

## BÁO CÁO MÔ TẢ SOURCE CODE

### LỊCH SỬ THAY ĐỔI TÀI LIỆU

Ngày thay đổi	Phiên bản	Mô tả	Tác giả/Nhóm tác giả
30/12/2020	V1.0	Tạo mới	Hoang
07/01/2021	V1.1	Thêm mô tả phần classification	Hoang
10/01/2021	V1.2	Bổ sung cho zingnews và kênh14	Hoang
11/01/2021	V1.3	Bổ sung mô tả chạy tự động + Finalize	Hoang

## Mục lục

1. Cấu trúc thư mục.....	3
2. Thư mục news_system (Mã nguồn chính).....	3
2.1. Cấu trúc chi tiết thư mục.....	3
2.2. Mô tả.....	5
3. Thư mục web-demo .....	6
3.1. Cấu trúc thư mục.....	6
3.2. Mô tả.....	7

## 1. Cấu trúc thư mục

Đường dẫn: <https://bitbucket.org/pinetreevietnam/ai-news>

Nhánh develop: dev, web-crawler

Nhánh kiểm thử trước khi deploy trên server: web-crawler-server

Nhánh đã deploy trên server: web-crawler-server-2

Mã nguồn gồm 3 phần chính:

- docs: Gồm các mô tả ban đầu của từng phần; hướng dẫn và quy chuẩn chung cho các luồng xử lý.
- news\_system: Chứa mã nguồn của các phần chính gồm: thu thập tin tức (Hoàn thành), tóm tắt tin tức (Hoàn thành), phân loại tin tức (Hoàn thành), sắp xếp tin tức (Đang cập nhật).
- web-demo: Chứa mã nguồn của web hiển thị tin tức (Đang cập nhật).

## 2. Thư mục news\_system (Mã nguồn chính)

### 2.1. Cấu trúc chi tiết thư mục

\*\*\*\*\* Directory Tree \*\*\*\*\*

```
news_system/
├── classifier/
│   ├── __init__.py
│   ├── classifier.py
│   ├── README.md
│   ├── test_classifier.py
│   └── utils.py
├── clustering/
│   └── __init__.py
├── crawler/
│   ├── __init__.py
│   ├── docs/
│   │   └── Crawling_scroll_webpage_tutorial.pdf
│   ├── README.md
│   ├── requirements.txt
│   ├── scraper/
│   │   ├── __init__.py
│   │   ├── classifier/
│   │   │   ├── __init__.py
│   │   │   ├── classifier.py
│   │   │   ├── models/
│   │   │   ├── README.md
│   │   │   ├── test_classifier.py
│   │   │   └── utils.py
│   │   ├── debug.log
│   │   ├── international_crawler.sh
│   │   ├── m1_banking.sh
│   │   ├── m2_real_estate.sh
│   │   ├── m3_energy.sh
│   │   ├── m4_vingroup.sh
│   │   ├── m5_retails.sh
│   │   ├── m6_others.sh
│   │   ├── macro_crawler.sh
│   │   └── news/
│   └── news/
```



```

├── __init__.py
├── requirements.txt
├── vi_stopwords.txt
├── ViTextRank.py
└── textrank_.py

```

## 2.2. Mô tả

### a) Mã nguồn thu thập tin tức (thư mục: **crawler**)

Các file thực thi để tự động crawl tin tức từ các nguồn khác nhau gồm:

- Daily news: macro\_crawler.sh, international\_crawler.sh, vnstock\_crawler.sh.
- Morning brief: m1\_banking.sh, m2\_real\_estate.sh, m3\_energy.sh, m4\_vingroup.sh, m5\_retails.sh, m6\_others.sh, m7\_socialtrend.sh.

Các file được lập lịch cứ mỗi 2 giờ tự động thực thi. Để tránh trường hợp bị vượt quá số tiến trình giới hạn của CPU, các file được thực thi theo từng khoảng thời gian khác nhau và chạy không quá 3 đoạn mã thực thi cùng lúc. Ví dụ: file m1\_banking.sh được thực thi lúc 0h00, file m2\_real\_estate.sh được thực thi lúc 0h20, ... Các dòng lệnh trong từng file được thực thi trong vòng 5 phút.

Mã nguồn thu thập tin tức theo từng website (thư mục con: **scraper/news/spiders**)

Bao gồm 17 file .py cho 17 website khác nhau. Công nghệ sử dụng:

STT	Thư viện	Mô tả
1	scrapy ( <a href="https://scrapy.org/">https://scrapy.org/</a> )	Scrapy là một khung thu thập dữ liệu web miễn phí và mã nguồn mở được viết bằng Python. Được thiết kế ban đầu để quét web, nó cũng có thể được sử dụng để trích xuất dữ liệu bằng API hoặc làm trình thu thập dữ liệu web đa năng.
2	selenium ( <a href="https://www.selenium.dev/">https://www.selenium.dev/</a> )	Selenium là một khung di động để thử nghiệm các ứng dụng web. Selenium cung cấp một công cụ phát lại để soạn thảo các bài kiểm tra chức năng mà không cần phải học ngôn ngữ kịch bản kiểm tra.
3	chromedriver ( <a href="https://chromedriver.chromium.org/">https://chromedriver.chromium.org/</a> )	ChromeDriver là một máy chủ độc lập, nó giúp thực hiện truyền tải các giao thức xử lý của WebDriver tới trình duyệt Chrome.
4	pyvirtualdisplay ( <a href="https://github.com/ponty/pyvirtualdisplay">https://github.com/ponty/pyvirtualdisplay</a> )	Giả lập giao diện trên cho hệ điều hành linux server
5	pymongo ( <a href="https://github.com/mongodb/mongo-python-driver">https://github.com/mongodb/mongo-python-driver</a> )	Thư viện hỗ trợ kết nối với cơ sở dữ liệu MongoDB

### b) Mã nguồn tóm tắt tin tức (thư mục: **summarizer**)

Bao gồm 2 thuật toán: Phân cụm kết hợp với BERT và Thuật toán TextRank.

Thư mục *summarizer*: Lưu mã nguồn thuật toán tóm tắt bằng phân cụm. Các file/thư mục chính:

- BertParent.py, ClusterFeatures.py, model\_processors.py: Mã nguồn của thuật toán phân cụm
- vncorenlp: Thư mục chứa model của thuật toán tokenizer (tách từ) cho tiếng Việt.

Thư mục *textrank*: Lưu mã nguồn thuật toán tóm tắt bằng TextRank. Các file chính:

- vi\_stopwords.txt: Lưu stopwords (các từ không có ý nghĩa) cho tiếng Việt – phục vụ tiền xử lý.
- ViTextRank.py: Mã nguồn thuật toán TextRank.

c) Mã nguồn phân loại tin tức (thư mục: *classifier*)

Sử dụng model BERT cho biểu diễn tiếng Việt.

Các file chính:

- classifier.py: Lưu class cho việc training/testing/predicting loại của 1 tin tức.
- test\_classifier.py: Thống kê/đánh giá kết quả trên bộ dữ liệu của nhóm.
- utils.py: Mã nguồn cho các hàm tiền xử lý tin tức.

Model: Bao gồm 2 model cho bản tin daily news và morning brief. Do model có dung lượng lớn nên sẽ được tải về trực tiếp từ server chứ không lưu trữ tại bitbucket.

Đường dẫn: [MODEL CLASSIFIER - Google Drive](#)

Công nghệ chính được sử dụng cho phần tóm tắt và phân loại gồm:

- [VinAIRsearch/PhoBERT: PhoBERT: Pre-trained language models for Vietnamese \(EMNLP-2020 Findings\) \(github.com\)](#)
- [vncorenlp/VnCoreNLP: A Vietnamese natural language processing toolkit \(NAACL 2018\) \(github.com\)](#)

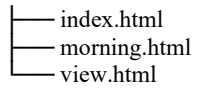
Hai mã nguồn cho tóm tắt và phân loại tin tức đã được tích hợp và chạy theo thời gian thực trong phần thu thập tin tức.

### 3. Thư mục web-demo

#### 3.1. Cấu trúc thư mục

\*\*\*\*\* Directory Tree \*\*\*\*\*

```
web-demo/
├── dummy_data.csv
├── main.py
├── static/
│   ├── script.js
│   └── style.css
└── templates/
```



```
graph LR; A[ ] --- B[index.html]; A --- C[morning.html]; A --- D[view.html];
```

### **3.2. Mô tả**

Các file/thư mục chính:

- main.py: Mã nguồn xử lý hiển thị dữ liệu trên website.
- static/: Thư mục lưu các hàm bổ sung và style của website.
- templates/: Thư mục lưu thiết kế cơ bản của website.