ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Sinh viên thực hiện:

Họ và tên: **Nguyễn Thị Ánh Tuyết**
MSSV: **20120422**
Lớp: **20_21**
Học phần: **Lập trình cho Khoa học dữ liệu**

Báo cáo
Bonus Assignment 1

Giáo viên hướng dẫn:
Thầy Bùi Tiến Lên
Thầy Lê Nhựt Nam
Thầy Lê Đại Chí

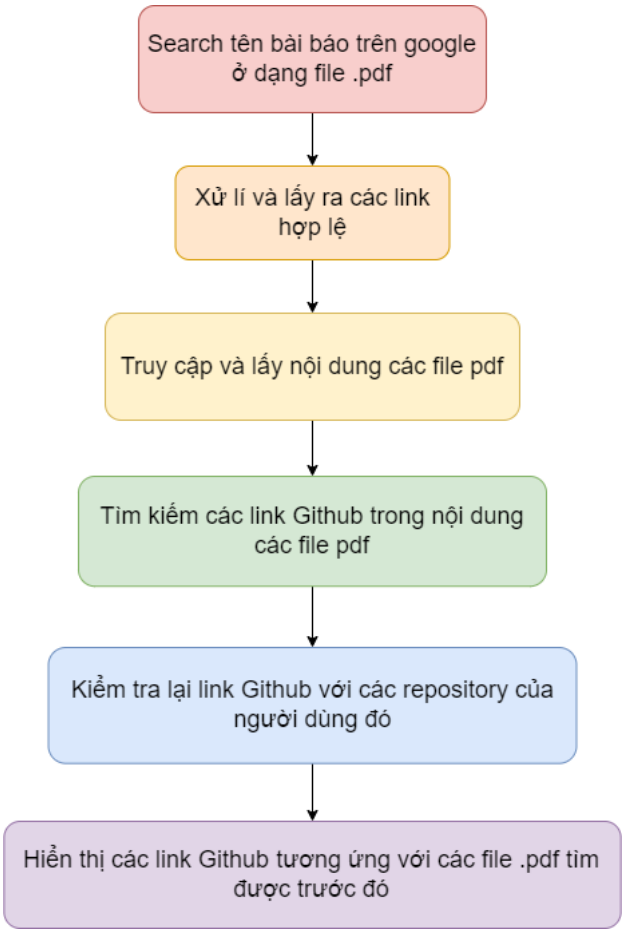
Thành phố Hồ Chí Minh – 12/2022

MỤC LỤC

BÁO CÁO

Yêu cầu: Kiểm tra repository của một bài báo khoa học

1. Quy trình thực hiện:



2. Các thư viện sử dụng:

Thư viện, công cụ	Mục đích
re	Tách chuỗi.
threading	Chạy nhiều tác vụ cùng lúc.
io → BytesIO	Chuyển từ bytes sang bytes stream (để làm input cho PyPDF2.PdfReader)
PyPDF2	Đọc file PDF
requests	Gửi http request và download nội dung pdf.
BeautifulSoup	Tách các thành phần của trang web.

3. Input/Output:

- Input: tên bài báo khoa học, kiểu string.
- Output:
  - Nếu tìm thấy link Github, chương trình sẽ trả về các link Github tương ứng với mỗi file pdf tìm được (dựa vào việc search google)
  - Nếu không tìm thấy link Github nào, chương trình sẽ thông báo: "**There is no Github source!**"

4. Các quá trình xử lí chính:

- Tìm các bài pdf trên google chứa tên giống hoặc gần giống với bài báo cần tìm.
- Truy cập vào các link file pdf để lấy nội dung file pdf.
- Tìm kiếm và hiển thị các link github tìm thấy được trong nội dung các file pdf.

Các quá trình được ghi chú trong file ***find_paper_github.ipynb***

5. Kiểm thử chương trình:

- Với input:

```
In [2]: # pdf_name = 'Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model'
        pdf_name = input('Nhap ten bai bao: ')

        Nhap ten bai bao: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
```

- Kết quả:

Kết quả

```
In [10]: if having_github_link:
             pdf_link = recheck(pdf_link)
             print_table_format(pdf_link)
         else:
             print('There is no Github source!')

         There is no Github source!
```

- Với input:

```
In [2]: # pdf_name = 'Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model'
        pdf_name = input('Nhap ten bai bao: ')

        Nhap ten bai bao: Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model
```

- Kết quả:

Kết quả

```
In [10]: if having_github_link:
             pdf_link = recheck(pdf_link)
             print_link(pdf_link)
         else:
             print('There is no Github source!')

         Link pdf: https://www.robots.ox.ac.uk/~vedaldi/assets/pubs/ulyanov20deep.pdf
              Link Github:
              ---> https://github.com/gfacciol/bm3d
         Link pdf: https://arxiv.org/pdf/2212.00490
              Link Github:
              ---> https://github.com/wyhuai/DDNM
```

6. Đánh giá:

- Chương trình chạy khá ổn, có thể tìm ra được link Github cho bài báo khoa học. Hoặc nếu không tìm thấy thì có thông báo cho người dùng.
- Tuy nhiên là kết quả có thể ra được nhiều link Github và người dùng sẽ cần thử truy cập tất cả để tìm thấy link mà mình cần.
- Trong các quá trình xử lí ở mục **4.** quá trình chạy lâu nhất là truy cập vào các link file pdf để lấy nội dung file pdf, tuy là em đã sử dụng thread để có thể truy cập và tải nội dung nhiều file 1 lần.

NGUỒN THAM KHẢO

Slide bài giảng môn học Lập trình cho Khoa học dữ liệu.

https://stackoverflow.com/questions/47801564/is-it-possible-to-input-pdf-bytes-straight-into-pypdf2-instead-of-making-a-pdf-f

https://pypdf2.readthedocs.io/en/latest/modules/PdfReader.html

https://docs.python.org/3/library/re.html