#_Toc2714)

[Bảng 3. 1 Bảng ministry\_info 28](#_Toc25598)

[Bảng 3. 2. Bảng article\_category\_type 29](#_Toc28013)

[Bảng 3. 3. Bảng legislation\_category\_type 29](#_Toc25469)

[Bảng 3. 4. Bảng article\_category\_info 30](#_Toc4259)

[Bảng 3. 5. legislation\_category\_info 30](#_Toc18406)

[Bảng 3. 6. Bảng ministry\_article\_category\_configuration 31](#_Toc24556)

[Bảng 3. 7. Bảng ministry\_article\_detail\_configuration 32](#_Toc12444)

[Bảng 3. 8. Bảng ministry\_legislation\_category\_configuration 32](#_Toc15795)

[Bảng 3. 9.Bảng ministry\_legislation\_detail\_configuration 33](#_Toc31918)

[Bảng 3. 10. Bảng ministry\_legislation\_detail\_configuration 34](#_Toc8684)

[Bảng 3. 11. Bảng legislation\_info 35](#_Toc3285)

CHƯƠNG 1 – GIỚI THIỆU

* 1. Giới thiệu đề tài:

Trong thời đại mà các thông tin gây nhiễu xuất hiện tràn lan và khó kiểm soát, ta sẽ khó có thể nắm bắt được những tin thật từ những trang tin tức chính thống từ các cơ quan bộ ngành. Để mọi người có thể cập nhật kịp thời các tin tức và các văn bản mới, chúng em cho ra đời hệ thống xử lý song song thu thập dữ liệu từ các trang web của các tổ chức, cơ quan thuộc các bộ và các ngành.

Các trang tin tức chính thống của các bộ và các ngành được thu thập bao gồm:

|  |  |  |
| --- | --- | --- |
| STT | BỘ\NGÀNH | LINK |
| 1 | Bộ Công Thương | <http://www.moit.gov.vn/web/guest> |
| 2 | [Bộ Công an](https://www.moha.gov.vn/danh-muc/bo-cong-an-9982.html) | <http://bocongan.gov.vn/> |
| 3 | [Bộ Giáo dục và Đào tạo](https://www.moha.gov.vn/danh-muc/bo-giao-duc-va-dao-tao-9997.html) | <https://moet.gov.vn/Pages/home.aspx> |
| 4 | [Bộ Giao thông vận tải](https://www.moha.gov.vn/danh-muc/bo-giao-thong-van-tai-9990.html) | <https://www.mt.gov.vn/vn/Pages/Trangchu.aspx> |
| 5 | [Bộ Kế hoạch và Đầu tư](https://www.moha.gov.vn/danh-muc/bo-ke-hoach-va-dau-tu-9987.html) | [http://www.mpi.gov.vn/Pages/default.aspx](http://www.mpi.gov.vn/Pages/default.aspx" \o "http://www.mpi.gov.vn/Pages/default.aspx) |
| 6 | [Bộ Khoa học và Công nghệ](https://www.moha.gov.vn/danh-muc/bo-khoa-hoc-va-cong-nghe-9994.html) | <https://www.most.gov.vn/vn/Pages/Trangchu.aspx> |
| 7 | [Bộ Lao động - Thương Binh và Xã hội](https://www.moha.gov.vn/danh-muc/bo-lao-dong-thuong-binh-va-xa-hoi-9995.html) | [http://www.molisa.gov.vn/Pages/trangchu.aspx](http://www.molisa.gov.vn/Pages/trangchu.aspx" \o "http://www.molisa.gov.vn/Pages/trangchu.aspx) |
| 8 | [Bộ Ngoại giao](https://www.moha.gov.vn/danh-muc/bo-ngoai-giao-9983.html) | [https://lanhsuvietnam.gov.vn/default.aspx](https://lanhsuvietnam.gov.vn/default.aspx" \o "https://lanhsuvietnam.gov.vn/default.aspx) |
| 9 | [Bộ Nội vụ](https://www.moha.gov.vn/danh-muc/bo-noi-vu-9984.html) | [https://www.moha.gov.vn/](https://www.moha.gov.vn/" \o "https://www.moha.gov.vn/) |
| 10 | [Bộ Nông nghiệp và Phát triển nông thôn](https://www.moha.gov.vn/danh-muc/bo-nong-nghiep-va-phat-trien-nong-thon-9989.html) | [https://www.mard.gov.vn/Pages/default.aspx](https://www.mard.gov.vn/Pages/default.aspx" \o "https://www.mard.gov.vn/Pages/default.aspx) |
| 11 | [Bộ Quốc phòng](https://www.moha.gov.vn/danh-muc/bo-quoc-phong-9981.html) | [http://www.mod.gov.vn/wps/portal](http://www.mod.gov.vn/wps/portal" \o "http://www.mod.gov.vn/wps/portal) |
| 12 | [Bộ Tài chính](https://www.moha.gov.vn/danh-muc/bo-tai-chinh-9988.html) | [https://www.mof.gov.vn/webcenter/portal/btc?\_afrLoop=604986911223584#%40%3F\_afrLoop%3D604986911223584%26\_adf.ctrl-state%3Du42aoybdh\_206](https://www.mof.gov.vn/webcenter/portal/btc?_afrLoop=604986911223584" \l "@?_afrLoop=604986911223584&_adf.ctrl-state=u42aoybdh_206" \o "https://www.mof.gov.vn/webcenter/portal/btc?_afrLoop=604986911223584#%40%3F_afrLoop%3D604986911223584%26_adf.ctrl-state%3Du42aoybdh_206) |
| 13 | [Bộ Tài nguyên và Môi trường](https://www.moha.gov.vn/danh-muc/bo-tai-nguyen-va-moi-truong-9992.html) | [https://monre.gov.vn/](https://monre.gov.vn/" \o "https://monre.gov.vn/) |
| 14 | [Bộ Thông tin và Truyền thông](https://www.moha.gov.vn/danh-muc/bo-thong-tin-va-truyen-thong-9993.html) | <https://www.mic.gov.vn/mic_2020/Pages/trangchu_2020.aspx> |
| 15 | [Bộ Tư pháp](https://www.moha.gov.vn/danh-muc/bo-tu-phap-9986.html) | [https://moj.gov.vn/Pages/home.aspx](https://moj.gov.vn/Pages/home.aspx" \o "https://moj.gov.vn/Pages/home.aspx) |
| 16 | [Bộ Văn hóa - Thể thao và Du lịch](https://www.moha.gov.vn/danh-muc/bo-van-hoa-the-thao-va-du-lich-9996.html) | [https://bvhttdl.gov.vn/](https://bvhttdl.gov.vn/" \o "https://bvhttdl.gov.vn/) |
| 17 | [Bộ Xây dựng](https://www.moha.gov.vn/danh-muc/bo-xay-dung-9991.html) | [https://moc.gov.vn/vn/Pages/Trangchu.aspx](https://moc.gov.vn/vn/Pages/Trangchu.aspx" \o "https://moc.gov.vn/vn/Pages/Trangchu.aspx) |
| 18 | [Bộ Y tế](https://www.moha.gov.vn/danh-muc/bo-y-te-9998.html) | [https://moh.gov.vn/](https://moh.gov.vn/" \o "https://moh.gov.vn/) |
| 19 | Chính phủ | [http://chinhphu.vn/portal/page/portal/chinhphu/trangchu](http://chinhphu.vn/portal/page/portal/chinhphu/trangchu" \o "http://chinhphu.vn/portal/page/portal/chinhphu/trangchu) |
| 20 | Ủy ban Dân tộc | [http://cema.gov.vn/home.htm](http://cema.gov.vn/home.htm" \o "http://cema.gov.vn/home.htm) |
| 21 | Ngân hàng Nhà nước Việt Nam | [https://www.sbv.gov.vn/webcenter/portal/vi/menu/trangchu;jsessionid=F5X8390vySpcPZ7QrftzbhKbWebL3\_-GSPGOln8XcoOOq\_7SYsV\_!1777602058!103147707?\_afrLoop=21080536281629224#%40%3F\_afrLoop%3D21080536281629224%26centerWidth%3D80%2525%26leftWidth%3D10%2525%26rightWidth%3D10%2525%26showFooter%3Dfalse%26showHeader%3Dfalse%26\_adf.ctrl-state%3D1afvgqb405\_4](https://www.sbv.gov.vn/webcenter/portal/vi/menu/trangchu;jsessionid=F5X8390vySpcPZ7QrftzbhKbWebL3_-GSPGOln8XcoOOq_7SYsV_!1777602058!103147707?_afrLoop=21080536281629224" \l "@?_afrLoop=21080536281629224&centerWidth=80%25&leftWidth=10%25&rightWidth=10%25&showFooter=false&showHeader=false&_adf.ctrl-state=1afvgqb405_4" \o "https://www.sbv.gov.vn/webcenter/portal/vi/menu/trangchu;jsessionid=F5X8390vySpcPZ7QrftzbhKbWebL3_-GSPGOln8XcoOOq_7SYsV_!1777602058!103147707?_afrLoop=21080536281629224#%40%3F_afrLoop%3D21080536281629224%26centerWidth%3D80%2525%26leftWidth%3D10%2525%26rig) |
| 22 | Văn phòng Chính phủ | <http://vpcp.chinhphu.vn/> |
| 23 | Bảo hiểm Xã hội Việt Nam | <https://baohiemxahoi.gov.vn/Pages/default.aspx> |
| 24 | Viện Hàn lâm Khoa học và Công nghệ Việt Nam | <https://vast.gov.vn/> |
| 25 | Viện Hàn lâm Khoa học Xã hội Việt Nam | <https://vass.gov.vn/Pages/Index.aspx> |
| 26 | Ủy ban quản lý vốn nhà nước tại doanh nghiệp | <http://cmsc.gov.vn/> |

Bảng 1. 1. Danh sách thông tin các bộ ngành tại Việt Nam

Hầu hết các trang web đều có mục tin tức gồm nhiều loại tin khác nhau, mục văn bản pháp quy và văn bản điều hành. Tuy nhiên có những trang không thể hiện các mục này, như trang web của Chính phủ, Bộ Tư pháp,…

* 1. Kết quả mong đợi:

Kết quả cần đạt được:

* Thành công crawl dữ liệu là tin tức và hai loại văn bản hiện thị trên các trang web.
* Xây dựng trang web để tải các dữ liệu đã thu thập được lên.
* Xây dựng hệ thống xử lý song song để thu thập tin tức nhanh hơn

CHƯƠNG 2 – TỔNG QUAN

* 1. Tổng quan hệ thống:

Ứng dụng thu thập dữ liệu nhằm mục đích thu thập những bài báo và những văn bản mới nhất được đăng tải ở trang web của các bộ ngành. Mục đích để cho người dùng có thể cập nhật tin tức chính thống nhanh và chính xác hơn. Ứng dụng được xây dựng trên nền tảng Web:

Các chức năng của ứng dụng:

* Crawl thông tin từ các trang web của bộ: sau mỗi 12 tiếng, hệ thống sẽ thu thập lại các bài báo mới và các văn bản mới.
* Phân loại tin tức: người dùng có thể phân loại tin tức theo bộ và theo loại theo từng mục tin tức và văn bản.
* Tìm kiếm: tìm kiếm văn bản, bài báo theo nhãn được gắn sẵn.
* Hiển thị chi tiết bài báo: hiển thị hình ảnh và nội dung bài báo/văn bản thu thập được từ trang web các bộ ngành

## 2.2. Người dùng hệ thống:

Admin (Quản lý ứng dụng): Kiểm soát các hoạt động thu thập và hiển thị dữ liệu lên trang web.

User (Người dùng): Sử dụng trang web để tìm kiếm thông tin và xem thông tin

* 1. Mô tả chức năng hệ thống:
* Crawl thông tin: trang web sử dụng thư viện Scrapy của Python dể thu thập thông tin của các bài báo và văn bản. Ở các bài báo, bộ thu thập sẽ lấy các thông tin như: hình ảnh, nội dung, tiêu đề, thời gian, tác giả. Ở các văn bản bộ thu thập sẽ lấy các thông tin như: số hiệu văn bản, tên, ngày ký, ngày có hiệu lực, trích yếu, người ký, cơ quan ban hành, tình trạng, thể loại và link download. Các thông sẽ được lấy dưới dạng xpath qua các trang web chính thống của bộ.
* Phân loại: các tin tức sau khi thu thập sẽ được đưa lên web, hiển thị theo các loại tin hoặc các loại văn bản. Các bài báo và văn bản này dượd phân loại vào từng mục nhất định theo các bộ - ngành ban hành. Người dùng có thể sử dụng chức năng này để tìm thông tin cần thiết nhanh hơn.
* Tìm kiếm: người dùng có thể tìm kiếm thông tin bằng cách nhập chi tiết tiêu đề hoặc các thuộc tính của bài báo hay thông qua các từ khóa đã được gắn nhãn.
* Hiển thị thông tin chi tiết: trên trang web sẽ hiển thị thông tin nội dung, hình ảnh, ngày đăng và tác giả như trên trang web của chính phủ. Người dùng có thể cập nhật thông tin ngay trên trang web mà không cần di chuyển đến trang web chính.
  1. Công nghệ sử dụng:
     1. Python:

Python là một [ngôn ngữ lập trình](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh) bậc cao cho các mục đích lập trình đa năng, do [Guido van Rossum](https://vi.wikipedia.org/w/index.php?title=Guido_van_Rossum&action=edit&redlink=1" \o "Guido van Rossum (trang chưa được viết)) tạo ra và lần đầu ra mắt vào năm [1991](https://vi.wikipedia.org/wiki/1990" \o "1990). Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu.

Python là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ, được tạo ra bởi Guido van Rossum. Nó dễ dàng để tìm hiểu và đang nổi lên như một trong những ngôn ngữ lập trình nhập môn tốt nhất cho người lần đầu tiếp xúc với ngôn ngữ lập trình. Python hoàn toàn tạo kiểu động và sử dụng cơ chế cấp phát bộ nhớ tự động. Python có cấu trúc dữ liệu cấp cao mạnh mẽ và cách tiếp cận đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp lệnh của Python là điểm cộng vô cùng lớn vì sự rõ ràng, dễ hiểu và cách gõ linh động làm cho nó nhanh chóng trở thành một ngôn ngữ lý tưởng để viết script và phát triển ứng dụng trong nhiều lĩnh vực, ở hầu hết các nền tảng.

* + - 1. Tính năng chính của Python

Ngôn ngữ lập trình đơn giản, dễ học: Python có cú pháp rất đơn giản, rõ ràng. Nó dễ đọc và viết hơn rất nhiều khi so sánh với những ngôn ngữ lập trình khác như C++, Java, C#. Python làm cho việc lập trình trở nên thú vị, cho phép bạn tập trung vào những giải pháp chứ không phải cú pháp.

* Miễn phí, mã nguồn mở: Bạn có thể tự do sử dụng và phân phối Python, thậm chí là dùng nó cho mục đích thương mại. Vì là mã nguồn mở, bạn không những có thể sử dụng các phần mềm, chương trình được viết trong Python mà còn có thể thay đổi mã nguồn của nó. Python có một cộng đồng rộng lớn, không ngừng cải thiện nó mỗi lần cập nhật.
* Khả năng di chuyển: Các chương trình Python có thể di chuyển từ nền tảng này sang nền tảng khác và chạy nó mà không có bất kỳ thay đổi nào. Nó chạy liền mạch trên hầu hết tất cả các nền tảng như Windows, macOS, Linux.
* Khả năng mở rộng và có thể nhúng: Giả sử một ứng dụng đòi hỏi sự phức tạp rất lớn, bạn có thể dễ dàng kết hợp các phần code bằng C, [C++](https://quantrimang.com/cplusplus" \o "Lập trình C++) và những ngôn ngữ khác (có thể gọi được từ C) vào code Python. Điều này sẽ cung cấp cho ứng dụng của bạn những tính năng tốt hơn cũng như khả năng scripting mà những ngôn ngữ lập trình khác khó có thể làm được.
* Ngôn ngữ thông dịch cấp cao: Không giống như [C](https://quantrimang.com/lap-trinh-c" \o "Lập trình C)/C++, với Python, bạn không phải lo lắng những nhiệm vụ khó khăn như quản lý bộ nhớ, dọn dẹp những dữ liệu vô nghĩa,... Khi chạy code Python, nó sẽ tự động chuyển đổi code sang ngôn ngữ máy tính có thể hiểu. Bạn không cần lo lắng về bất kỳ hoạt động ở cấp thấp nào.
* Thư viện tiêu chuẩn lớn để giải quyết những tác vụ phổ biến: Python có một số lượng lớn thư viện tiêu chuẩn giúp cho công việc lập trình của bạn trở nên dễ thở hơn rất nhiều, đơn giản vì không phải tự viết tất cả code. Ví dụ: Bạn cần kết nối [cơ sở dữ liệu](https://quantrimang.com/co-so-du-lieu" \o "Cơ sở dữ liệu) MySQL trên Web server? Bạn có thể nhập thư viện MySQLdb và sử dụng nó. Những thư viện này được kiểm tra kỹ lưỡng và được sử dụng bởi hàng trăm người. Vì vậy, bạn có thể chắc chắn rằng nó sẽ không làm hỏng code hay ứng dụng của mình.

Hướng đối tượng: Mọi thứ trong Python đều là hướng đối tượng. [Lập trình hướng đối tượng](https://quantrimang.com/steve-jobs-dinh-nghia-lap-trinh-huong-doi-tuong-khien-ca-the-gioi-than-phuc-131900" \t "_blank" \o "Steve Jobs định nghĩa lập trình hướng đối tượng khiến cả thế giới thán phục) (OOP) giúp giải quyết những vấn đề phức tạp một cách trực quan. Với OOP, bạn có thể phân chia những vấn đề phức tạp thành những tập nhỏ hơn bằng cách tạo ra các đối tượng.

* + - 1. Ứng dụng của Python:

Lập trình ứng dụng web: Bạn có thể tạo web app có khả năng mở rộng (scalable) được bằng cách sử dụng framework và CMS (Hệ thống quản trị nội dung) được tích hợp trong Python. Vài nền tảng phổ biến để tạo web app là: Django, Flask, Pyramid, Plone, Django CMS. Các trang như Mozilla, Reddit, Instagram và PBS đều được viết bằng Python.

Khoa học và tính toán: Có nhiều thư viện trong Python cho khoa học và tính toán số liệu, như SciPy và NumPy, được sử dụng cho những mục đích chung chung trong tính toán. Và, có những thư viện cụ thể như: EarthPy cho khoa học trái đất, AstroPy cho Thiên văn học,... Ngoài ra, Python còn được sử dụng nhiều trong machine learning, khai thác dữ liệu và deep learning.

Tạo nguyên mẫu phần mềm: Python chậm hơn khi so sánh với các ngôn ngữ được biên dịch như C++ và Java. Nó có thể không phải là lựa chọn tốt nếu nguồn lực bị giới hạn và yêu cầu về hiệu quả là bắt buộc. Tuy nhiên, Python là ngôn ngữ tuyệt vời để tạo những nguyên mẫu (bản chạy thử - prototype). Ví dụ, bạn có thể sử dụng Pygame (thư viện viết game) để tạo nguyên mẫu game trước. Nếu thích nguyên mẫu đó có thể dùng C++ để viết game thực sự.

Ngôn ngữ tốt để dạy lập trình: Python được nhiều công ty, trường học sử dụng để dạy lập trình cho trẻ em và những người mới lần đầu học lập trình. Bên cạnh những tính năng và khả năng tuyệt vời thì cú pháp đơn giản và dễ sử dụng của nó là lý do chính cho việc này.

* Cú pháp đơn giản:

Lập trình bằng Python rất thú vị. Nó dễ dàng để hiểu và code bằng Python. Tại sao? Cú pháp của Python khá giống với ngôn ngữ tự nhiên, ví dụ như đoạn code dưới đây:

a = 2

b = 3

sum = a + b

print(sum)

Ngay cả khi chưa lập trình bao giờ, bạn có thể dễ dàng đoán được đoạn code này thêm vào hai số a, b, tính tổng và in tổng của chúng.

* Không quá khắt khe:

Bạn không cần xác định kiểu của một biến trong Python, không cần thêm dấu chấm phẩy vào cuối câu lệnh. Python buộc bạn tuân theo những bài tập có sẵn (như chỉ dẫn đúng). Điều nhỏ nhặt này giúp cho việc học Python dễ dàng với người mới hơn rất nhiều.

* Viết code ít hơn:

Python cho phép viết những chương trình có nhiều chức năng tốt hơn với ít dòng code hơn. Bạn có thể tham khảo mã nguồn game Tic-tac-toe (pastebin.com/7LTkj2V5) với giao diện đồ họa và đối thủ máy tính thông minh mà chỉ chưa đến 500 dòng code. Đây chỉ là một ví dụ. Bạn có thể sẽ ngạc nhiên về những gì mà Python có thể làm được khi tìm hiểu sâu hơn về nó.

* Cộng đồng lớn, hỗ trợ tốt.
  + 1. Scrapy:

Scrapy là một thư viện Python giúp ta thu thập dữ liệu từ web. Lấy ví dụ như khi ta làm một web cần lấy thông tin từ các bài báo ở trang web khác thì ta sẽ dùng scrapy để làm điều đó.



Hình 2. 1. Hình ảnh thư viện Scrapy trong Python

*(Nguồn: https://ichi.pro/vi/cach-cao-bang-scrapy-cua-tran-va-danh-bai-captcha-124640164858391*)

Các thành phần của Scrapy như sau:

* Scrapy Engine: có nhiệm vụ kiểm soát dữ liệu giữa các thành phần trong hệ thống và đồng thời sẽ kích hoạt một số sự kiện khi có một hành động nào đó xảy ra.
* Scheduler: Giúp sắp xếp thứ tự các Url mà chúng ta cần lấy dữ liệu về.
* Dowloader: Thực hiện việc download dữ liệu và truyền dữ liệu ấy cho Scrapy Engine.
* Spiders: Đây là lớp mà chúng ta sẽ định nghĩa để thực hiện công việc tách dữ liệu cần thiết khi download về và có thể cung cấp các url mới để thêm vô phần Scheduler.
* Item pipeline: Sau khi dữ liệu được lấy và qua xử lý thì sẽ được đưa vào đây và item pipeline có nhiệm vụ xử lý và lưu vào cơ sở dữ liệu.
* Middlewares: Là các thành phần nằm giữa Scrapy Engine và các thành phần khác. Chúng giúp ta có thể mở rộng khả năng xử lý cho các thành phần.
  + Spider middlewares: Nằm giữa Scrapy Engine và Spiders, chúng xử lý các dữ liệu trước và sau khi đưa vào Spiders.
  + Dowloader middlewares: Nằm giữa Scrapy Engine và Downloader, ở đây chúng xử lý các request được đẩy vào Downloader từ Scrapy Engine và đồng thời xử lý các response từ Downloader.
  + Scheduler moddlewares: Nằm giữa Scrapy Engine và Scheduler để xử lý các request giữa hai thành phần.

Cách hoạt động của Scrapy như sau:

1. Khi muốn lấy dữ liệu từ một website thì Scrapy Engine sẽ yêu cầu Spiders cung cấp danh sách mà người dùng đã nạp.
2. Sau khi Scrapy Engine có được danh sách URL thì sẽ gửi cho Scheduler để sắp xếp thứ tự các url cần tải dữ liệu về.
3. Sau đó Scrapy Engine nhận được nhanh sách các url theo thứ tự từ Scheduler và đưa danh sách ấy vào Dowloader thông qua Dowloader Middleware (theo hướng đầu vào).
4. Dowloader nhận được các url ấy và bắt đầu thực hiện tải dữ liệu về, cứ sau mỗi một url tải xong thì sẽ tạo một response trả về cho Scrapy Engine thông qua Dowloader Middleware (theo hướng đầu ra).
5. Sau khi Scrapy Engine nhận được kết quả trả về thì sẽ đẩy kết quả ấy cho Spiders thông qua Spider Middleware (theo hướng đầu vào) để xử lý kết quả trả về.
6. Spiders sẽ xử lý dữ liệu và trả lại kết quả sau khi xử lý đồng thời sẽ cung cấp các url mới có thể được nạp vào và đẩy lại cho Scrapy Engine thông qua Spider Middleware (theo hướng đầu ra).
7. Scrapy Engine sau khi nhận được kết quả trả về từ Spiders thì sẽ thực hiện hai việc là một là đẩy dữ liệu sau khi được xử lý cho Item Pipeline để xử lý và lưu vào cơ sở dữ liệu, hai là đẩy danh sách các url mới cho Scheduler và thực hiện lại bước thứ 3.
   * 1. SQL:

SQL Server là hệ thống dùng quản lý cơ sở dữ liệu và ngôn ngữ được sử dụng trong SQL Server là SQL. Ngôn ngữ SQL là ngôn ngữ rất dễ sử dụng và ít mắc lỗi. Nó cung cấp các câu lệnh rất đa dạng và hữu ích có thể kết hợp được với nhau. Các câu lệnh của SQL thực hiện các công việc như thêm, xóa, chỉnh sửa dữ liệu có thể với một hoặc nhiều điều kiện.



Hình 2. 2. Hình ảnh Microsoft SQL Server

* + - 1. Lịch sử phát triển:

SQL Server được phát triển từ ngôn ngữ SEQUEL2 bởi IBM theo mô hình Codd tại trung tâm nghiên cứu của IBM ở California, vào những năm 70 cho hệ thống quản trị CSDL lớn.

Đầu tiên, SQL Server được sử dụng trong các ngôn ngữ quản lý CSDL và chạy trên các máy đơn lẻ. Song do sự phát triển nhanh chóng của những nhu cần xây dựng những CSDL lớn theo mô hình khách chủ (Trong mô hình này, toàn bộ CSDL được tập trung trên máy chủ (Server)). Mọi thao tác xử lý dữ liệu được thực hiện trên máy chủ bằng các lệnh SQL, máy trạm chỉ dùng để cập nhật hoặc lấy thông tin từ máy chủ). Ngày nay, trong các ngôn ngữ lập trình bậc cao, đều có sự trợ giúp của SQL. Nhất là trong lĩnh vực phát triển của Internet, ngôn ngữ SQL càng đóng vai trò quan trọng hơn. Nó được sử dụng để nhanh chóng tạo các trang Web động…

SQL đã được viện tiêu chuẩn quốc gia Mỹ (ANSI) và tổ chức tiêu chuẩn quốc tế (ISO) chấp nhận như một ngôn ngữ chuẩn cho CSDL quan hệ. Nhưng cho đến nay chuẩn này chưa đưa ra đủ 100%. Nên các SQL nhúng trong các ngôn ngữ lập trình khác nhau đã được bổ xung mở rộng cho SQL chuẩn cho phù hợp với các ứng dụng của mình. Do vậy có sự khác nhau rõ ràng giữa các SQL.

* + - 1. Đặc điểm và các ấn bản của SQL:
* Đặc điểm:

SQL là ngôn ngữ phi thủ tục, nó không yêu cầu ta cách thức truy nhập CSDL như thế nào. Tất cả các thông báo của SQL đều rất dễ sử dụng và ít khả năng mắc lỗi.

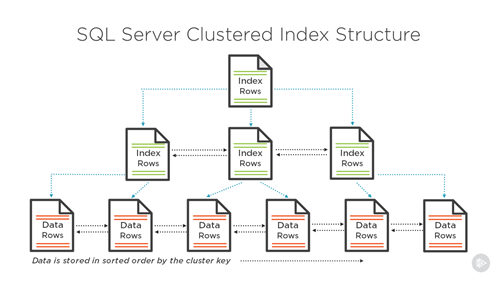
SQL cung cấp tập lệnh phong phú cho các công việc hỏi đáp dữ liệu:

* Chèn, cập nhật, xoá các hàng trong một quan hệ
* Tạo, sửa đổi, thêm và xoá các đối tượng trong của CSDL.
* Điều khiển việc truy nhập tới CSDL và các đối tượng của CSDL để đảm bảo tính bảo mật của CSDL.
* Đảm bảo tính nhất quán và sự ràng buộc của CSDL.

Yêu cầu duy nhất để sử dụng cho các hỏi đáp là phải nắm vững được các cấu trúc CSDL của mình.

* Các ấn bản của SQL Server:
* Enterprise: chứa tất cả cá đặc điểm nổi bật của SQL Server, bao gồm nhân bộ máy CSDL và các dịch vụ đi kèm cùng với các công cụ cho tạo và quản lý phân cụm SQL Server. Nó có thể quản lý các CSDL lớn tới 524 [petabytes](https://en.wikipedia.org/wiki/Petabyte" \o "Petabyte) và đánh địa chỉ 12 [terabytes](https://en.wikipedia.org/wiki/Terabyte" \o "Terabyte) bộ nhớ và hỗ trợ tới 640 bộ vi xử lý(các core của cpu)
* Standard: Rất thích hợp cho các công ty vừa và nhỏ vì giá thành rẻ hơn nhiều so với Enterprise Edition, nhưng lại bị giới hạn một số chức năng cao cấp (advanced features) khác, edition này có thể chạy tốt trên hệ thống lên đến 4 CPU và 2 GB RAM.
* Developer: Có đầy đủ các tính năng của Enterprise Edition nhưng được chế tạo đặc biệt như giới hạn số lượng người kết nối vào Server cùng một lúc…. Ðây là phiên bản sử dụng cho phát triển và kiểm tra ứng dụng. Phiên bản này phù hợp cho các cá nhân, tổ chức xây dựng và kiểm tra ứng dụng. MIỄN PHÍ
* Workgroup: ấn bản SQL Server Workgroup bao gồm chức năng lõi CSDL nhưng không có các dịch vụ đi kèm. Chú ý phiên bản này không còn tồn tại ở SQL Server 2012.
* Express: SQL Server Express dễ sử dụng và quản trị CSDL đơn giản. Được tích hợp với Microsoft Visual Studio, nên dễ dàng để phát triển các ứng dụng dữ liệu, an toàn trong lưu trữ, và nhanh chóng triển khai. SQL Server Express là phiên bản miễn phí, không giới hạn về số cơ ở dữ liệu hoặc người sử dụng, nhưng nó chỉ dùng cho 1 bộ vi xử lý với 1 GB bộ nhớ và 10 GB file CSDL. SQL Server Express là lựa chọn tốt cho những người dùng chỉ cần một phiên bản SQL Server 2005 nhỏ gọn, dùng trên máy chủ có cấu hình thấp, những nhà phát triển ứng dụng không chuyên hay những người yêu thích xây dựng các ứng dụng nhỏ. Miễn phí
  + - 1. Mục đích sử dụng SQL:
* Tạo CSDL.
* Duy trì CSDL.
* Phân tích dữ liệu bằng SSAS – SQL Server Analysis Services.
* Tạo báo cáo bằng SSRS – SQL Server Reporting Services.
* Thực hiện quá trình ETL (Extract-Transform-Load) bằng SSIS – SQL Server Integration Services.
  + - 1. Tìm hiểu thêm về cấu trúc sử dụng SQL Server:

SQL Server cơ bản dựa trên một cấu trúc bảng biểu, bao gồm các dòng kết nối các phần tử [dữ liệu](https://www.semtek.com.vn/du-lieu-la-gi/" \l "Du_lieu_la_gi" \t "_self) có liên quan trong các bảng khác nhau, do đó loại bỏ nhu cầu lưu trữ [dữ liệu](https://www.semtek.com.vn/du-lieu-la-gi/" \l "Du_lieu_la_gi" \t "_self) ở nhiều vị trí trong cơ sở [dữ liệu](https://www.semtek.com.vn/du-lieu-la-gi/" \l "Du_lieu_la_gi" \t "_self). Mô hình quan hệ cũng cung cấp các tham chiếu và ràng buộc toàn vẹn nhằm duy trì độ chính xác của [dữ liệu](https://www.semtek.com.vn/du-lieu-la-gi/" \t "_self); các kiểm tra này là một phần thúc đẩy sự tuân thủ các nguyên tắc về tính nhất quán, tính độc lập và độ tin cậy – được gọi chung là các thuộc tính ACID và được thiết kế để đảm bảo cho các giao dịch cơ sở [dữ liệu](https://www.semtek.com.vn/du-lieu-la-gi/" \t "_self) được xử lý ổn thỏa.



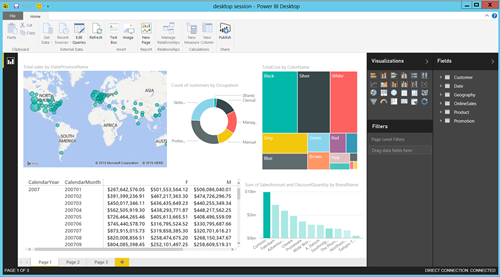
Hình 2. 3. Minh họa cấu trúc sử dụng SQL Server

Thành phần cơ bản của Microsoft SQLServer là SQL Server Database Engine Công cụ này kiểm soát việc lưu trữ, xử lý và bảo mật [dữ liệu](https://www.semtek.com.vn/du-lieu-la-gi/" \t "_self). Thành phần này bao gồm một công cụ quan hệ có chức năng xử lý các lệnh và truy vấn, một công cụ lưu trữ quản lý các tệp, bảng, trang, chỉ mục, bộ đệm dữ liệu và giao dịch [CSDL](https://www.semtek.com.vn/co-so-du-lieu/" \t "_self). Các nhiệm vụ, trigger, trình xem và các đối tượng dữ liệu lưu trữ khác cũng được Database Engine khởi tạo và xử lý.

Lớp phía dưới Database Engine là Hệ điều hành SQL Server – viết tắt SQLOS. Hệ điều hành xử lý các chức năng ở cấp độ thấp hơn như quản lý bộ nhớ và I/O, lên lịch nhiệm vụ và khóa dữ liệu để tránh các xung đột xảy ra khi update. Một lớp giao diện mạng nằm trên lớp Database Engine và sử dụng một giao thức gọi là Tabular Data Stream của Microsoft để các yêu cầu và phản hồi tương tác với máy chủ [CSDL](https://www.semtek.com.vn/co-so-du-lieu/" \t "_self) thuận tiện hơn. Ở cấp độ user, SQLServer DBAs và developers viết các câu lệnh T-SQL để xây dựng và sửa đổi cấu trúc [CSDL](https://www.semtek.com.vn/co-so-du-lieu/" \t "_self), thao tác, thiết lập các bảo vệ, sao lưu CSDL, cùng với nhiều nhiệm vụ khác

* + - 1. Các dịch vụ, công cụ và phiên bản SQL Server:
* Tính năng doanh nghiệp chuyên biệt

Microsoft hiện nay đã có thêm các tính năng quản lý dữ liệu đa dạng, tính năng dành cho doanh nghiệp (business intelligence – BI), và các công cụ phân tích SQL Server. Bên cạnh các dịch vụ Machine Learning mới được tích hợp lần đầu tiên trong phiên bản SQLServer 2016, các dịch vụ phân tích dữ liệu bao gồm SQL Server Analysis Services, công cụ phân tích xử lý dữ liệu sử dụng trong BI, các ứng dụng trực quan hóa dữ liệu và các dịch vụ SQL Server Reporting, hỗ trợ tạo và [phân phối](https://www.semtek.com.vn/chien-luoc-phan-phoi/" \t "_self) các báo cáo BI.



Hình 2. 4. Minh họa tính năng doanh nghiệp SQL Server

* Tính năng quản lý:

Về phương diện quản trị, Microsoft SQL Server gồm có các dịch vụ tích hợp SQLServer, dịch vụ SQL Server Data Quality và dịch vụ SQL Server Master. Hai bộ công cụ dành riêng cho DBAs và developer: SQLServer Data Tools, sử dụng trong việc phát triển CSDL và SQL Server Management Studio, nhằm mục đích triển khai, giám sát và quản lý CSDL.

* Các phiên bản phổ biến của Microsoft SQL Server:

Microsoft hiện phát hành bốn phiên bản SQL Server chính dựa trên cấp độ khác nhau của các dịch vụ đi kèm. Trong đó hai phiên bản miễn phí: một phiên bản bao gồm đầy đủ các tính năng cho phép phát triển và thử nghiệm CSDL, phiên bản Express có thể sử dụng để chạy các CSDL nhỏ với dung lượng lưu trữ lên đến 10 GB.

Đối với các ứng dụng có dung lượng lớn hơn, Microsoft phát hành một phiên bản Enterprise mất phí bao gồm tất cả các tính năng của SQLServer, cùng với một phiên bản Standard sở hữu một phần hệ tính năng, giới hạn số lượng lõi bộ xử lý và kích thước bộ nhớ khi cấu hình máy chủ CSDL của người dùng.

* + 1. Web - Django:

Django là một trong số những [web framework](https://bizfly.vn/techblog/web-framework-la-gi.html) bậc cao miễn phí, là mã nguồn mở được tạo ra bởi ngôn ngữ Python dựa trên mô hình mô hình MTV (gồm Model-Template-Views). Hiện framework này được phát triển, quản lý bởi Django Software Foundation.

Django ra đời với mục tiêu hỗ trợ thiết kế các website phức tạp dựa trên những CSDL có sẵn. Nó hoạt động dựa theo nguyên lý ‘cắm’ các thành phần và và tái sử dụng để tạo nên các website với ít code, ít khớp nối, có khả năng phát triển và không bị trùng lặp.



Hình 2. 5. Hình ảnh Django Python

* + - 1. Ưu điểm:

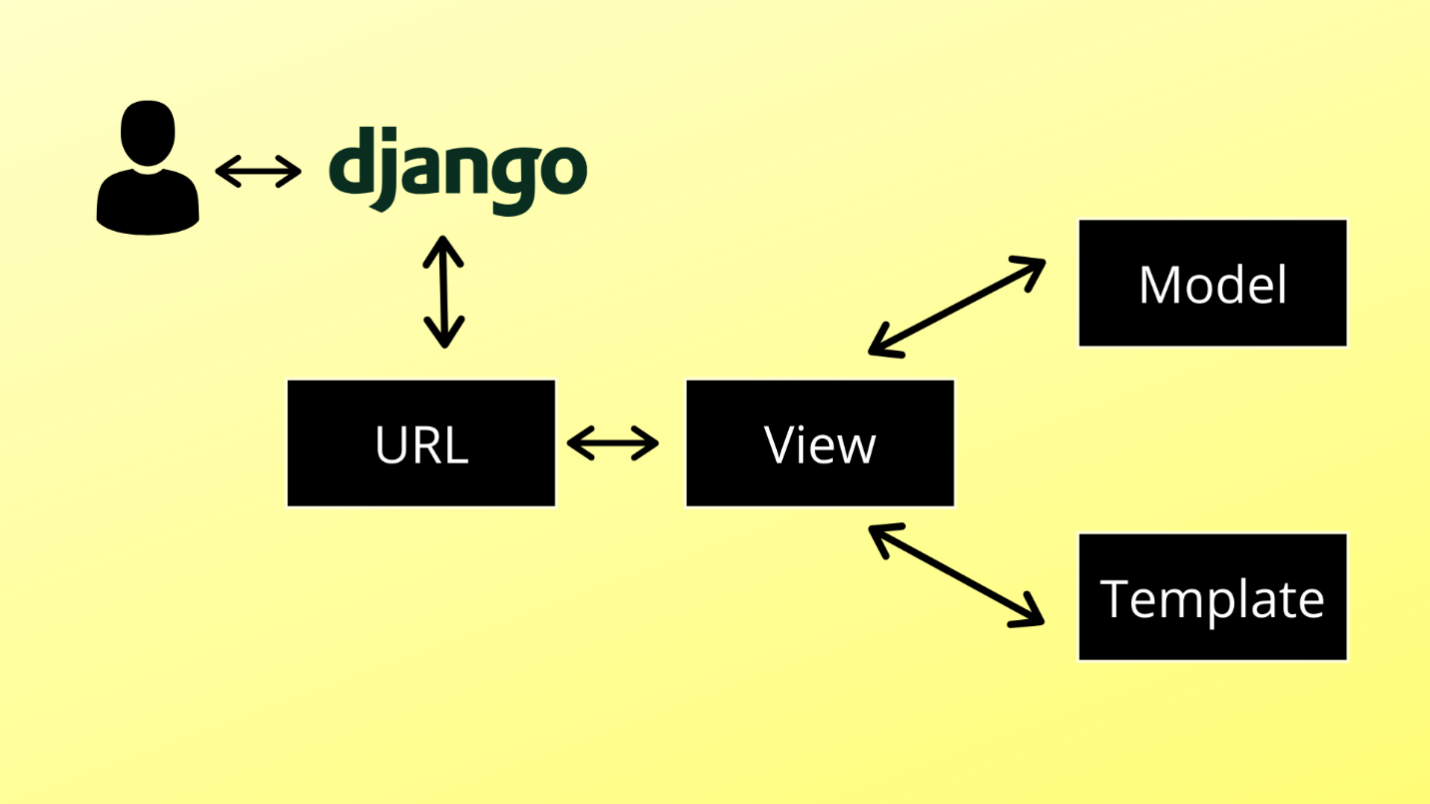
Lợi thế hàng đầu của Django là khả năng thiết kế, tạo lập website và các ứng dụng nhanh chóng đến bất ngờ. Ngoài ra, dưới đây Bizfly liệt kê những điểm cộng khiến framework này trở nên nổi bật và được lập trình viên sử dụng rộng rãi.

* Bảo mật tốt: Làm việc với Django, các lập trình viên gần như không có không gian để phạm bất cứ sai lầm về an ninh nào. Nó giúp các developer tránh được tất cả các lỗi thường gặp như nhấp chuột, kịch bản chéo trang, SQL tiêm, giả mạo yêu cầu, ... Nhờ đó, sản phẩm được tạo bởi framework này có khả năng bảo mật cực tốt.
* Mở rộng thỏa thích: Django có sẵn tính năng mở rộng nhằm hỗ trợ các lập trình viên quản lý lưu lượng người truy cập, thích hợp với các trang có traffic lớn.
* Dễ sử dụng: Django được tạo ra bởi ngôn ngữ lập trình [Python](https://bizfly.vn/techblog/python-la-gi.html) và mô hình MVC nên rất dễ ứng dụng trong các dự án. Đa ngôn ngữ và được hỗ trợ Multi-Site
* Dễ học: Có nhiều tài liệu hỗ trợ việc học Django, bao gồm cả tài liệu miễn phí trên mạng và sách in. Cộng đồng sử dụng Django hiện nay đang phát triển mạnh mẽ, newbie có thể tìm kiếm sự giúp đỡ trên các trang facebook, diễn đàn, blog, ...
  + - 1. Nhược điểm:

Bên cạnh những ưu điểm kể trên, Django cũng tồn tại một số nhược điểm mà bạn cần cân nhắc trước khi sử dụng chúng. Cụ thể sau đây:

* Có thể xảy ra một vài vấn đề khi phát triển các ứng dụng/website quy mô nhỏ.
* Định tuyến tương đối khó
* Không đưa ra các cảnh báo khi xuất hiện lỗi trong mẫu.
  + - 1. Mô hình MTV (Model - Template - View):

Django sử dụng mô hình MTV tương tự như mô hình MVC (Model-View-Controller) trong các framework khác



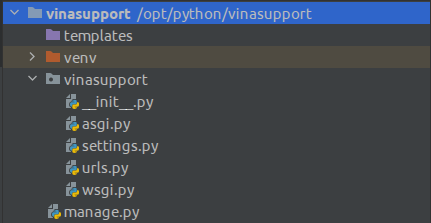
Hình 2. 6. Hình ảnh mô hình MTV (Model - Template - View) Django

| Thành phần | Mô tả |
| --- | --- |
| Model | Nơi thiết kế ra những table cho database, từ đó Django ORM đã cung cấp những phương thức xử lý, nghiệp vụ lên database |
| Template | Là những template được thiết kế ra và xử lý output ra mã html/css cho trang web |
| Views | Các function để xử lý khi có request từ người dùng |

Bảng 1. 2. Mô tả mô hình MTV (Model - Templat - View) Django

* + - 1. Cấu trúc thư mục của một project Django:

Cấu trúc thư mục ban đầu giống như là:



Hình 2. 7. Ví dụ cấu trúc thư mục của một project Django

Trong đó

\_\_init\_\_.py: Nói với trình thông dịch python là thư mục nên được coi là một python package. Tập tin này chủ yếu là trống.

asgi.py: Cho phép các máy chủ web tương thích ASGI phục vụ dự án của bạn (chỉ Django 3).

settings.py: Tập tin cấu hình

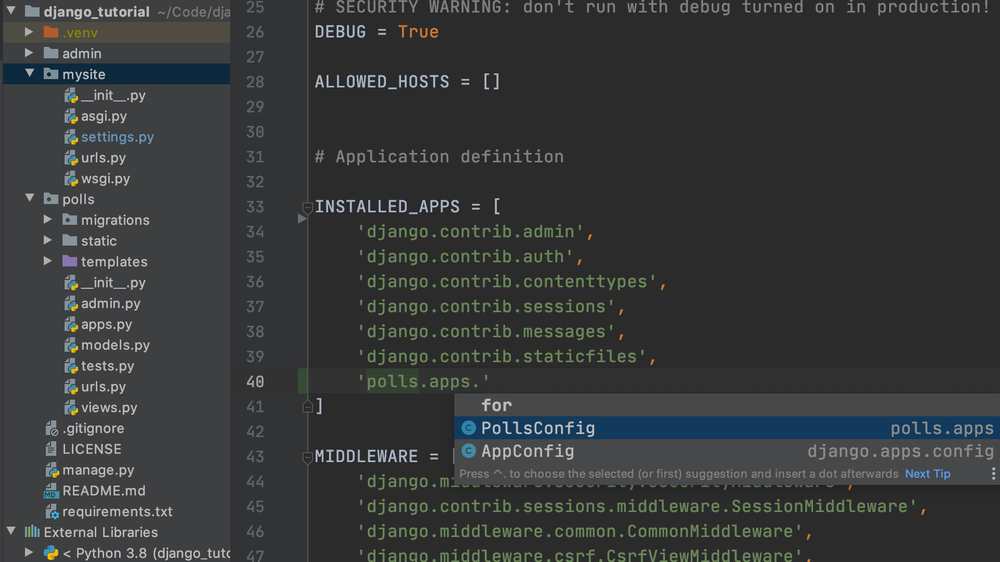
urls.py: Bao gồm tất cả khai báo URL đến các views để xử lý request

wsgi.py: Đây là lối vào cho các máy chủ web tương thích WSGI

manage.py: Cho phép bạn quản lý, tương tác với Django bằng dòng lệnh

* + - 1. Lí do nên chọn Django:

Đây là câu hỏi của khá nhiều newbie khi bắt đầu tìm hiểu Django là gì. Có nhiều nguyên nhân khiến framework này được ưu ái. Có thể kể đến vài lợi ích sau.



Hình 2. 8. Minh họa project sử dụng Django

Nhanh chóng, đơn giản trong lập trình web/ứng dụng

Đa dạng về tài liệu

Sử dụng nguyên tắc DRY - không trùng lặp

Hỗ trợ ORM, mang lại sự tương thích, khả năng hoạt động với các CSDL thường gặp như Oracle và SQL

Sở hữu nhiều tính năng, giải pháp độc đáo, hữu ích như GPS, chatbot, ứng dụng CSDL, nguồn cung cấp dữ liệu RSS,... giúp phát triển, [quản trị website](https://bizfly.vn/techblog/quan-tri-website-la-gi.html), chăm sóc khách hàng,...

Mang lại khả năng sửa đổi (thêm, xóa) một cách liền mạch nhiều thành phần

Thân thiện với Seo

Linh hoạt cao: Xây dựng nhiều dạng website, ứng dụng cần thiết cho mọi lĩnh vực, tổ chức, cơ quan.

* + - 1. Các nền tảng sử dụng Django:

Các nền tảng nổi bật đang sử dụng Django là vấn đề cuối cùng bạn cần biết khi tìm hiểu Django là gì. Framework này được ứng dụng để tạo lập, phát triển và hỗ trợ rất nhiều nền tảng lớn như Instagram, Spotify, YouTube, Dropbox, trình duyệt Mozilla, Disqus, …

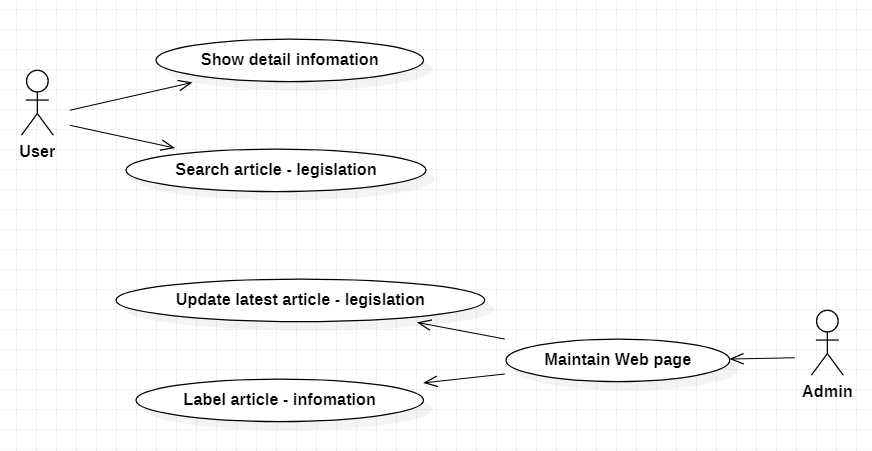
Trong những nền tảng trên, Django giữ những vai trò khác nhau, khi thì hỗ trợ quản lý và xử lý dữ liệu, khi thì lưu trữ thông tin hoặc cũng có lúc hỗ trợ triển khai các thành phần mới, … Với các nền tảng lớn, nhiều người truy cập thì nó là framework lý tưởng để tăng hiệu suất làm việc.

Những ‘ông lớn’ trong thế giới số đều đã nắm bắt cơ hội để sử dụng ứng dụng Django. Bởi một website lớn được thiết kế dưới sự hỗ trợ của framework này sẽ mang lại hiệu suất hoạt động vô cùng mạnh mẽ. Tìm hiểu Django là gì và cân nhắc sử dụng nó cho website mới của bạn ngay bây giờ.

* 1. PySpark - kỹ thuật song song:

CHƯƠNG 3 – THỰC NGHIỆM

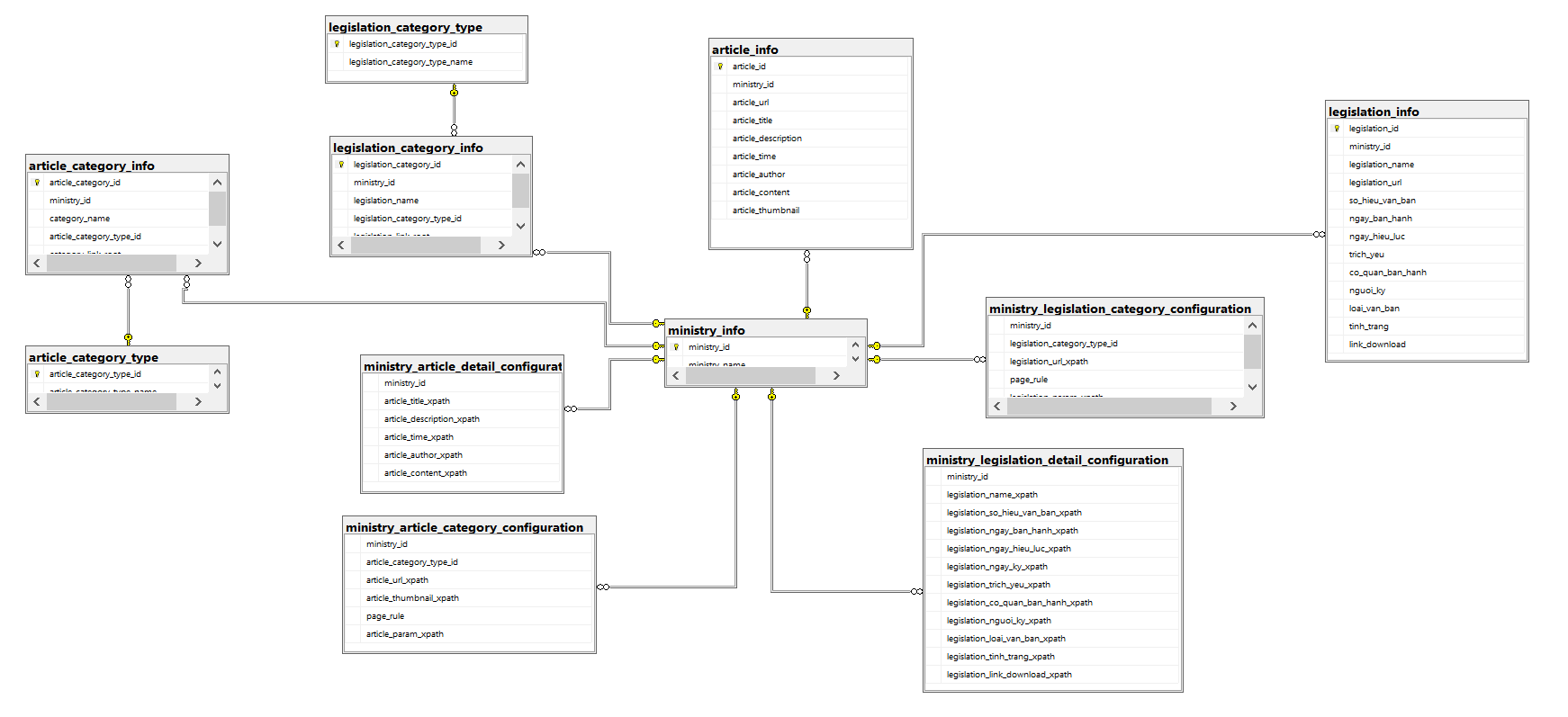
* 1. Các sơ đồ đặc tả hệ thống:
     1. Sơ đồ usecase:



Hình 3. 1. Sơ đồ usecase

* + 1. Sơ đồ tuần tự:
    2. Sơ đồ lớp:
    3. Sơ đồ ERD:
  1. Phân tích cơ sở dữ liệu:

Tổng quan cơ sở dữ liệu:



Hình 3. 2. Tổng quan cơ sở dữ liệu

* + 1. Bảng ministry\_info:

Bảng lưu thông tin của các bộ ngành:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | ministry\_id | int | Id của mỗi bộ |
|  | ministry\_name | nvarchar(100) | Tên của các bộ ngành |

Bảng 3. 1 Bảng ministry\_info

* + 1. Bảng article\_category\_type:

Bảng lưu thông tin các loại tin tức.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | article\_category\_type\_id | int | Id các loại tin tức |
|  | article\_category\_type\_name | nvarchar(100) | Tên của các mục tin tức được phân loại |

Bảng 3. 2. Bảng article\_category\_type

* + 1. Bảng legislation\_category\_type:

Bảng lưu thông tin các loại văn bản.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | legislation\_category\_type\_id | int | Id các loại văn bản |
|  | legislation\_category\_type\_name | nvarchar(100) | Tên các loại văn bản |

Bảng 3. 3. Bảng legislation\_category\_type

* + 1. Bảng article\_category\_info:

Bảng lưu thông tin chi tiết các loại tin tức

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | article\_category\_id | int | Id các loại tin tức của từng bộ - ngành |
| FPK | ministry\_id | int | Id các bộ ngành |
|  | category\_name | Nvarchar(200) | Tên các loại tin tức của từng bộ - ngành |
|  | article\_category\_type\_id | Int | Id các loại tin tức |
|  | category\_link\_root | Varchar(1000) | Link truy cập đến danh sách hiển thị của từng loại tin tức |

Bảng 3. 4. Bảng article\_category\_info

* + 1. Bảng legislation\_category\_info:

Bảng lưu thông tin chi tiết các loại tin tức

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | legislation\_category\_id | int | Id các loại văn bản của từng bộ - ngành |
| FPK | ministry\_id | int | Id các bộ ngành |
|  | legislation\_name | Nvarchar(200) | Tên các loại văn bản theo từng bộ - ngành |
|  | legislation\_category\_type\_id | Int | Id các loại văn bản |
|  | legislation\_link\_root | Varchar(1000) | Link truy cập đến danh sách hiển thị của từng loại vản bản |

Bảng 3. 5. legislation\_category\_info

* + 1. Bảng ministry\_article\_category\_configuration:

Bảng lưu thông tin cấu hình các loại tin tức

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| FPK | ministry\_id | int | Id các bộ ngành |
|  | article\_category\_type\_id | Int | Id các loại tin tức |
|  | article\_url\_xpath | Varchar(200) | Xpath của đường link dẫn đến chi tiết bài báo |
|  | article\_thumbnail\_xpath | Varchar(200) | Xpath của ảnh thumbnail |
|  | page\_rule | int |  |
|  | article\_param\_xpath | Varchar(200) |  |

Bảng 3. 6. Bảng ministry\_article\_category\_configuration

* + 1. Bảng ministry\_article\_detail\_configuration:

Bảng lưu thông tin cấu hình các chi tiết của bài báo:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| FPK | ministry\_id | int | Id các bộ ngành |
|  | article\_tittle\_xpath | Varchar(200) | Xpath tiêu đề bài báo |
|  | article\_description\_xpath | Varchar(200) | Xpath mô tả bài báo |
|  | article\_time\_xpath | Varchar(200) | Xpath thời gian đăng tải bài báo |
|  | article\_author\_xpath | Varchar(200) | Xpath của tác giả viết bài |
|  | article\_content\_xpath | Varchar(200) | Xpath nội dung của bài báo |

Bảng 3. 7. Bảng ministry\_article\_detail\_configuration

* + 1. Bảng ministry\_legislation\_category\_configuration:

Bảng lưu thông tin cấu hình của văn bản:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| FPK | ministry\_id | int | Id các bộ ngành |
|  | legislation\_category\_id | int | Id các loại văn bản của từng bộ - ngành |
|  | legislation\_url\_xpath | Varchar(200) | Xpath dẫn đến văn bản chi tiết |
|  | page\_rule | int |  |
|  | legislation\_param\_xpath | Varchar(200) |  |

Bảng 3. 8. Bảng ministry\_legislation\_category\_configuration

* + 1. Bảng ministry\_legislation\_detail\_configuration:

Bảng lưu thông tin cấu hình các chi tiết của một văn bản:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| FPK | ministry\_id | int | Id các bộ ngành |
|  | legislation\_name\_xpath | Varchar(200) | Xpath tên văn bản |
|  | legislation\_so\_hieu\_van\_ban\_xpath | Varchar(200) | Xpath số hiệu văn bản |
|  | legislation\_ngay\_hieu\_luc\_xpath | Varchar(200) | Xpath ngày hiệu lực |
|  | legislation\_ngay\_ky\_xpath | Varchar(200) | Xpath ngày ký |
|  | legislation\_trich\_yeu\_xpath | Varchar(200) | Xpath trích yếu văn bản |
|  | legislation\_co\_quan\_ban\_hanh\_xpath | Varchar(200) | Xpath cơ quan ban hành văn bản |
|  | legislation\_nguoi\_ky\_xpath | Varchar(200) | Xpath người ký văn bản |
|  | legislation\_loai\_van\_ban\_xpath | Varchar(200) | Xpath tên loại văn bản |
|  | legislation\_tinh\_trang\_xpath | Varchar(200) | Xpath tình trạng hiệu lực văn bản |
|  | legislation\_link\_download\_xpath | Varchar(200) | Xpath link download văn bản |

Bảng 3. 9.Bảng ministry\_legislation\_detail\_configuration

* + 1. Bảng article\_info:

Bảng để lưu trữ thông tin tin tức sau khi crawl:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | article\_id | int | Id của bài báo |
| FPK | ministry\_id | int | Id của bộ - ngành |
|  | article\_url | Varchar(1000) | Url của bài báo |
|  | article\_title | Nvarchar(200) | Tiêu đề bài báo |
|  | article\_description | Nvarchar(500) | Mô tải bài báo |
|  | article\_time | Varchar(10) | Thời gian đăng bài báo |
|  | article\_author | Nvarchar(100) | Tác giả bài báo |
|  | article\_content | Nvarchar(max) | Nội dung bài báo |
|  | article\_thumbnail | Varchar(100) | Ảnh thumbnail bài báo |

Bảng 3. 10. Bảng ministry\_legislation\_detail\_configuration

* + 1. Bảng legislation\_info:

Bảng để lưu trữ thông tin văn bản sau khi crawl:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Mô tả |
| PK | legislation\_id | int | Id của văn bản |
| FPK | ministry\_id | int | Id của bộ - ngành |
|  | legislation\_name | Nvarchar(500) | Tên văn bản |
|  | legislation\_url | Varchar(500) | Link dẫn đến chi tiết văn bản |
|  | so\_hieu\_van\_ban | Varchar(20) | Số hiệu văn bản |
|  | ngay\_ban\_hanh | Varchar(100) | Ngày ban hành văn bản |
|  | ngay\_hieu\_luc | Nvarchar(100) | Ngày văn bản có hiệu lực |
|  | trich\_yeu | Nvarchar(500) | Trích yếu của văn bản |
|  | co\_quan\_ban\_hanh | Nvarchar(100) | Cơ quan ban hành văn bản |
|  | nguoi\_ky | Nvarchar(50) | Người ký văn bản |
|  | loai\_van\_ban | Nvarchar(50) | Thể loại văn bản |
|  | tinh\_trang | Nvarchar(50) | Tình trạng hiệu lực của văn bản |
|  | link\_download | Nvarchar(300) | Link tải văn bản |

Bảng 3. 11. Bảng legislation\_info

**TÀI LIỆU THAM KHẢO**

**Tiếng Việt**

1. <https://bizfly.vn/techblog/django-la-gi.html>
2. <https://vinasupport.com/django-la-gi-tong-quan-ve-django-framework/>
3. <https://www.semtek.com.vn/sql-server-2/>
4. <https://help.pacisoft.com/knowledgebase/sql-server-la-gi-sql-server-dung-lam-gi-co-may-phien-ban/>
5. <https://www.semtek.com.vn/sql-server-2/>
6. <https://viblo.asia/p/gioi-thieuhuong-dan-ve-crawler-voi-scrapy-framework-ByEZkWoEZQ0>
7. https://doc.scrapy.org/en/0.10.3/topics/architecture.html

**PHỤ LỤC**

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung luận văn như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của luận văn