Introduction to Information Retrieval

Assignment 01: Web Crawler

Web Crawler for Extracting Data from Tuoitre.vn

Student ID: 23125067

Student Name: Nguyen Le Thinh Phuc

1. Approach

This project implements a large-scale web crawler designed to automatically collect structured news article data from the Vietnamese online newspaper tuoitre.com. The system is divided into two main components: category & URL collection and detailed article crawling with multimedia extraction. This modular design improves scalability, maintainability, and fault tolerance.

1.1 Category and Article URL Collection

The first stage focuses on identifying valid article URLs efficiently. The crawler begins by sending a request to the homepage and extracting all available news categories from the navigation menu. Each category is converted into both a main page URL and a corresponding RSS feed URL for future extensibility. Category filtering is supported via command-line arguments to allow selective crawling.

Within each category page, article URLs are collected from three main sections:

- Focus list (hot news)
- Sub-listing blocks
- Main article blocks

To ensure data quality, a strict URL validation function filters out invalid links such as search pages, print views, videos, and non-HTML content. Additionally, the crawler extracts a timeline ID embedded in the page to enable automatic pagination through AJAX-based timeline

endpoints. This allows the crawler to continuously load older articles until a predefined limit is reached.

All collected results are stored in a structured JSON file, where each category contains a list of article URLs ready for deep crawling.

1.2 Deep Article Crawling and Data Extraction

The second stage performs in-depth crawling of each article using a hybrid approach combining Requests + BeautifulSoup for static data and Playwright for JavaScript-rendered content (audio and reactions). A custom TLS adapter is used to handle SSL issues and improve connection stability.

For each article, the crawler extracts:

- Post ID
- Title, author, and publication date
- Cleaned article content
- Images with captions
- Audio podcasts
- User reactions (star, like, love)
- Threaded user comments via Tuổi Trẻ's official comment API

The text content is carefully cleaned by filtering advertisement blocks, related-news elements, empty paragraphs, and non-content placeholders, ensuring high-quality textual data.

1.3 Multimedia Download and Data Storage

All images and audio files are downloaded and organized into separate subfolders using the article ID as the directory name. The final result of each article is stored as a standalone JSON file, containing both metadata and local file paths to the downloaded media.

1.4 Logging and Fault Tolerance

A centralized logging system is implemented to track crawling progress, errors, and failed URLs without polluting the terminal output. Failed requests, network issues, and parsing errors

are recorded for later reprocessing. The crawler also avoids re-crawling finished URLs, improving efficiency.

2. Problems and Solutions

2.1 Ensuring Article Limit

Problem:

The category pages of tuoitre.com apply an infinite scrolling mechanism to progressively load older articles. As a result, a simple combination of HTTP requests and HTML parsing using BeautifulSoup can only retrieve the initially rendered content. This makes it impossible to guarantee the collection of a predefined number of articles using traditional static crawling methods.

Solution:

Through traffic inspection of the website, a hidden timeline endpoint was identified with the format: *https://tuoitre.vn/timeline/{timeline_id}/trang-{page_index}.htm*,

where the timeline_id is fixed for each category. By automatically extracting this timeline ID and iteratively increasing the page index, the crawler systematically requests additional pages until the required lower bound of article URLs is reached. This approach ensures deterministic control over the number of articles collected per category.

2.2. Handling Dynamically Rendered JavaScript Content

Problem:

Several components of the article pages, such as audio players and user reaction counters, are dynamically rendered through JavaScript after the initial HTML document is loaded. Consequently, traditional crawling with requests and BeautifulSoup only captures static content and fails to retrieve these dynamic elements.

Solution:

To address this limitation, the Playwright browser automation framework is integrated into the crawling pipeline. Playwright simulates a real browser environment, allowing JavaScript execution to complete before data extraction. This enables accurate retrieval of dynamically loaded multimedia content and reaction statistics.

2.3 Automatic Retrieval of Paginated User Comments

Problem:

User comments are not embedded directly in the article HTML. Instead, they are loaded asynchronously via a pop-up interface and therefore cannot be extracted using standard DOM parsing with BeautifulSoup.

Solution:

By analyzing the network requests triggered during comment loading, an internal API endpoint was discovered: *...api/getlist-comment.api?pageindex={page_index}&objId={post_id}*.

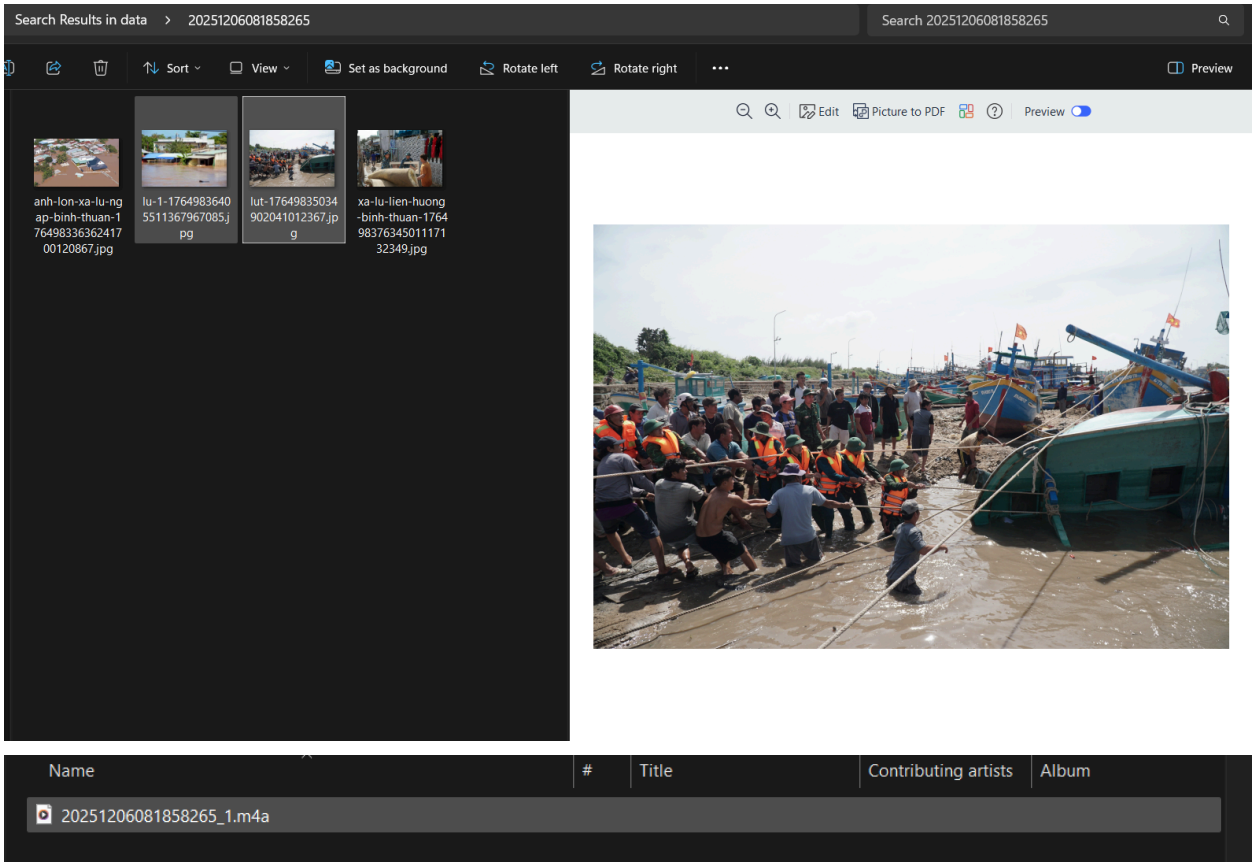
The crawler programmatically iterates through all available comment pages by incrementing the page index until no additional comments are returned. This ensures that the complete hierarchical comment structure, including nested replies and reactions, is fully collected.

3. Sample Output

3.1 Sample Output Json:


```
{
  "postId": "20251206081858265",
  "category": "Thời sự",
  "url": "https://tuoitre.vn/xa-lu-ky-luc-o-lam-dong-vi-sao-chi-thong-bao-truoc-2-gio-20251206081858265.htm",
  "title": "Xả lũ kỷ lục ở Lâm Đồng: Vì sao chỉ thông báo trước 2 giờ?",
  "content": "Những nơi thiệt hại nặng nhất là Liên Hương và vùng giáp Phan Thiết cũ như Hàm Liêm, Hàm Thuận, Hàm Thắng. Nguyên nhân ban đầu được ghi nhận là doxả lũõ ạt từ các hồ thủy lợi trư
  "author": "ĐỨC TRONG",
  "date": "2025-12-06T10:00:00+07:00",
  "audio_podcast": [
    {
      "url": "https://tts.mediacdn.vn/2025/12/06/tuoitre-nu-1-20251206081858265.m4a",
      "local_path": "data\\audio\\20251206081858265\\20251206081858265_1.m4a"
    }
  ],
  "images": [
    {
      "url": "https://cdn2.tuoitre.vn/471584752817336320/2025/12/6/anh-lon-xa-lu-ngap-binh-thuan-17649833636241700120867.jpg",
      "caption": "xả lũ - Ảnh 1.",
      "local_path": "data\\images\\20251206081858265\\anh-lon-xa-lu-ngap-binh-thuan-17649833636241700120867.jpg"
    },
    {
      "url": "https://cdn2.tuoitre.vn/471584752817336320/2025/12/6/lut-17649835034902041012367.jpg",
      "caption": "xả lũ - Ảnh 2.",
      "local_path": "data\\images\\20251206081858265\\lut-17649835034902041012367.jpg"
    },
    {
      "url": "https://cdn2.tuoitre.vn/471584752817336320/2025/12/6/lu-1-17649836405511367967085.jpg",
      "caption": "xả lũ - Ảnh 3.",
      "local_path": "data\\images\\20251206081858265\\lu-1-17649836405511367967085.jpg"
    },
    {
      "url": "https://cdn2.tuoitre.vn/471584752817336320/2025/12/6/xa-lu-lien-huong-binh-thuan-17649837634501117132349.jpg",
      "caption": "xả lũ - Ảnh 4.",
      "local_path": "data\\images\\20251206081858265\\xa-lu-lien-huong-binh-thuan-17649837634501117132349.jpg"
    }
  ],
  "comments": [
    {
      "commentId": "ee501820-8167-4b6d-8ee4-87c57d88d871",
      "author": "Ana",
```

3.2 Sample Downloaded Images and Audio:



3.3 Nested Comments


```
{
  "commentId": "86b88905-f085-4962-9743-0462ca7e1fbb",
  "author": "daid****@gmail.com",
  "text": "Đề nghị nhà nước nên có một bộ luật hình sự để răn đe những người điều hành nhà máy thủy điện",
  "date": "2025-12-06T14:50:09.7500116",
  "vote_reactions": {
    "like": 0,
    "love": 1,
    "wow": 0,
    "sad": 0,
    "angry": 0
  },
  "replies": [
    {
      "commentId": "49a8097a-9730-4346-b62c-7d6898186ad6",
      "author": "dung****@gmail.com",
      "text": "Trách nhiệm nhà nước phải có nghĩa vụ bảo vệ tài sản và tính mạng của nhân dân. ",
      "date": "2025-12-06T16:44:18.4326377",
      "vote_reactions": {
        "like": 0,
        "love": 1,
        "wow": 0,
        "sad": 0,
        "angry": 0
      },
      "replies": []
    }
  ]
},
```

3.3 Posts with at least 20 comments

```
(venv) PS D:\projects\tuoitre-crawler> python .\check_comments.py
[MATCH] 59 comments | 2 depth | Pháp luật | https://tuoitre.vn/vu-11-anh-em-ruot-kien-nhau-thua-ke-dat-toa-tuyen-bi-don-duoc-so-huu-manh-dat-tranh-chap-20250917165953208.htm
[MATCH] 21 comments | 2 depth | Pháp luật | https://tuoitre.vn/sinh-vien-bong-dung-thanh-con-no-cua-app-vay-tien-online-20251118145608691.htm
[MATCH] 32 comments | 2 depth | Xe | https://tuoitre.vn/bo-cong-an-de-xuat-bo-bai-thi-mo-phong-trong-sat-hach-giay-phep-lai-xe-o-to-20251125100239 4.htm
[MATCH] 43 comments | 2 depth | Xe | https://tuoitre.vn/vu-o-to-con-tong-truc-dien-xe-dau-keo-4-nguoi-chet-tai-xe-nao-phai-boi-thuong-20251126215617888.htm
[MATCH] 23 comments | 2 depth | Pháp luật | https://tuoitre.vn/bat-khan-cap-nguoi-cha-om-con-nho-2-tuoi-va-6-tuoi-nhay-xuong-song-tra-khuc-tu-tu-20251127101908173.htm
[MATCH] 22 comments | 2 depth | Pháp luật | https://tuoitre.vn/thu-giu-12-sieu-xe-300-cay-vang-100-so-do-400-000-usd-cua-chu-tham-my-vien-mailisa-20251128155114772.htm
[MATCH] 21 comments | 2 depth | Pháp luật | https://tuoitre.vn/cong-ty-dang-ky-luat-thuc-tap-sinh-len-mang-gay-xon-xao-20251129095532354.htm
[MATCH] 21 comments | 2 depth | Pháp luật | https://tuoitre.vn/vu-tranh-chap-cho-do-o-to-o-tang-ham-chung-cu-bat-truong-ban-quan-tri-va-5-nguoi-20251130100032959.htm
[MATCH] 100 comments | 2 depth | Xe | https://tuoitre.vn/khong-yeu-cau-o-to-kinh-doanh-van-tai-hanh-khach-trang-bi-ghe-ngoi-cho-tre-em-20251130110908759.htm
[MATCH] 22 comments | 2 depth | Pháp luật | https://tuoitre.vn/vuong-day-cap-sa-xuong-duong-mot-can-bo-cong-an-tu-vong-20251130124857 03.htm
[MATCH] 23 comments | 2 depth | Pháp luật | https://tuoitre.vn/tong-vang-nguoi-khac-len-via-he-con-lao-vao-danh-toi-tap-tam-giu-hinh-su-1-nghi-can-20251201002221549.htm
[MATCH] 28 comments | 2 depth | Pháp luật | https://tuoitre.vn/vu-an-o-lang-son-bat-tam-giam-doan-van-sang-cuu-doi-pho-quan-ly-thi-truong-ve-toi-giet-nguoi-20251203124726484.htm
[MATCH] 24 comments | 2 depth | Thời sự | https://tuoitre.vn/co-hay-khong-tram-sac-xe-dien-trong-chung-cu-phu-thuoc-chu-dau-tu-va-ban-quan-ly-20251204110804099.htm
[MATCH] 22 comments | 1 depth | Xe | https://tuoitre.vn/honda-dream-ncx-2026-dau-tien-ve-viet-nam-gia-hon-100-trieu-dong-20251204171041114.htm
[MATCH] 21 comments | 2 depth | Pháp luật | https://tuoitre.vn/cong-an-viet-nam-phoi-hop-cong-an-campuchia-bat-nhom-lua-dao-lon-gia-mao-vingroup-doji-20251204184437504.htm
[MATCH] 46 comments | 2 depth | Pháp luật | https://tuoitre.vn/cong-an-ly-giai-viec-phat-750-000-dong-vi-chua-dang-ky-thuong-tru-cho-tre-20251207152541088.htm
[MATCH] 23 comments | 2 depth | Thời sự | https://tuoitre.vn/dan-thiet-hai-nang-vi-ho-xa-lu-cong-ty-khong-co-tien-den-20251208074530014.htm
[MATCH] 27 comments | 1 depth | Thời sự | https://tuoitre.vn/tp-hcm-chap-thuan-trien-khai-du-an-duong-sat-ben-thanh-can-gio-du-kien-khoi-cong-ngay-19-12-20251208124017257.htm
```

3.4 Detailed Log File

```
2025-12-09 00:22:59 | ERROR | ArticleCrawler | article_crawler.py:309 | Failed extracting reactions from https://tuoitre.vn/honda-city-doi-mo
Traceback (most recent call last):
  File "D:\projects\tuoitre-crawler\article_crawler.py", line 294, in extract_reactions
    self.page.wait_for_selector(
  File "D:\projects\tuoitre-crawler\venv\Lib\site-packages\playwright\sync_api\_generated.py", line 8182, in wait_for_selector
    self._sync(
  File "D:\projects\tuoitre-crawler\venv\Lib\site-packages\playwright\_impl\_sync_base.py", line 115, in _sync
    return task.result()
           ^^^^^^^^^^^^^
  File "D:\projects\tuoitre-crawler\venv\Lib\site-packages\playwright\_impl\_page.py", line 422, in wait_for_selector
    return await self._main_frame.wait_for_selector(**locals_to_params(locals()))
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "D:\projects\tuoitre-crawler\venv\Lib\site-packages\playwright\_impl\_frame.py", line 369, in wait_for_selector
    await self._channel.send(
  File "D:\projects\tuoitre-crawler\venv\Lib\site-packages\playwright\_impl\_connection.py", line 69, in send
    return await self._connection.wrap_api_call(
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "D:\projects\tuoitre-crawler\venv\Lib\site-packages\playwright\_impl\_connection.py", line 559, in wrap_api_call
    raise rewrite_error(error, f"{parsed_st['apiName']}: {error}") from None
playwright._impl._errors.TimeoutError: Page.wait_for_selector: Timeout 10000ms exceeded.
Call log:
  - waiting for locator("#main-detail > div.sendstarauthor > div > div > div.reactinfo") to be visible

2025-12-09 00:22:59 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251122091601689
2025-12-09 00:23:09 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251208085313474
2025-12-09 00:23:18 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251129093313689
2025-12-09 00:23:29 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251208011206201
2025-12-09 00:23:39 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251201150252038
2025-12-09 00:23:51 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251129131544834
2025-12-09 00:24:05 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251204171041114
2025-12-09 00:24:32 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251130110908759
2025-12-09 00:24:47 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251202161042943
2025-12-09 00:25:01 | INFO | ArticleCrawler | article_crawler.py:82 | Saved article 20251126165734993
```