HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Đề tài: Máy tìm kiếm chủ đề động vật

Học phần: Tìm kiếm thông tin
GVHD: TS. Nguyễn Bá Ngọc
Nhóm PHM:

Nguyễn Hải Minh 20210611
Phạm Văn Phong 20215448
Đỗ Văn Hoàng 20215378

ONE LOVE. ONE FUTURE.

Mục lục:

1. Giới thiệu bài toán:

- **Vai trò của động vật:** Động vật quan trọng trong hệ sinh thái, nghiên cứu khoa học, giáo dục và giải trí.
- **Thách thức trong tìm kiếm thông tin:** Tìm kiếm thông tin về động vật trên Internet ngày càng khó khăn vì sự đa dạng và phức tạp của dữ liệu.
- **Công cụ tìm kiếm chuyên biệt:** Cần thiết để tiết kiệm thời gian và nâng cao hiệu quả tìm kiếm thông tin chính xác, đáng tin cậy về động vật.
- **Mục tiêu dự án:** Xây dựng công cụ tìm kiếm tối ưu, cung cấp thông tin về các loài động vật, bao gồm đặc điểm nhận dạng, môi trường sống, chế độ ăn uống và tình trạng bảo tồn.

2. Thu thập dữ liệu:

2.1. Công nghệ sử dụng

2.2. Phương pháp thu thập dữ liệu

2.3. Kết quả thu thập dữ liệu

2. Thu thập dữ liệu:

2.1. Công nghệ sử dụng

- **Python** là ngôn ngữ lập trình chính cho việc thu thập dữ liệu.
- Các thư viện chính sử dụng:
  - **Selenium**: Giả lập và tự động hóa trình duyệt web.
  - **Asyncio & aiohttp**: Xử lý yêu cầu HTTP/HTTPS bất đồng bộ.
  - **BeautifulSoup**: Phân tích cú pháp HTML.
  - Các thư viện hỗ trợ xử lý thời gian chờ và JSON.

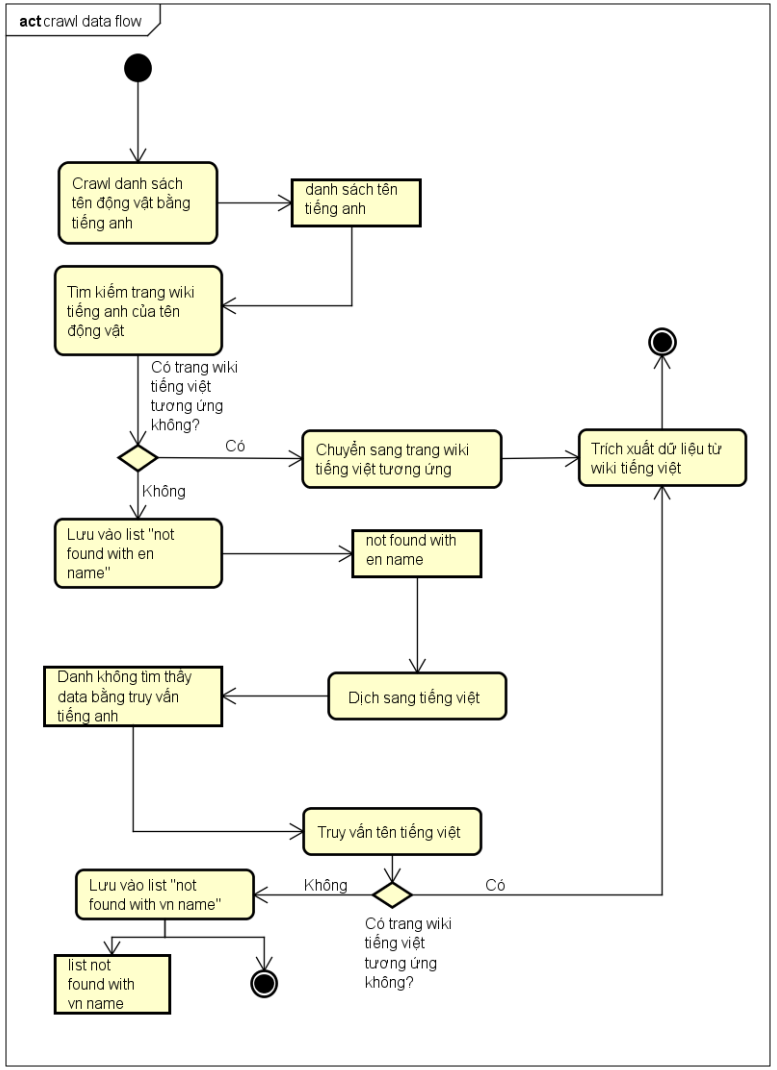
2. Thu thập dữ liệu:

2.2. Phương pháp thu thập dữ liệu

- Nguồn tài liệu:
  - **Animalia**: Danh sách tên động vật bằng tiếng Anh.
  - **Wikipedia**: Thu thập thông tin từ Wikipedia.
- Luồng thu thập dữ liệu:
  - Sử dụng bất đồng bộ để tối ưu tài nguyên.
  - Quá trình thu thập dữ liệu động vật từ các nguồn trên.

2. Thu thập dữ liệu:

2.2. Phương pháp thu thập dữ liệu



Hình 1. Luồng thu thập dữ liệu

2. Thu thập dữ liệu:

2.2. Phương pháp thu thập dữ liệu

* Quá trình thu thập dữ liệu:

- **Lập trình bất đồng bộ** giúp tăng hiệu quả.
- **Hạn chế**: Phụ thuộc vào tài nguyên mạng, nếu số lượng yêu cầu quá lớn, có thể làm chậm quá trình.
- **Lưu ý:** Không thu thập dữ liệu từ các trang web chặn bot hoặc không cho phép lưu trữ.

2. Thu thập dữ liệu:

2.3. Kết quả thu thập dữ liệu

- Thu thập 22600 tài liệu động vật từ Wikipedia tiếng Việt.
- Các trường dữ liệu bao gồm:
  - title, url, description, content, classify, categories.
- Hình ảnh mẫu dữ liệu trong bộ dữ liệu.
- Dữ liệu đã thu thập
  - Từ 8 phân ngành động vật, chọn 125 tài liệu có độ dài lớn nhất.
  - Tổng cộng 1000 tài liệu được chọn lọc và đưa vào bộ dữ liệu.

```
{
    "title": "Hàu Mỹ",
    "url": "https://vi.wikipedia.org/wiki/H%C3%A0u_M%E1%BB%B9",
    "classify": [
        "Animalia",
        "Mollusca",
        "Bivalvia",
        "Ostreoida (trang không tồn tại)",
        "Ostreidae",
        "Crassostrea",
        "Loài"
    ],
    "description": "Hàu Mỹ (Danh pháp khoa học: Crassostrea virginica) là một loài hàu trong họ Ostreidae phân bố ở Mỹ. Đây là một loại hàu có giá trị kinh tế và được nuôi nhiều ở Mỹ. ",
    "first_img": "https://upload.wikimedia.org/wikipedia/commons/e/e5/OysterBed.jpg",
    "content": "Hàu nuôi Hàng triệu con hàu đang được nuôi tại các vùng biển ở New York và các thành phố khác nhằm làm sạch môi trường nước bị ô nhiễm. Con hàu và các loài thủy sản có vỏ khác có thể làm sạch các chất độc và bụi bẩn. Việc khôi phục số lượng loài hàu tại sông Hudson gần Yonkers, bắc New York vì loài hàu giúp cải thiện môi trường sống thủy sinh, có tác dụng thu hút các loài thủy sản và sinh vật biển khác vào khu vực chúng sống. Các con hàu này chỉ nên sử dụng để làm sạch ô nhiễm, không nên ăn hay thu hoạch để bán. Một mẫu (0,4 ha mặt nước nuôi trồng) với 1 triệu con hàu cần khoản chi phí ít nhất là 50.000 USD. Mỗi con hàu có khả năng lọc khoảng 189,26 lít nước bẩn mỗi ngày.Thị trường Hiện nay Mỹ nhập khẩu hầu hết các sản phẩm hàu nuôi, khai thác, hun khói và các loại khác cùng tăng trưởng từ 16 - 123%, riêng sản phẩm hàu hun khói được Mỹ nhập khẩu nhiều nhất và chiếm đến 46,3% tổng giá trị, trong khi sản phẩm hàu khai thác chỉ chiếm chưa đến 2% thị phần. Xuất khẩu hàu của Mỹ cũng tăng trưởng nhẹ 1% về khối lượng và 6% về giá trị so với cùng kỳ năm trước, mặt hàng hàu tươi sống chiếm tỷ trọng cao nhất trên 70% tổng xuất khẩu mặt hàng này của Mỹ với khối lượng 1.820 tấn và giá trị 13,3 triệu USD, trong khi sản phẩm hàu đông lạnh và các loại khác của Mỹ lại xuất chủ yếu sang Hồng Kông.Trước đây Trung Quốc đã từng nhập khẩu khoảng 4.140 con Hàu các loại từ bang Washington Mỹ, nay Trung Quốc dừng nhập khẩu Hàu biển từ Mỹ do nhiễm khuẩn, việc nuôi dưỡng và sản xuất Hàu ở khu vực eo biển Hood bị nhiễm vi khuẩn Vibrio parahaemolyticus do ô nhiễm. Đồng thời một số lượng nhỏ loại thực phẩm này đã được phía Mỹ tiến hành xuất khẩu sang các nước như Trung Quốc, Thái lan và Indonexia. Trung Quốc đã nhập khẩu khoảng 4.140 con Hàu nhiễm bệnh. Hiện nay trên thị trường hải sản tươi sống tại thành phố Bắc Kinh không còn phát hiện thấy việc kinh doanh Hàu nhiễm bệnh có nguồn gốc từ Mỹ.Loại hàu mang tên Cape Neddick/Blue Point Oysters được khách hàng ưa thích dùng ăn sống vừa bị thu hồi tại Mỹ do nghi nhiễm vi khuẩn Vibrio parahaemolyticus. Loại khuẩn này dễ gây bệnh cho những người có hệ miễn dịch yếu, triệu chứng nhiễm khuẩn bao gồm tiêu chảy, buồn nôn, nôn, sốt và ớn lạnh. Thông thường các triệu chứng xảy ra trong vòng 24 giờ, kéo dài ba ngày.",
    "categories": [
        "Crassostrea",
        "Động vật được mô tả năm 1791",
        "Hàu",
        "Động vật Mỹ",
        "Động vật thân mềm thương mại",
        "Động vật thân mềm ăn được"
    ]
}
```

crawl_data > content_final_no_cate_3field > { } 0. Acanthonus armatus.json > ...


```
{
    "title": "Acanthonus armatus",
    "link": "https://vi.wikipedia.org/wiki/Acanthonus_armatus",
    "content": "Acanthonus armatus là một loài cá được tìm thấy ở vùng đại dương nhiệt đới và cận nhiệt đới ở độ sâu từ 1.171 đến 4.415 mét (3.842 đến 14.485 ft). Loài này phát triển đến chiều dài 37,5 cm (14,8 in) SL. Nó là thành viên duy nhất được biết trong chi của nó. Nội dung:  Phân loài: Động vật, Động vật có dây sống, Actinopterygii, Ophidiiformes, Ophidiidae (trang không tồn tại), Albert Günther, Loài,  first_img: https://upload.wikimedia.org/wikipedia/commons/9/93/Acanthonus_armatus.jpg"
}
```

3. Cấu hình trên Elasticsearch:

3.1. Analysis

3.2. Similarity

3.3. Mapping

3.1. Analysis

Analyzer:

1. **vn_text_analyzer:** custom từ những tùy chỉnh có sẵn trên Elasticsearch như: tokenize dựa trên khoảng trắng và ký tự đặc biệt, chuyển tất cả chữ cái thành chữ thường, loại bỏ các từ dừng (stopword) dựa trên danh sách định nghĩa.
2. **vn_text_analyzer_no_accent:** tương tự như analyzer trên nhưng với trường hợp đã loại bỏ dấu tiếng Việt.

```
"analyzer": {
    "vn_text_analyzer": {
        "type": "custom",
        "tokenizer": "standard",
        "filter": ["lowercase", "my_stop_filter"]
    },
    "vn_text_analyzer_no_accent": {
        "type": "custom",
        "tokenizer": "standard",
        "filter": ["lowercase", "asciifolding", "
           my_stop_filter"]
    }
},
```

3.1. Analysis

Tokenizer

- **type: "ngram"** - chia văn bản thành các chuỗi nhỏ (n-gram).
- **min_gram: 3** - Độ dài nhỏ nhất của n-gram
- **max_gram: 4** - Độ dài lớn nhất của n-gram

```
"tokenizer": {
    "my_ngram_tokenizer": {
        "type": "ngram",
        "min_gram": 3,
        "max_gram": 4
    }
},
```

3.1. Analysis

Filter

Bộ lọc để loại bỏ các từ dừng (stopword) được định nghĩa và lưu trữ trong file stopword.txt

```
"filter": {
    "my_stop_filter": {
        "type": "stop",
        "stopwords_path": "stopword.txt"
    }
}
```

3.2. Similarity

Sử dụng thuật toán **BM25** để tính toán độ tương đồng văn bản với các thông số tùy chỉnh:
- k1: Tham số kiểm soát độ nhạy với tần suất từ.
- b: Tham số kiểm soát độ nhạy với độ dài tài liệu.

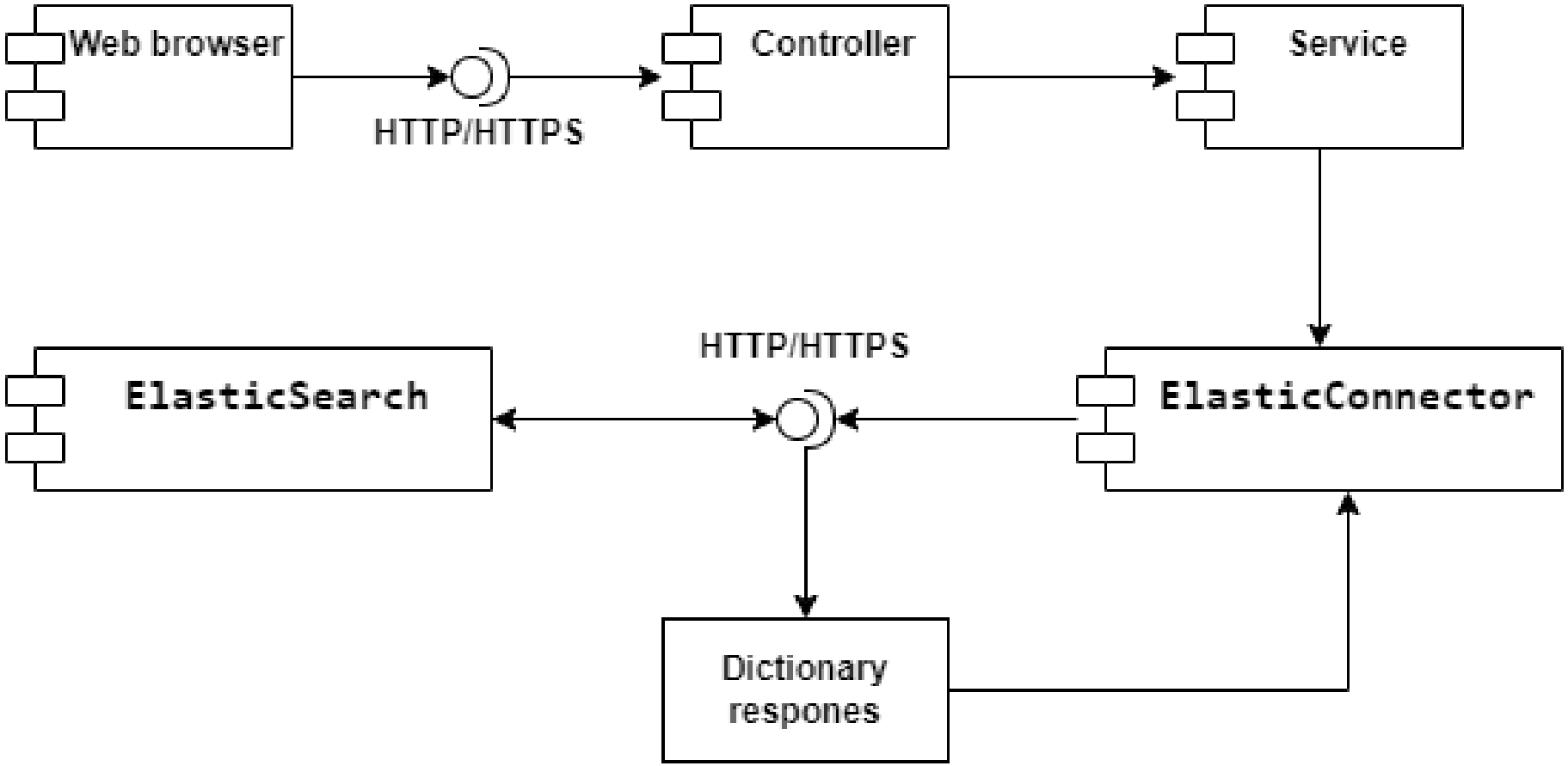
```
"similarity": {
  "bm25_custom_content": {
      "type": "BM25",
      "k1": 2.0
  }
}
```

3.3. Mapping

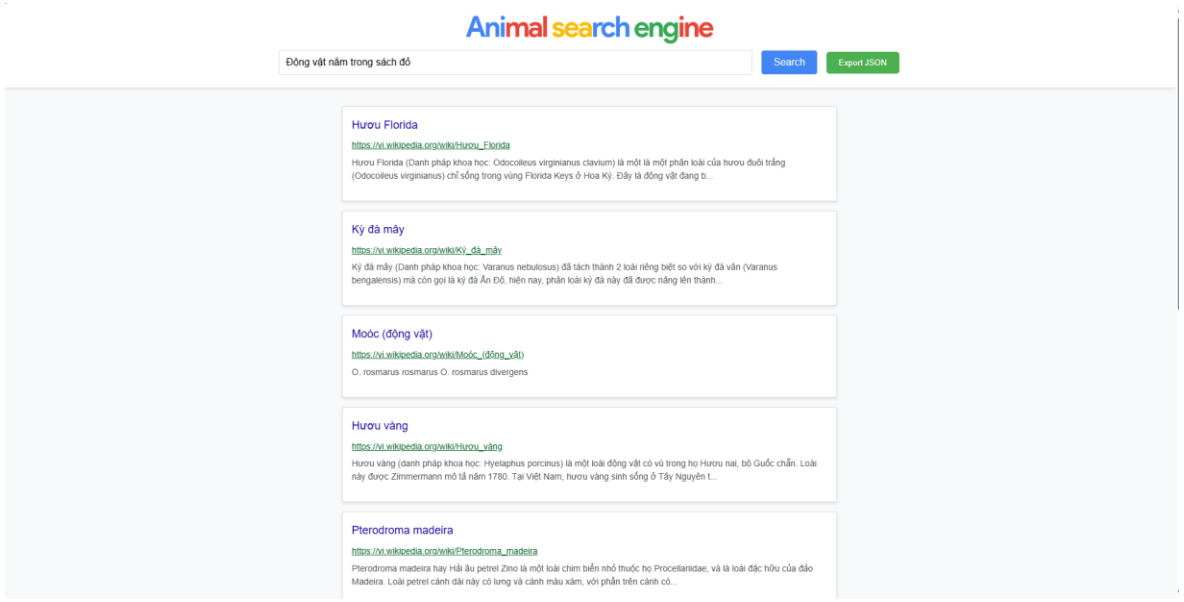
Trường **title, description, classify, content, categories:**

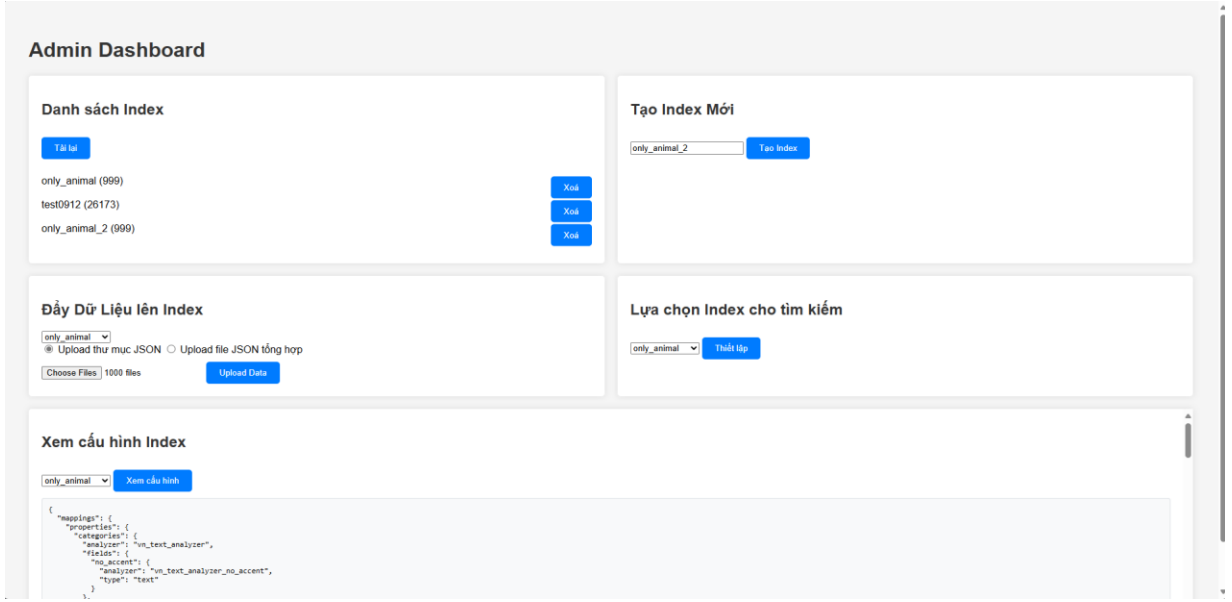
- Sử dụng các analyzer "vn_text_analyzer" và "vn_text_analyzer_no_accent" đã cấu hình.
- Cấu hình type là "text" để Elasticsearch tự động đánh chỉ mục ngược cho trường **content**.
- Trường **link, first_img:**
- Chọn type là "keyword" để chứa chuỗi ký tự ngắn không cần phân tích như URL.

Kiến trúc


Web browser
HTTP/HTTPS
Controller
Service
HTTP/HTTPS
ElasticSearch
ElasticConnector
Dictionary
respones

4. Xây dựng hệ thống:





Giao diện tìm kiếm

Admin Dashboard

Hình ảnh minh họa giao diện

5. Kết quả thực nghiệm:

Đánh giá kết quả tìm kiếm dựa trên độ chính xác trung bình AP
Thực hiện trên tập dữ liệu có 50 tài liệu, trong đó:

- Có **5** tài liệu phù hợp với "Động vật sống dưới nước"
- Có **8** tài liệu phù hợp với "Động vật sách đỏ Việt Nam"
- Có **12** tài liệu phù hợp với "Động vật ăn thịt"

MAP = 0.306

Truy vấn	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	AP
Động vật sống dưới nước	0	0	0	1	0	0	0	0	0	0	0.050
Động vật ăn thịt	1	1	1	1	1	0	0	1	1	1	0.611
Động vật sách đỏ Việt Nam	0	1	1	0	1	1	0	0	1	0	0.499
ĐỘng vậ sống dướis nướ	0	0	0	0	0	0	1	0	0	0	0.029
động vật ăng thịy	1	1	0	0	0	1	0	0	0	1	0.710
Động vậ sácg đỏ việt nam	0	1	1	0	1	1	1	1	1	1	0.697
Dong vat song duoi nuoc	0	0	0	0	0	1	0	0	0	1	0.054
Dong vat an thit	1	1	1	1	1	0	1	1	1	1	0.710
Dong vat sach do Viet Nam	1	0	1	0	0	0	1	0	0	1	0.042

6. Kết luận:

Đạt được

- Đáp ứng cơ bản yêu cầu của đề tài.
- Vận dụng được các kiến thức về máy tìm kiếm, đánh giá kết quả tìm kiếm, thu thập dữ liệu từ web.
- Có khả năng mở rộng trên các tập dữ liệu với nhiều chủ đề khác nhau.

Hạn chế

- Quá trình thu thập dữ liệu còn thô sơ, chưa tận dụng được tối đa các thư viện, tiện ích có sẵn
- Kết quả tìm kiếm chưa tối ưu và đáp ứng nhu cầu tìm kiếm của người dùng trong một số trường hợp cụ thể.

6. Kết luận:

Hướng phát triển

- Tiếp tục cải tiến hệ thống tìm kiếm bằng một số cách: Thử nghiệm các cấu hình chỉ mục tốt hơn giúp tối ưu độ chính xác của kết quả.
- Cải thiện tìm kiếm mờ (truy vấn không dấu, sai chính tả, tìm kiếm đồng nghĩa).
- Ngoài ra còn có thể ứng dụng trí tuệ nhân tạo vào trong xử lí truy vấn giúp đáp ứng nhu cầu người sử dụng.

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !