|  |  |
| --- | --- |
| **CÔNG TY**  **TNHH MTV MY JOB365** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập - Tự do - Hạnh phúc**  *Hà Nội, ngày 16 tháng 7 năm 2025* |

**BÁO CÁO BÀN GIAO CODE**

Bài toán phân tích key

Kính Gửi: Ban giám đốc công ty TNHH MTV MY JOB 365

Tôi tên là: Nguyễn Mạnh Cường

Phòng ban: Phòng 4 – Phòng AI

Số điện thoại liên hệ: 0329144716

Tôi xin báo cáo bàn giao code cụ thể như sau:

1. Bài toán yêu cầu:

- Vào mục blog của đối thủ ví dụ: https://www.topcv.vn/blog

- Bước 1: Quét Tách hết các URL ra thành 1 file tức danh sách các bài viết

- Bước 2: Từ danh sách các bài viết đó lọc ra danh sách các tile

- Bước 3: Từ các key title đó lọc ra key chính, tức chủ đề của bài viết, tách ra key trọng tâm. Ví dụ:

https://www.topcv.vn/nghe-it-la-gi

Tile: Ngành IT là gì? Mô tả chi tiết công việc của ngành IT

Key chính: Công việc của ngành IT

https://www.topcv.vn/cau-hoi-phong-van-tieng-trung

Tile: Top 18 câu hỏi phỏng vấn tiếng Trung thường gặp hiện nay

Key chính (chủ đề chính): câu hỏi phỏng vấn tiếng Trung

- Bước 4: Lọc trùng key, xem key đã được cho timviec365.vn chưa có 2 cách

Đọc các bài blog trên site timviec365 để xem chủ đề được viết chưa, API lấy chỗ Bình.

-Bước 5:

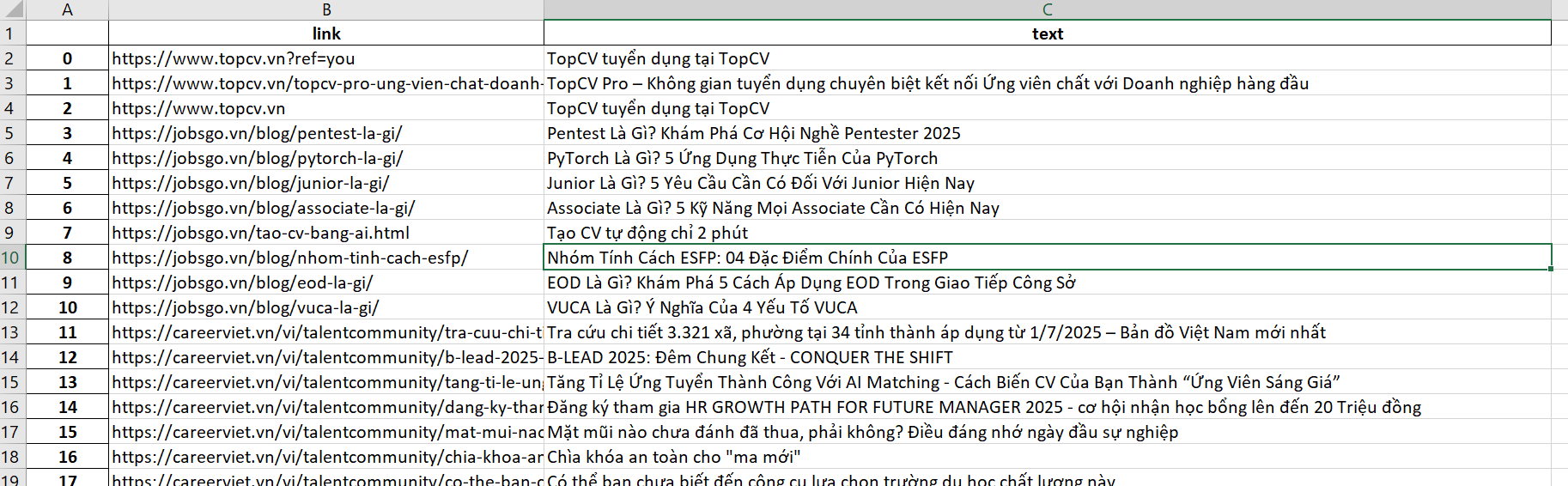
Tách các chủ để chưa được viết thành 1 file gửi đ.c Phượng

- Chú ý quan trọng: Bài toán đã bỏ bước trích xuất key từ title, chỉ lấy những title có chủ đề chưa được viết, xuất ra file

- Input



- Outut



1. Thuật toán hướng xử lý

+Tiền xử lý: lấy toàn bộ các link, titles của công ty lưu vào file blog\_list.xlsx (cụ thể ở III. Diễn giải code)

+Tải model để xử lý độ trùng lặp

- Bước 1: Sử dụng selenium + google chrome để lấy các dữ liệu trên web -> raw\_titles.

- Bước 2: Gom toàn bộ dữ liệu thô (raw\_titles) của từng trang vào làm 1 -> all\_raw\_titles.

- Bước 3: Sử dụng mô hình (model) so sánh ngữ nghĩa câu (cụ thể ở III. Diễn giải code) để so sánh độ tương đồng của từng raw\_title với titles của công ty, loại những câu có độ tương đồng cao

- Bước 4: Xuất file excel các link và title độ tương đồng thấp

*Chú ý: có sơ đồ khối càng tốt với bài toán yêu cầu phức tạp.*

1. Diễn giải code

- Đảm bảo cây thư mục có dạng như sau (Nếu thay đổi thì sửa lại các đường dẫn):

./

│ .env

│ .git/

│ .gitattributes

│ .gitignore

│ chromedriver-win64/

│ │ chromedriver.exe

│ │ LICENSE.chromedriver

│ │ THIRD\_PARTY\_NOTICES.chromedriver

│ database/

│ │ blog\_list.xlsx

│ │ blog\_list\_new.xlsx

│ driver/

│ input.txt

│ main.py

│ model/

│ │ paraphrase-multilingual-MiniLM-L12-v2/

│ output.xlsx

│ print\_tree.py

│ README.md

│ requirements.txt

│ source/

│ │ crawl\_title/

│ │ │ crawl\_title\_from\_link.py

│ │ │ filter\_title.py

│ │ │ selenium\_driver.py

│ │ │ \_\_init\_\_.py

│ │ database/

│ │ │ get\_title\_database.py

│ │ │ \_\_init\_\_.py

│ │ download\_model/

│ │ │ downloadModel.py

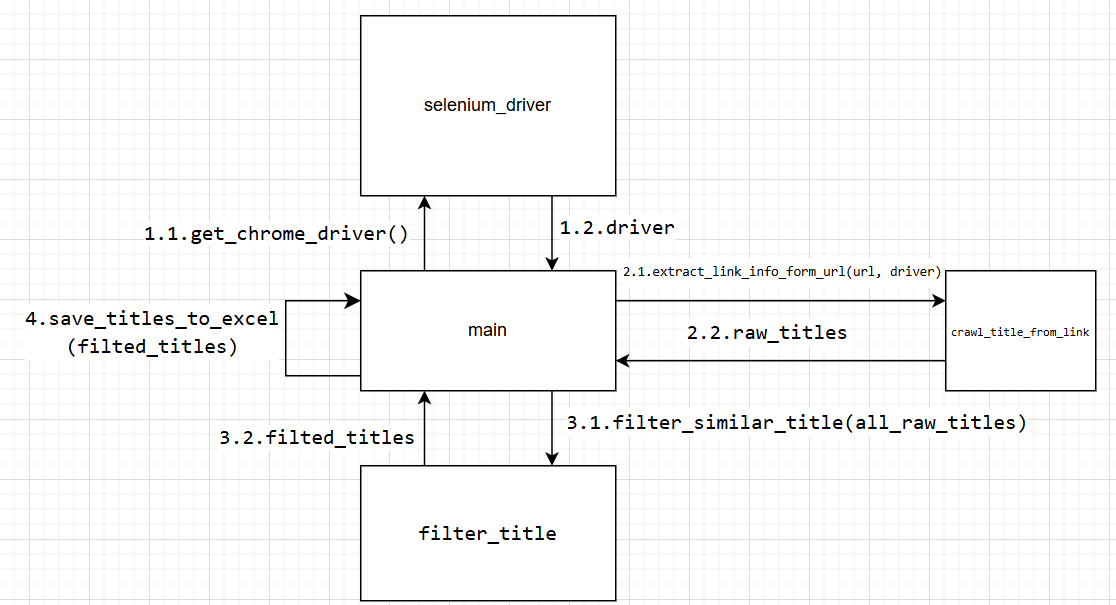
│ │ │ \_\_init\_\_.py

│ │ \_\_init\_\_.py

│ venv/

-

-Sơ đồ giao tiếp (Communication diagram)



Chú ý:

- .//source//crawl\_title//selenium\_driver.py : chứa hàm get\_chrome\_driver() trả về driver của selenium cho google chrome. Nếu dùng nền tảng, trình duyệt khác thì viết lại hàm này.

- .//source//crawl\_title//filter\_title.py : chứa hàm filter\_similar\_title(items, threshold=0.65), đổi model sử dụng thì viết lại cả hàm để trả về danh sách link, title (do mỗi model có 1 cách sử dụng khác nhau)

- .//source//database//get\_title\_database.py : có sẵn hàm chạy để tải database về. To\_date là lần cuối cập nhật database, chỉ cần đổi from\_date = to\_date, to\_date = timestamp hiện tại

IV. Link code Git

*https://github.com/HungHa365/Phantichkey*

**Nhân viên**

**Họ tên**

*Cường*

*Nguyễn Mạnh Cường*