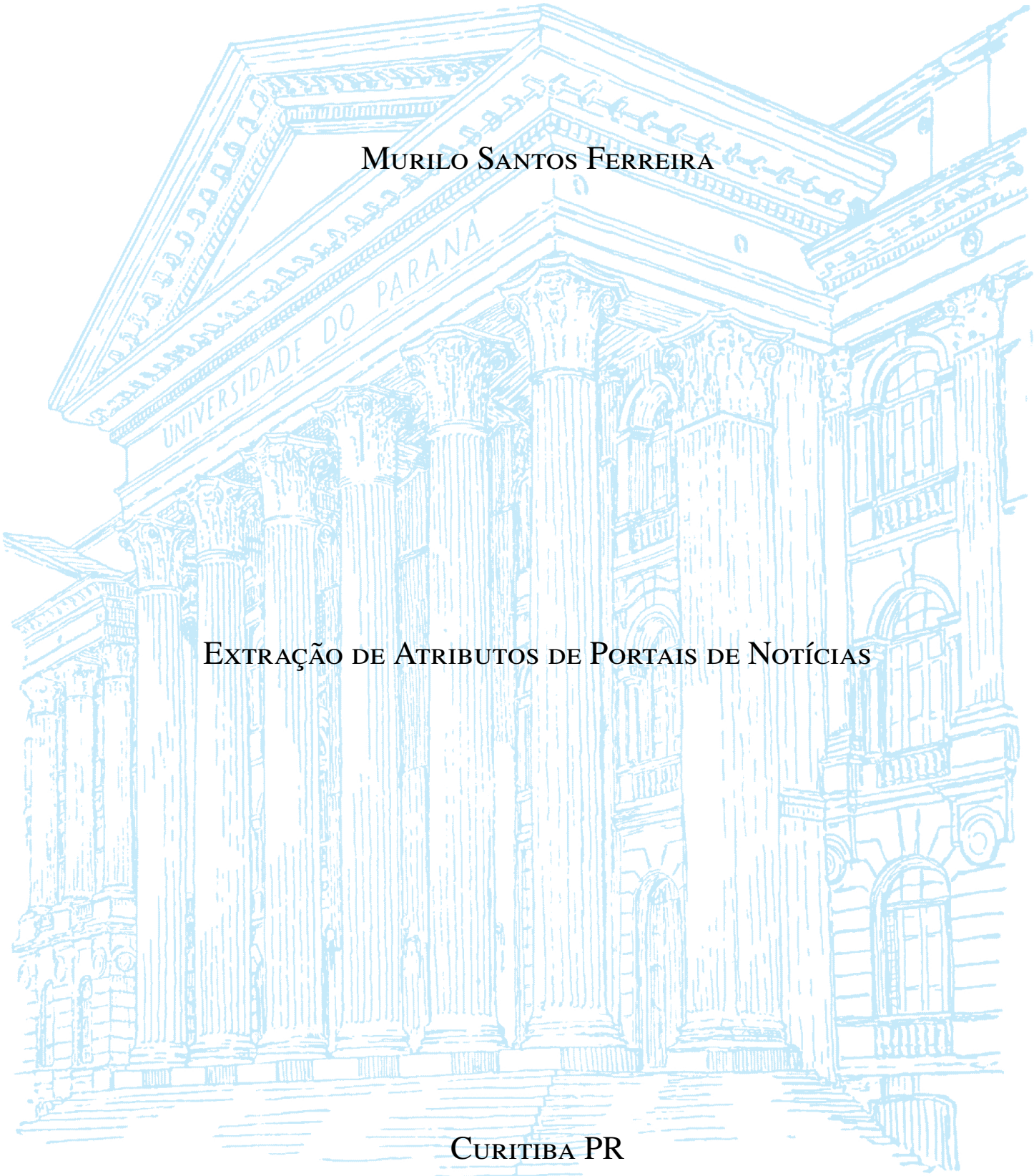
Universidade Federal do Paraná

Murilo Santos Ferreira



Extração de Atributos de Portais de Notícias

Curitiba PR
2025

Murilo Santos Ferreira

Extração de Atributos de Portais de Notícias

Trabalho de graduação apresentado como requisito à obtenção do grau de Bacharel em Ciência da Computação, no Departamento de Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Carmem Satie Hara.

Curitiba PR
2025

Resumo

Há uma crescente demanda por soluções que possibilitem o acesso estruturado a informações provenientes de portais de notícias. Dentre as ferramentas voltadas à coleta automatizada de conteúdos jornalísticos online existentes, o ENoW (Extrator de Dados de Notícias da Web) destaca-se por ser de código aberto e armazenar as informações extraídas em um banco de dados relacional. No entanto, para que a extração seja possível, é necessário que o usuário manualmente defina os caminhos dentro da página HTML que devem ser percorridos para obter os atributos de interesse. Nesse contexto, esta monografia apresenta melhorias no sistema ENoW, propondo duas abordagens para facilitar a extração dos atributos: a extensão ENoW Selector e a integração da biblioteca Newspaper. O ENoW Selector foi desenvolvido como uma extensão para navegadores que facilita a configuração do sistema, permitindo a seleção visual dos elementos HTML desejados, como título, data e conteúdo da notícia. Essa abordagem visa eliminar a necessidade de intervenção manual, reduzindo a barreira técnica para novos usuários e acelerar o cadastro de novos portais. Já a integração com a biblioteca Newspaper tem como objetivo automatizar ainda mais o processo de extração, aplicando heurísticas para identificar automaticamente os principais atributos das notícias.

Palavras-chave: Raspagem de Dados, Notícias, Extração.

Abstract

There is a growing demand for solutions that enable structured access to information from news portals. Among the tools focused on the automated collection of online journalistic content, ENoW (Web News Data Extractor) stands out for being open source and for storing extracted information in a relational database. However, in order for extraction to be possible, users must manually define the paths within the HTML page that should be followed to retrieve the desired attributes. In this context, this monograph presents improvements to the ENoW system, proposing two approaches to facilitate attribute extraction: the ENoW Selector extension and the integration of the Newspaper library. The ENoW Selector was developed as a browser extension that simplifies system configuration by allowing visual selection of desired HTML elements such as the title, date, and article content. This approach aims to eliminate the need for manual intervention, lower the technical barrier for new users, and accelerate the registration of new portals. Meanwhile, the integration with the Newspaper library seeks to further automate the extraction process by applying heuristics to automatically identify the main attributes of news articles.

Keywords: Web Scraping, News, Extraction.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

API	Interface de Programação de Aplicações (Application Programming Interface)
CAPTCHA	Teste de Turing Público Completamente Automatizado para Diferenciar Computadores de Humanos
CSS	Folhas de Estilo em Cascata (Cascading Style Sheets)
CSV	Valores Separados por Vírgula (Comma-Separated Values)
ENoW	Extrator de Dados de Notícias da Web
HTML	Linguagem de Marcação de Hipertexto (HyperText Markup Language)
HTTP	Protocolo de Transferência de Hipertexto (HyperText Transfer Protocol)
UFPR	Universidade Federal do Paraná
URL	Localizador Uniforme de Recursos (Uniform Resource Locator)
XML	Linguagem de Marcação Extensível (eXtensible Markup Language)

Sumário

Capítulo 1

Introdução

A crescente digitalização da informação e a popularização da internet resultaram em uma explosão no volume de conteúdos publicados online. Portais jornalísticos atualizam-se em tempo real, oferecendo aos leitores uma ampla cobertura de temas diversos e, ao mesmo tempo, são considerados fontes confiáveis de informação, conforme abordado por KALOGEROPOULOS, A. et al. [2019]. Com isso, surge também a demanda por ferramentas capazes de coletar, organizar e analisar essas informações de forma automatizada, especialmente em contextos de pesquisa acadêmica e monitoramento de mídia. A obtenção estruturada de dados jornalísticos pode contribuir para estudos em diversas áreas, como ciência política (ANSOLABEHERE, S. et al. [2006]), estudo de mudanças climáticas (VARGAS-SOLAR, G. et al. [2021]), predição de turismo (PARK, E. et al. [2021]), sociologia, jornalismo e vários outros, fornecendo subsídios para análises qualitativas e quantitativas de eventos e tendências sociais.

Nesse contexto, destaca-se o desenvolvimento do sistema ENoW (Extrator de Dados de Notícias da Web) REIPS, L. et al. [2023], que tem como principal finalidade a coleta automatizada de notícias a partir de portais previamente cadastrados, organizando os dados em um banco relacional para posterior análise.

1.1 Motivação

O ENoW é um sistema gratuito, personalizável e voltado à comunidade acadêmica. No entanto, uma das principais limitações observadas foi a necessidade de configurar manualmente como é feita a extração dos atributos HTML de cada portal. Esta tarefa pode demandar tempo e conhecimentos técnicos específicos. A identificação dessa limitação motivou a busca por soluções que reduzissem a barreira técnica e aumentassem a eficiência no uso do sistema.

1.2 Proposta

Este trabalho propõe a incorporação de duas soluções ao ENoW como forma de aprimorar sua usabilidade, reduzir o tempo de configuração manual e ampliar o número de portais compatíveis. É proposto o *ENoW Selector*, uma extensão para navegadores que permite a seleção visual dos elementos HTML, e a integração da biblioteca *Newspaper*, capaz de extrair automaticamente atributos como título, texto e data com base em heurísticas de estrutura e conteúdo.

1.3 Organização do documento

Este trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta os trabalhos relacionados ao tema e faz um comparativo. O Capítulo 3 descreve o funcionamento do ENoW, bem como as soluções propostas, o ENoW Selector e a utilização da biblioteca Newspaper. Por fim, o capítulo 4 apresenta as conclusões, discute as soluções apresentadas e apresenta sugestões de trabalhos futuros.

Capítulo 2

Trabalhos relacionados

Dentro deste capítulo, serão analisadas ferramentas de raspagem de dados Web. O foco desse capítulo é entender as características das ferramentas, assim como suas vantagens, desvantagens e casos de uso preferíveis. A escolha das ferramentas para a comparação buscou maximizar a gama de funcionalidades dentre as opções mais populares do mercado atual e também outras ferramentas criadas em ambiente acadêmico.

2.1 ParseHub

O ParseHub [1] é uma ferramenta de raspagem de dados com foco na utilização por usuários não especializados. Ele possui uma interface gráfica que permite, a partir de uma página inicial pré-definida, extrair textos, imagens e outros elementos HTML de forma estruturada em arquivos JSON e tabulares (Excel e CSV). É também possível consumir o conteúdo extraído via uma API.

Como mencionado anteriormente, a partir de uma interface visual, o usuário pode mapear o caminho que o aplicativo deve fazer na página Web e os dados para serem extraídos. Essa aplicação faz o caminho a partir de uma página inicial e pode navegar de maneira linear ou também fazer paginação de forma iterativa quando os dados não tiverem tamanho fixo. Através de uma interface de apontar e clicar, ele é capaz de identificar padrões e sugerir a possível estrutura, sem a necessidade de clicar repetidas vezes em objetos similares presentes na página.

Uma grande desvantagem do ParseHub se deve à sua baixa adaptabilidade a mudanças no layout do site, requerendo reconfiguração total do caminho em caso de mudanças na estrutura da página. A ferramenta também não possui um sistema de autoidentificação de informações, o que impossibilita a fácil inclusão de outras fontes disponíveis na Internet.

A versão paga do ParseHub resolve algumas limitações de raspagem presentes na versão gratuita:

- para evitar possíveis bloqueios de sistemas de proteção de robôs, desfere requisições de múltiplos endereços IP;
- reduz o limite de requisições por segundo, acelerando o processo de raspagem em múltiplas vezes;
- número ilimitado de projetos simultâneos.

A Figura 2.1 apresenta a interface da ferramenta.

[1] https://www.parsehub.com/

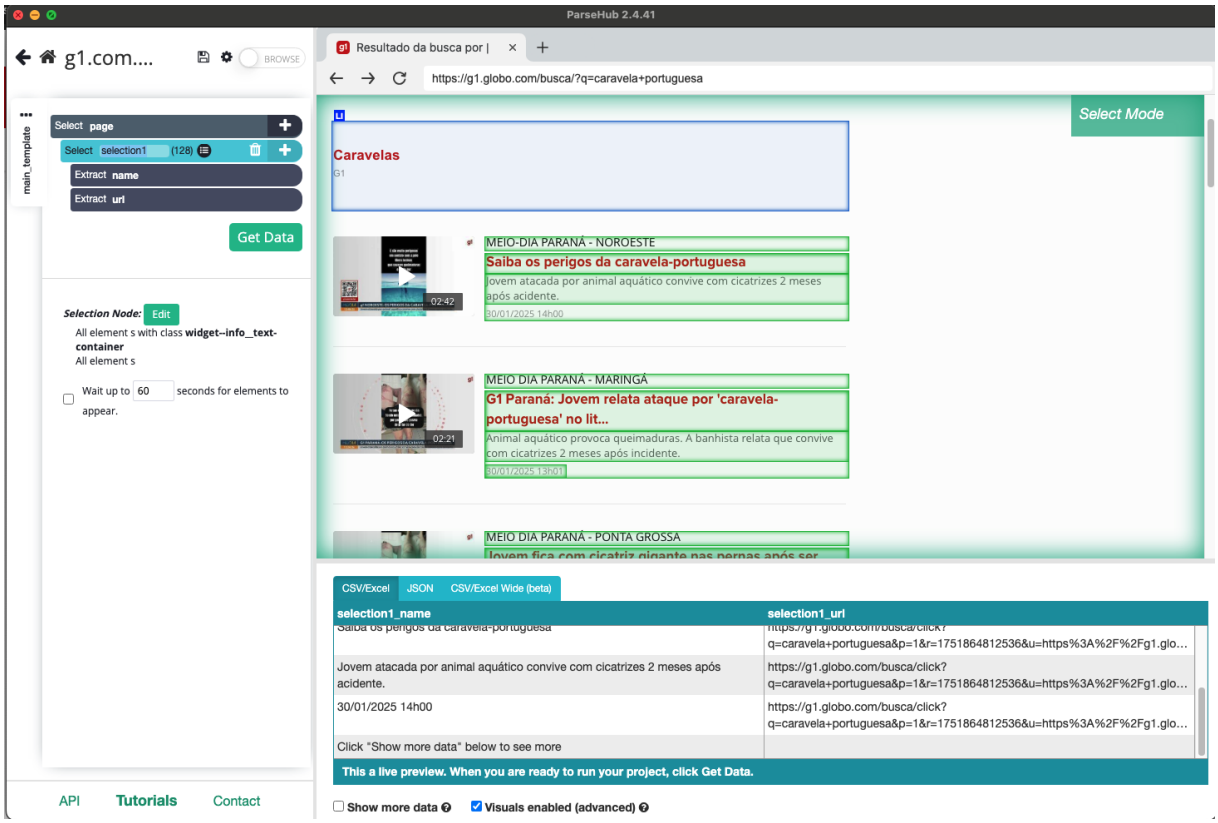


Figura 2.1: Utilização do ParseHub

2.2 80legs

O 80legs[2] é uma plataforma de raspagem de dados, com alta adaptabilidade e, por isso, maior alcance de raspagem dentro da Internet. Diferente de outras ferramentas com foco na experiência do usuário, essa ferramenta prioriza a complexidade da raspagem, necessitando de configuração mais técnica e aprofundada.

Por se tratar de uma aplicação configurável e escalável, torna-se muito desejável para processar grandes volumes de dados. Totalmente baseada em APIs, o utilizador precisa definir, através de script, o comportamento da ferramenta e as informações que serão buscadas dentro da página. Os resultados podem ser resgatados em múltiplos formatos estruturados, como diretamente da API, mas também em arquivos tabulares e JSON. Ainda assim, ela é sensível a mudanças de estrutura na página e pode falhar em casos de mudanças estruturais bruscas nos websites escolhidos.

A ferramenta tem pouca utilidade em sua versão gratuita, restringindo boa parte da sua API. Usuários pagantes podem explorar funcionalidades como:

- paralelismo na execução da raspagem;
- raspagem de infinitas páginas e requisições HTTP;
- extensão da API para raspagens, fornecendo funcionalidades úteis ao usuário.

[2]https://80legs.com/

2.3 OctoParse

O OctoParse[3] também se trata de uma ferramenta de apontar e clicar, sem necessidade de programação, com interface gráfica voltada para usuários sem experiência prévia em áreas de tecnologia. Ele permite definir regras de extração após o usuário mostrar os padrões que deseja capturar na página e, de forma similar ao ParseHub, identifica padrões e relacionamentos repetidos nas páginas visitadas.

A partir de uma página inicial, o OctoParse começa a extração e faz um caminho pré-definido, navegando para outras páginas e até mesmo podendo interagir com JavaScript e recursos que utilizam AJAX, aumentando a possibilidade de páginas que podem ser extraídas através dessa ferramenta. Na saída de dados, a ferramenta oferece formatos via API, dados tabulares e outros, como JSON, XML e HTML.

Com o propósito de facilitar a experiência do usuário, o OctoParse fornece templates prontos para sites onde a raspagem é mais comum: X (antigo Twitter), Google Maps, e-mails, sites de vagas de emprego, Amazon e vários outros.

Assim como as outras ferramentas descritas anteriormente, há diversas limitações na versão gratuita, com funcionalidades avançadas sendo disponibilizadas apenas após o pagamento de uma assinatura na plataforma. Algumas das funcionalidades na versão paga são:

- maior limite de tarefas criadas, tarefas executadas simultaneamente e requisições por segundo;
- execução através do servidores OctoParse, e não do computador do usuário. Isso possibilita utilização de múltiplos endereço IP para burlar bloqueios;
- resolução de CAPTCHAs;
- funcionalidades avançadas da API para integração com outros sistemas;
- backups automáticos.

A Figura 2.2 apresenta a interface da ferramenta.

2.4 Biblioteca Newspaper3k

Newspaper3k[4] é uma biblioteca implementada na linguagem Python, de código aberto, disponível através do repositório padrão da linguagem. Ela utiliza outras bibliotecas disponíveis na linguagem para requisições HTTP, raspagem de dados e processamento de linguagem natural para unificar o processo de processamento de notícias e artigos disponíveis online em uma só ferramenta, reduzindo o número de chamadas a funções e encapsulando a lógica de extração, análise e processamento para o usuário final.

Com poucas chamadas, o Newspaper3k é capaz de fazer download de uma página de notícia para a memória interna do programa, e então, usar NLP para extrair atributos desejados da página, tais como nome do autor, data de postagem, título, subtítulo, texto, lista de imagens e vários outros, após processar e limpar o código HTML de possíveis distrações, como anúncios e barras laterais. Além disso, consegue gerar resumos e listas de palavras-chave.

[3] https://www.octoparse.com/

[4] https://github.com/codelucas/newspaper

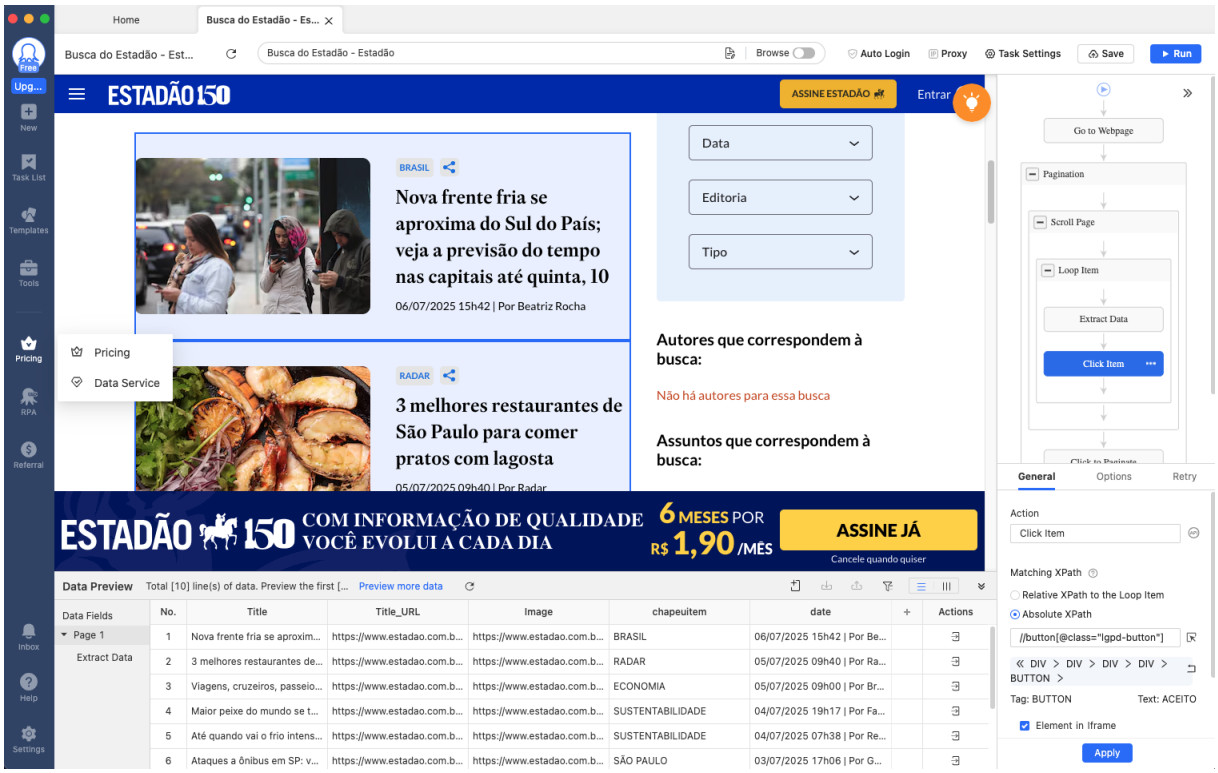


Figura 2.2: Utilização do OctoParse

Devido ao uso de NLP, o Newspaper3k não necessita de configuração manual para diferentes layouts de página web. No entanto, ele pode apresentar limitações em websites com maior complexidade no código-fonte, não possui um controle granular para selecionar atributos e deixa a possibilidade de errar a extração de um atributo, gerando um resultado incorreto para o usuário final.

2.5 FactExtract

O FactExtract SARR, E. N. et al. [2018], uma ferramenta de raspagem de dados focada em artigos de jornais, foi criada em ambiente acadêmico e segue uma abordagem diferente do ferramental mencionado anteriormente. Diferente da forma de navegação através da pré-definição de um caminho para a ferramenta percorrer, o FactExtract funciona baseado em dois núcleos: o módulo de extração de artigos (MEA) e o módulo de fusão de artigos e afirmações (MFA). O primeiro percorre páginas web e artigos através da definição de palavras-chave de entrada, enquanto o segundo faz a limpeza dos dados coletados e os agrega, gerando um banco de dados para análise posterior.

A entrada dada para a ferramenta é uma lista de websites pré-definidos e as palavras-chave desejadas, diferente da definição precisa e trabalhosa que é requisitada pelas outras aplicações. Por se tratar de uma ferramenta com foco em páginas textuais, usa a vantagem de ferramentas de processamento de texto para maior automatização do processo. A saída possui elementos estruturados para a rotulagem dos dados, tais como timestamps, autores e outros valores qualitativos.

Por ser uma ferramenta acadêmica, o FactExtract é gratuito e foi construído usando outras bibliotecas, como a Newspaper3k anteriormente mencionada, dependendo apenas da obtenção de código, compilação e recursos computacionais do usuário para ser utilizado. No

entanto, não possui recursos comerciais avançados que foram mencionados anteriormente, como a resolução de CAPTCHAs, requisições distribuídas e outras facilidades fornecidas por empresas especializadas.

2.6 Projeto ENoW

Assim como o FactExtract, o ENoW REIPS, L. et al. [2023] também foi idealizado e criado em ambiente acadêmico. Motivado pela limitação de ferramentas pagas, o ENoW é um projeto codificado em Python que utiliza bibliotecas como o BeautifulSoup para raspagem de dados e o Newspaper3k para análise de notícias. Essa ferramenta funciona com a pré-definição dos metadados para portais de notícias e outros tipos de websites e, a partir disso, a definição de uma string de busca a ser realizada nestas fontes de informação.

A partir do esforço manual de cadastro de metadados, o sistema é capaz de realizar buscas automatizadas a partir desse ponto: percorrer os resultados paginados, acessar individualmente cada notícia retornada e extrair informações relevantes como título, descrição, conteúdo completo, data, imagem e localização. Todos esses dados são armazenados de forma estruturada em um banco de dados relacional, possibilitando a criação de um acervo consultável e reutilizável.

Além da extração, o ENoW também registra falhas de coleta, como páginas indisponíveis, e evita a duplicação de dados já coletados, garantindo maior integridade ao conjunto de dados. Por ter sido projetado com uma arquitetura modular, permite a criação de múltiplos projetos com diferentes fontes e termos de busca, oferecendo flexibilidade e escalabilidade ao processo. Além disso, por ser uma ferramenta de código aberto e gratuita, o ENoW se apresenta como uma alternativa viável para pesquisas que necessitam de grandes volumes de dados jornalísticos de forma automatizada, sem depender de soluções comerciais com custo de operação.

2.7 Comparação

As ferramentas relacionadas que foram analisadas apresentam abordagens variadas para o mesmo problema, a raspagem de dados. Cada uma apresenta pontos positivos e negativos e, dependendo do contexto do uso, demonstram desempenhos diferentes. A escolha ideal irá depender da necessidade do usuário e de seu projeto.

O ParseHub e o Octoparse destacam-se por sua interface gráfica intuitiva, o que concede a escolha de atributos de maneira visual e, por consequência, permite que usuários sem conhecimentos de programação façam uso das ferramentas para seus casos de uso específicos. Além disso, outros facilitadores, como a identificação automática de padrões nos atributos escolhidos, criam atratividade através da usabilidade. No entanto, para projetos de médio e grande porte, a falta de adaptabilidade para mudanças e as limitações da versão gratuita podem se tornar um fator limitante.

Já o 80legs assume a forma de um raspador com o enfoque nos usuários técnicos e projetos complexos. Ele oferece maior flexibilidade, escalabilidade e desempenho para grandes volumes de dados. Seu modelo é altamente configurável, mas sua versão gratuita é bastante limitada. Essas características apontam para vantagem no uso comercial, oferecendo desempenho e funcionalidades exclusivas para projetos que requerem grandes extrações de dados periodicamente.

Ferramentas acadêmicas como o FactExtract e ENoW representam abordagens mais especializadas, focadas na coleta de notícias e portais de artigos. O FactExtract tem uma lógica de extração baseada em palavras-chave e fusão de afirmações, o que facilita o tratamento e

agregação dos dados coletados. Já o ENoW propõe uma arquitetura modular que facilita a reutilização dos dados. Por terem nascido em um âmbito acadêmico, ambos se destacam por sua ausência de custos e acesso ao código-fonte, apesar de demandarem maior esforço para configurações iniciais.

A Tabela 2.1 apresenta uma análise comparativa das ferramentas apresentadas neste capítulo.

Tabela 2.1: Comparação entre ferramentas de web scraping analisadas

Ferramenta	**Navegação Usuário**	**Entrada**	**Seleção de Atributos**	**Adaptabilidade**	**Fontes de Dados**	**Versão paga (benefícios)**	**Saída**
ParseHub	Visual (point-and-click)	Página inicial e caminho de navegação	Interface gráfica com sugestão automática de padrões	Baixa (sensível a mudanças estruturais no site)	Definida manualmente pelo usuário	IPs múltiplos, mais requisições por segundo, projetos ilimitados	JSON, CSV, Excel, API
80legs	Configuração via API (script)	Scripts de configuração via API	Programático, via script	Média (boa para grandes volumes, mas sensível a mudanças bruscas)	URLs e comportamento definidos via script	Paralelismo, requisições ilimitadas, extensão da API	JSON, formatos tabulares, API
Octoparse	Visual (point-and-click)	Página inicial e regras de extração	Interface gráfica com sugestão de padrões, templates prontos	Média (suporta AJAX, mas sensível a grandes mudanças)	Definida manualmente ou via templates	Mais tarefas simultâneas, IPs múltiplos, resolução de CAPTCHAs, backups, execução em nuvem	JSON, Excel, API, XML, HTML
Newspaper3k	Programática (Python)	URL da página de notícia	Extração automática via NLP	Alta (graças ao NLP, mas pode errar com páginas complexas)	Qualquer URL de artigo/notícia	Não possui (open source)	JSON (via código), objetos Python
FactExtract	Programática (projeto acadêmico)	Lista de websites + palavras-chave	Definida pelo módulo MEA, focado em artigos de texto	Alta (usa palavras-chave e fusão de dados)	Pré-definida via lista de sites e palavras-chave	Não possui (open source)	Estruturado com timestamps, autores, conteúdo limpo
ENoW	Programática (projeto acadêmico)	Metadados cadastrados + string de busca	Pré-definida por metadados + NLP via Newspaper3k	Alta (arquitetura modular, suporta múltiplos projetos)	Bancos de fontes cadastradas manualmente	Não possui (open source)	Banco de dados relacional, formato consultável

Capítulo 3

Desenvolvimento do Extrator de Atributos de Portais de Notícias

O cenário contemporâneo é marcado por uma profunda transformação digital que alterou de maneira significativa a produção, disponibilização e consumo de notícias e dados jornalísticos. Dentro do cenário acadêmico, a obtenção desses dados pode se mostrar útil em várias áreas de conhecimento, utilizando técnicas de Big Data para extração e análise de dados disponíveis na web.

3.1 ENoW

O ENoW é um sistema automatizado projetado para coletar, organizar, pré-processar e armazenar notícias provenientes de portais online. Utilizando técnicas de web scraping, o ENoW possibilita ao usuário extrair notícias a partir de palavras-chave e armazenar essas informações em um banco de dados estruturado, facilitando análises posteriormente. Diferente de muitas ferramentas pagas disponíveis no mercado, o ENoW oferece uma solução gratuita e flexível, proporcionando maior autonomia aos usuários. No contexto acadêmico, destaca-se como uma ferramenta que viabiliza e impulsiona pesquisas que demandam a coleta massiva e estruturada de informações jornalísticas.

3.1.1 Funcionamento da coleta

O funcionamento do ENoW tem várias etapas, incluindo etapas de pós-processamento e análise. No entanto, este capítulo se limitará à coleta e ao armazenamento, que são as partes relevantes para este trabalho. Dentro desse contexto, podemos definir o funcionamento em duas etapas.

O processo de utilização do ENoW começa com o cadastro das informações necessárias para a operacionalização do sistema. Nesta etapa, o administrador ou usuário responsável insere os metadados referentes aos portais de notícias a serem monitorados, contemplando informações como nome do site, URL, localização geográfica (estado e cidade) e detalhes técnicos relativos à estrutura das páginas, tais como as tags HTML que delimitam elementos como título, conteúdo, data e demais campos relevantes das notícias. Esses campos são essenciais para o funcionamento do scraping, com a definição de onde os elementos desejados podem ser localizados dentro dos websites cadastrados.

Na segunda parte, o usuário procede à criação de projetos no ENoW, cada um deles associado a um objetivo de pesquisa ou monitoramento específico. Dentro de um projeto, são

definidas uma ou mais palavras-chave (strings de busca), que orientam o processo de extração de notícias ao delimitar os temas de interesse. A associação entre projetos e portais de notícias é realizada de modo a possibilitar a seleção das fontes a serem consideradas na coleta, o que dá flexibilidade na configuração do sistema conforme as necessidades do usuário.

Após as etapas de cadastro e configuração, o ENoW viabiliza a execução da coleta automatizada de notícias. O sistema utiliza técnicas de web scraping para acessar os portais previamente selecionados e identificar, de forma automática, as notícias que contenham as palavras-chave estabelecidas no projeto.

Durante a coleta, o ENoW percorre as páginas dos sites, localizando e extraindo informações pertinentes, tais como título, corpo do texto, data de publicação, localização e imagens presentes no artigo. O sistema é capaz de interpretar diferentes estratégias de paginação, garantindo a abrangência da coleta mesmo em casos em que as notícias estejam distribuídas em múltiplas páginas.

As notícias extraídas são armazenadas em um banco de dados relacional, o que favorece a organização, o acesso e o posterior processamento dos dados. O usuário pode acompanhar o progresso da coleta e consultar as notícias já armazenadas, mantendo controle sobre as informações coletadas. Um sumário em forma de diagrama desse funcionamento pode ser observado na figura 3.1

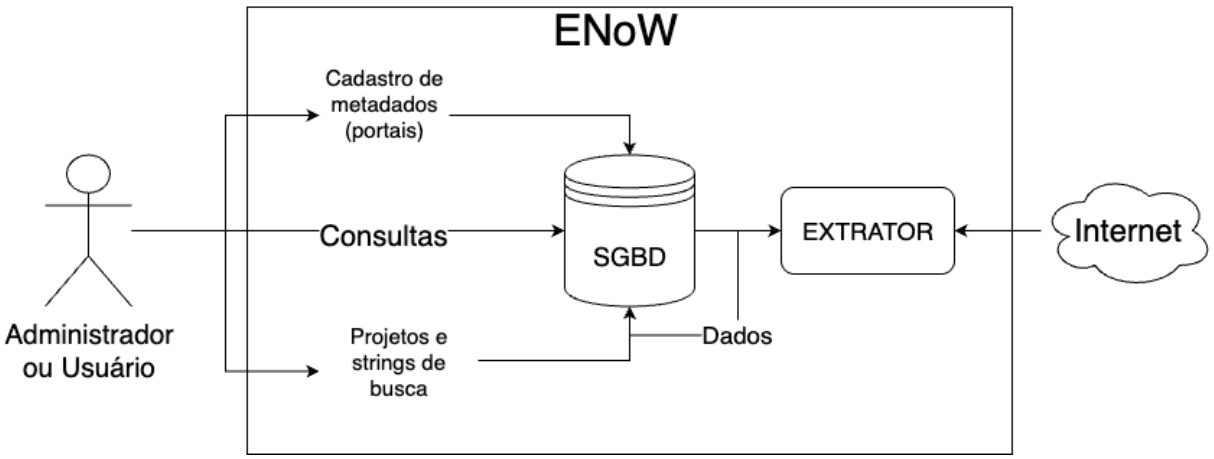


Figura 3.1: Diagrama sumarizado de funcionamento do ENoW

3.1.2 Cadastro de portais

Como já mencionado, o ENoW é um sistema completo e que realiza etapas que vão desde a extração da web até análises com os dados já armazenados. No entanto, uma das possíveis melhorias foi observada no processo de cadastro de metadados para portais de notícia. Esta é uma parte trabalhosa e manual e, portanto, passível de erros.

Entre outras tarefas, o cadastro de novos portais de notícia exige que o usuário, de maneira manual, vasculhe a estrutura HTML e registre na interface gráfica do ENoW os campos necessários, como exemplificado nas figuras 3.2 e 3.3.

Por ser um processo demorado, agilizar o cadastro de seleção de tags, além de melhorar a experiência do usuário ou administrador, também contribui na rápida expansão da base de portais cadastrados, contribuindo para que, com o mesmo tempo hábil, seja possível ampliar e extrair dados de outros portais, contribuindo para uma base de dados mais ampla.

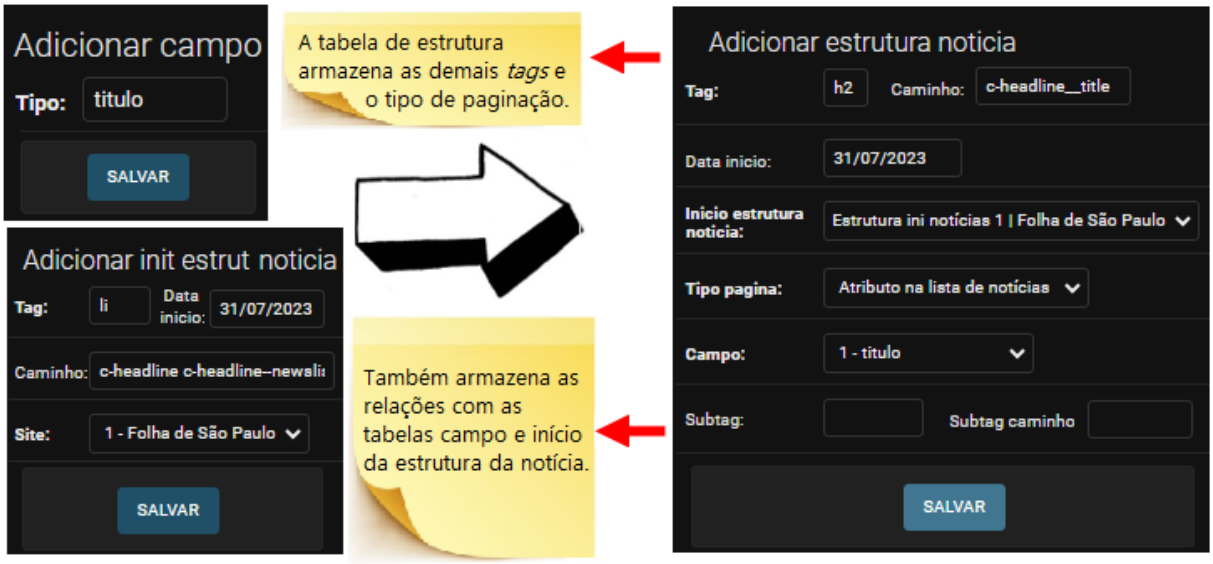


Figura 3.2: Cadastro da estrutura de notícia

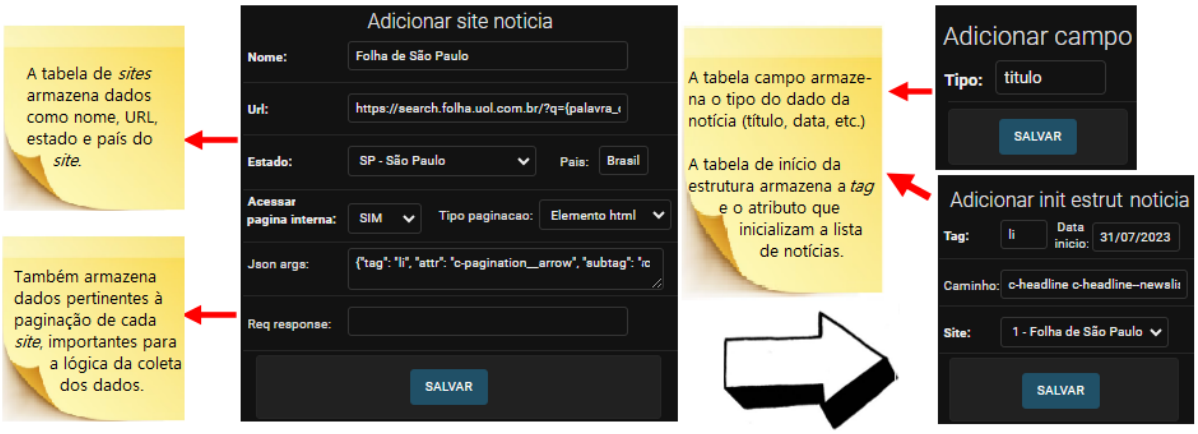


Figura 3.3: Cadastro da estrutura de notícia

3.2 Soluções

Com o objetivo de aprimorar o processo de cadastro de metadados para novos portais, foram consideradas duas opções para desenvolvimento e aprimoramento do ENoW. Ambas as abordagens tiveram desenvolvimento ativo para validação real da funcionalidade delas.

3.2.1 ENoW Selector

No processo de configuração de portais no sistema observou-se a necessidade recorrente de identificar e selecionar, com precisão, elementos HTML específicos das páginas web de interesse. A inspiração de ferramentas com interfaces visuais de apontar e clicar, tais como ParseHub e Octoparse, descritas nas seções 2.1 e 2.3 respectivamente, levou à ideia de ter uma funcionalidade similar, que permitisse a seleção visual dos atributos HTML vistos na página e facilitasse o fluxo de trabalho do usuário que cadastra novos portais de notícia, permitindo que até mesmo usuários de menor conhecimento técnico pudessem executar a tarefa, sem necessidade de vasculhar o código-fonte. Assim, o ENoW Selector foi idealizado.

O ENoW Selector é uma extensão para o navegador Google Chrome, e por consequência, todos os derivados do projeto open-source Chromium. Essa extensão permite que o usuário selecione de maneira visual um elemento HTML na página atual e, de forma automática, gere o caminho completo para ele, podendo ser usado para o cadastro no ENoW ou em outras ferramentas de raspagem de dados.

Ao ativar a extensão na barra de ferramentas do navegador, mostrada na figura 3.4, o usuário pode escolher entre dois modos de funcionamento:

Figura 3.4: Extensão dentro do navegador Google Chrome

- CSS: Gera caminho utilizado por sistemas CSS, podendo ser baseado nos identificadores únicos do elemento ou no caminho completo. Geralmente são mais performáticos na busca por um elemento;
- XPath: Uma linguagem de consulta padronizada para localizar elementos em documentos HTML e XML.

Depois da seleção do modo de funcionamento, o usuário habilita a extensão e volta a navegar na página, agora em um modo de inspeção. O cursor serve como seletor e, ao passar por cima de um elemento, o destaca na cor vermelha para melhor identificação do que está sendo selecionado, conforme mostra a figura 3.5.

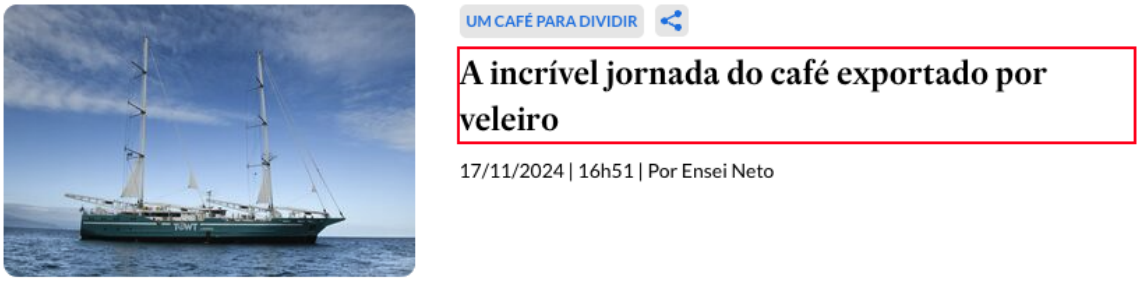


Figura 3.5: Elemento HTML destacado pelo ENoW Selector

Com o elemento identificado visualmente, um clique faz com que o caminho deste seja automaticamente copiado para a área de transferência do computador do usuário, mas também mostra de maneira visual, através de uma notificação pop-up, o texto que foi selecionado, ilustrado na figura 3.6. Essa funcionalidade permite que, de maneira ágil, o usuário possa alternar entre a

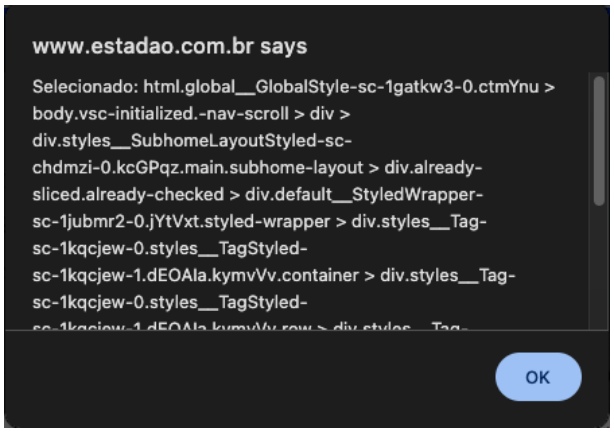


Figura 3.6: Caminho CSS para o elemento selecionado pela extensão

página de configuração do ENoW e o portal de notícias em análise, selecionando rapidamente os atributos e também removendo a necessidade de conhecimento do HTML.

A extensão foi construída a partir da API mais recente de extensões do Chrome (Manifest V3) e tem uma estrutura de projeto simples, exemplificada na figura 3.7. Ao adotar o ENoW

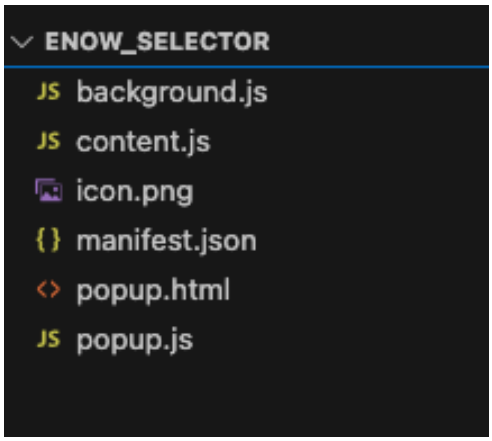


Figura 3.7: Estrutura de arquivos da extensão ENoW Selector

Selector como ferramenta auxiliar, o sistema ganha parte da usabilidade vista em ferramentas pagas, agilidade no cadastro de novas fontes de dados e também em acessibilidade para usuários menos técnicos, contribuindo em vários aspectos do projeto.

3.2.2 Aprimoramento com a biblioteca Newspaper

A biblioteca Newspaper foi apresentada na Seção 2.6 e é uma alternativa para fazer a extração de atributos dos portais de notícias, estendendo as funcionalidades do ENoW. O Newspaper é uma biblioteca consolidada para a extração automatizada de notícias e artigos em portais web, adotada em vários outros projetos de extração de dados. A biblioteca encapsula a lógica para funcionar de forma concisa. Com poucas linhas de código e utilizando poucos métodos, é possível fazer desde a requisição HTTP ao portal desejado até a extração de atributos como o título e análise com NLP.

O processo de coleta inicia-se com o módulo de *download*, responsável por requisitar e armazenar o HTML das páginas-alvo localmente. Para garantir robustez, mecanismos de *retry*,

timeout e detecção automática da codificação de caracteres são embutidos dentro do módulo, o que minimiza a possibilidade de falhas e garante maior compatibilidade.

Após ter o arquivo salvo localmente em memória, o módulo de *parsing* é ativado. Nessa etapa, a biblioteca lxml é usada para interpretar a estrutura do HTML. Diferentemente de abordagens baseadas apenas em seleção de tags, o Newspaper usa heurísticas que apostam na filtragem de conteúdo irrelevante da página, como menus, barras laterais e anúncios, e então, através de análises de densidade e aninhamento dos elementos, deduzir o núcleo de interesse da notícia: título, subtítulo, texto, imagens e outros.

As heurísticas de extração adotadas pelo Newspaper são o seu grande diferencial, projetadas para identificar e isolar automaticamente o conteúdo principal de artigos jornalísticos, mesmo diante de poucas semelhanças entre portais de notícias. Ao contrário de abordagens que dependem da demarcação prévia de atributos HTML, essa abordagem baseada em análise de padrões e características textuais confere maior autonomia e adaptabilidade ao processo de raspagem.

A Figura 3.8 ilustra os passos executados pelo processador de artigos da biblioteca, que são detalhados na sequência.

- Primeiramente é realizada a limpeza no documento, removendo elementos periféricos considerados irrelevantes, como scripts, menus de navegação, rodapés e anúncios. Essa filtragem preliminar é essencial para reduzir o ruído do documento nas etapas seguintes.

- Em seguida, são aplicados algoritmos de avaliação da densidade textual dos diferentes blocos HTML. Elementos com alta concentração de texto contínuo e poucas interrupções por links e imagens, centralizados no documento, recebem maior pontuação. Além da densidade, a hierarquia dos elementos e utilização de tags HTML específicas, assim como IDs do elemento, podem ser considerados na análise.

- São empregadas técnicas de normalização e limpeza, removendo itens como quebras de linha excessivas, espaços em branco e caracteres especiais que possam ter sido gerados pelo site da extração.

- Por fim, para cada elemento, são aplicados métodos reservas de extração, caso não seja possível obter o elemento pelo método favorito.

- Opcionalmente, é possível aplicar técnicas de NLP para sumarização ou classificação de conteúdo.

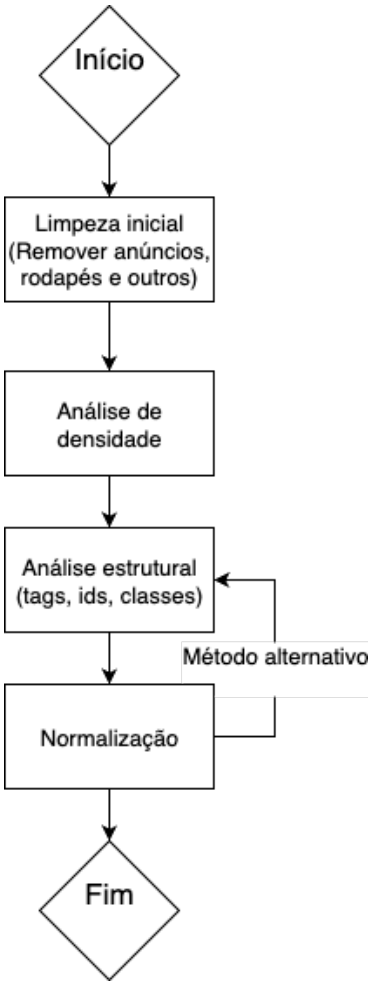


Figura 3.8: Diagrama sumarizado de funcionamento do Newspaper

Um exemplo, em pseudocódigo, da lógica de extração do título de uma notícia na biblioteca Newspaper se encontra no Algoritmo 1.

Algoritmo 1 Extração do título da notícia

1: Obtém o conteúdo da tag `<title>` e armazena em `title_text`
2: **if** `title_text` é inexistente **then**
3: **return** string vazia
4: **end if**
5: Extrai o maior texto de uma tag `<h1>` com mais de duas palavras e armazena em `title_text_h1`
6: Extrai o metadado `og:title`, se disponível, e armazena em `title_text_fb`
7: Normaliza as três versões (`title_text`, `title_text_h1`, `title_text_fb`) removendo pontuação e caixa alta
8: Compara as versões normalizadas e escolhe a mais informativa
9: **if** nenhuma substituição foi feita **e** `title_text` contém delimitadores comuns (ex.: "|", "-", "–") **then**
10: Divide o `title_text` usando `title_text_h1` como referência para selecionar o segmento mais relevante
11: **end if**
12: Aplica limpeza final ao texto selecionado
13: **if** versão final $\approx$ `title_text_h1` **then**
14: **return** `title_text_h1`
15: **else**
16: **return** versão final
17: **end if**

No contexto do ENoW, a biblioteca Newspaper pode complementar, ou até mesmo substituir a seleção de atributos no cadastro de website, permitindo a extração automática de informações como título, corpo do texto, data de publicação e imagem principal da notícia. Com isso, torna-se possível dispensar a configuração individual de seletores CSS ou XPath para cada portal, o que reduz significativamente o tempo de configuração e manutenção do sistema. Essa abordagem se mostra especialmente eficaz em sites com estrutura consistente e bem formatada, proporcionando maior escalabilidade ao sistema sem comprometer a qualidade dos dados coletados.

Contudo, a biblioteca Newspaper apresenta algumas limitações que devem ser consideradas. Sua eficácia depende fortemente da estrutura e da qualidade do HTML das páginas analisadas. Em sites com marcação inconsistente, uso excessivo de JavaScript, conteúdo carregado dinamicamente ou Paywalls, a extração pode falhar ou retornar informações incompletas.

Capítulo 4

Conclusão

Este trabalho apresentou o desenvolvimento e aprimoramento do ENoW, uma ferramenta gratuita e flexível voltada à coleta automatizada de notícias a partir de portais online. Por meio da aplicação de técnicas de web scraping e da organização estruturada das informações em banco de dados, o ENoW mostrou-se uma alternativa viável e eficaz para atender às demandas acadêmicas que envolvem análise de dados jornalísticos em larga escala, mesmo quando comparado a ferramentas pagas.

Durante o processo de desenvolvimento, foram propostas duas soluções complementares para melhorar a experiência do usuário e aumentar a escalabilidade do sistema: a extensão ENoW Selector e a integração com a biblioteca Newspaper. A primeira, inspirada em ferramentas visuais como o ParseHub e o Octoparse, viabilizou a seleção visual de atributos HTML, reduzindo o tempo necessário para configurar novos portais. Já a segunda trouxe uma abordagem heurística automatizada, capaz de extrair informações relevantes mesmo em sites sem estrutura uniforme, diminuindo a dependência da configuração manual.

A análise do funcionamento e das limitações de ambas as abordagens evidenciou que a combinação entre extração manual assistida e métodos automáticos oferece maior flexibilidade e robustez ao sistema. Assim, o ENoW passa a contar com um conjunto de ferramentas que ampliam sua capacidade de adaptação a diferentes cenários de coleta, fortalecendo seu papel como uma plataforma de apoio à pesquisa em tempos de abundância de informação disponível na internet.

4.1 Trabalhos futuros

Como continuação deste trabalho, pretende-se explorar de forma mais aprofundada o impacto prático das soluções implementadas no ENoW. Uma investigação que propõe a realização de uma análise comparativa entre os dados extraídos manualmente via seletores HTML e aqueles obtidos automaticamente pela biblioteca Newspaper. Essa comparação poderá oferecer evidências mais objetivas sobre a precisão e a completude de cada abordagem, além de indicar cenários em que o uso híbrido seja mais vantajoso.

Outra frente relevante consiste na mensuração do ganho de produtividade proporcionado pelo uso da extensão ENoW Selector. Por meio de testes controlados com usuários de diferentes níveis técnicos, seria possível quantificar a redução no tempo de cadastro de portais e avaliar o impacto na curva de aprendizagem do sistema.

Além disso, é considerado o desenvolvimento de uma ferramenta auxiliar capaz de detectar automaticamente mudanças de estrutura (schema) nas páginas web já cadastradas, alertando o usuário sempre que alterações relevantes possam comprometer a extração dos dados.

Essa funcionalidade aumentaria a robustez do sistema frente à natureza dinâmica da web e, lateralmente, ainda forneceria uma ferramenta de decisão ao sistema sobre qual abordagem escolher.

Por último, planeja-se a criação de uma camada de integração entre o ENoW, a biblioteca Newspaper e buscadores como Google News, permitindo a coleta automatizada de notícias a partir de consultas baseadas em palavras-chave, eliminando a necessidade de cadastrar manualmente cada portal e aumentando a abrangência do sistema para a mesma de um grande buscador.

Referências Bibliográficas

ANSOLABEHERE, S., LESSEM, R., and SNYDER, J. M. The orientation of newspaper endorsements in u.s. elections, 1940-2002. *Quarterly Journal of Political Science*, 1(4), 2006.

KALOGEROPOULOS, A., SUITER, J., UDRIS, L., and EISENEGGER, M. News media trust and news consumption: Factors related to trust in news in 35 countries. *International journal of communication*, 13:22, 2019.

PARK, E., PARK, J., and HU, M. Tourism demand forecasting with online news data mining. *Annals of Tourism Research*, 90:103273, 2021.

REIPS, L., MUSICANTE, M., VARGAS-SOLAR, G., POZO, A. T. R., and HARA, C. S. Enow-extrator de dados de notícias da web. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 78–83, Belo Horizonte, MG, Brasil, 2023. SBC.

SARR, E. N., SALL, O., and DIALLO, A. Factextract: automatic collection and aggregation of articles and journalistic factual claims from online newspaper. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 336–341, Valencia, Spain, 2018. IEEE.

VARGAS-SOLAR, G., ZECHINELLI-MARTINI, J.-L., ESPINOSA-OVIEDO, J. A., and VILCHES-BLÁZQUEZ, L. M. Laclichev: Exploring the history of climate change in latin america within newspapers digital collections. In *European Conference on Advances in Databases and Information Systems*, pages 121–132, Tartu, Estonia, 2021. Springer.