ENoW - Extrator de Dados de Notícias da Web

Lisiane Reips[1], Martin Musicante[2], Genoveva Vargas-Solar[3], Aurora T. R. Pozo[1], Carmem S. Hara[1]

[1] Universidade Federal do Paraná, DInf, Curitiba-PR

[2]Universidade Federal Rio Grande do Norte, DIMAp, Natal-RN

[3]CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69221, France
{lisiane.reips,aurora.pozo,carmemhara}@ufpr.br,
mam@dimap.ufrn.br, genoveva.vargas-solar@cnrs.fr

Abstract. *Data available on the Web play a determining role in decision-making both in personal and corporate life. Collecting and storing this data in a structured model helps integrate them with other sources and then use the dataset in various applications, such as event detection and sentiment monitoring. Online newspapers are essential sources of information, accessed daily by thousands of people. To facilitate the exploration of this data, this paper presents ENoW - News Data Extractor from the Web. ENoW receives search strings as input and stores in a relational database data extracted from the news as well as their full content. The system was implemented in Python, using Web scraping techniques. The demonstration comprises the three main functionalities of the tool: newspaper registration, project registration and news extraction.*

Resumo. *Os dados disponíveis na Web desempenham um papel determinante nas tomadas de decisão, sejam elas pessoais ou corporativas. A coleta e armazenamento destes dados de forma estruturada permite que eles sejam integrados com outras fontes e utilizados em diversas aplicações, tais como detecção de eventos e monitoramento de sentimento. Os jornais online são importantes fontes de informação, que são acessados diariamente por milhares de pessoas. Para facilitar a exploração destes dados, este artigo apresenta o ENoW - Extrator de Dados de Notícias da Web. O ENoW aceita como entrada strings de busca e armazena em uma base de dados relacional dados extraídos das notícias, bem como o texto da notícia em sua íntegra. O sistema foi implementado na linguagem Python, utilizando técnicas de Web Scraping. A demonstração apresenta as três principais funcionalidades da ferramenta: cadastro de jornais, cadastro de projetos e coleta de notícias.*

1. Introdução

Os dados tornaram-se essenciais na vida cotidiana. Sua utilização em diversas aplicações requer que eles sejam coletados, limpos e armazenados [Johnson 2014]. Na *Web*, existem diversos tipos de dados, dentre os quais estão os de jornais de notícias *online*. Estes dados podem ser aplicados em diferentes áreas, incluindo o turismo e identificação de áreas afetadas por desastres naturais [Park et al. 2021, Franceschini et al. 2022]. Independentemente da aplicação, a extração de informações relevantes é fundamental, podendo ela ser realizada manualmente ou de forma automatizada.

A extração de forma automatizada reduz a intervenção humana e facilita a coleta e análise de dados da *Web*, contando com o auxílio de ferramentas como o *Web Scraping*. Essas ferramentas extraem dados que podem ser armazenados em diferentes formatos, como em bancos de dados ou arquivos CSV [Salem and Mazzara 2020].

O processo de *Web Scraping* envolve a análise, rastreamento da fonte *online* e organização dos dados [Krotov et al. 2020]. A etapa de análise requer conhecimento básico da arquitetura HTML do *site*, enquanto a de rastreamento envolve o desenvolvimento de um código para extrair os dados desejados de forma automatizada. A etapa de organização permite limpar e analisar os dados coletados [Krotov et al. 2020].

Comumente, em casos de extração de dados de notícias *online*, realiza-se uma busca por uma *string* na caixa de pesquisa presente na página, resultando em uma lista de notícias com títulos, descrições, imagens e datas. O processo de *Web Scraping* pode localizar essa lista e identificar as *tags* do documento HTML que contêm as informações desejadas a fim de extraí-las.

Foi realizada uma investigação sobre as ferramentas de *Web Scraping* disponíveis a fim de determinar a existência de uma que atenda aos requisitos dos projetos em andamento na instituição. Estes requisitos são:

a) coletar dados de notícias de um conjunto de URLs;

b) obter toda a lista de notícias, com suas páginas subsequentes, retornadas a partir da *string* pesquisada;

c) acessar cada notícia individualmente e coletar informações ausentes na lista;

d) armazenar os metadados e os dados coletados em um banco de dados;

e) registrar no banco de dados a data de indisponibilidade de páginas *online*, evitando afetar outras coletas de fontes de notícias no conjunto de URLs;

f) programar coletas para atualizar periodicamente a base de dados.

A análise revelou que existem ferramentas que possuem funcionalidades de coleta de forma gratuita, e funcionalidades de agendamento de coletas de forma paga, como por exemplo, o *ParseHub*[1], *80legs*[2], *Octoparse*[3] e *FactExtract* [Sarr et al. 2018]. Mas elas armazenam apenas os dados coletados, e não os metadados das páginas de jornais. Isso motivou o desenvolvimento do Extrator de Dados de Notícias da *Web* - ENoW, uma ferramenta desenvolvida em *Python*, que permite extrair dados estruturados das notícias de páginas de jornais *online*, armazenando, além do conteúdo das notícias, metadados das páginas de jornais, mantendo a proveniência e o histórico dos dados coletados.

Este artigo demonstra as principais funções do ENoW, que são: o cadastro de jornais, o cadastro de projetos e a coleta de notícias. Os dados extraídos, incluindo título, descrição, conteúdo completo, data, localização e imagem, foram armazenados em um SGBD relacional. A demonstração do ENoW mostra a eficácia na extração de notícias de páginas de jornais *online*.

[1]https://www.parsehub.com/
[2]https://80legs.com/
[3]https://www.octoparse.com/

O restante do artigo está estruturado da seguinte forma. A Seção 2 aborda os trabalhos relacionados à extração de dados. A Seção 3 descreve a arquitetura do ENoW e sua implementação. A Seção 4 apresenta as funcionalidades da ferramenta. Finalmente, na Seção 5 são apresentadas as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Existem diversas ferramentas disponíveis na *Web* para extração de notícias de forma automatizada. Dentre elas podem ser citadas o *ParseHub*, *80legs*, *Octoparse* e *FactExtract*.

O *ParseHub* permite extrair dados de *sites*, que podem ser selecionados por meio de cliques e armazenados na nuvem, nos formatos JSON e CSV. Ele também permite a coleta com base em uma sequência de caracteres escolhida. Sua versão gratuita limita a extração a 200 *sites* por vez, e a programação de coletas agendadas está disponível somente na versão paga.

O *80legs* permite extrair dados de *sites* de forma sequencial. Sua versão gratuita oferece a extração de dados de até 10.000 URLs, com a opção de *download* nos formatos CSV e JSON. Recursos de personalização, como a possibilidade de extrair uma ampla variedade de dados, agendar coletas, obter *tags* HTML, acessar códigos de *script* e realizar buscas por *strings*, estão disponíveis apenas nas suas versões pagas.

O *Octoparse* permite extrair dados de *sites* por meio de cliques, resultando na criação de um fluxo de trabalho. Esse fluxo pode ser modificado durante o processo, conforme necessário. Ele também permite a inclusão de uma lista de URLs e uma lista de *strings* para pesquisa, extraindo tudo o que for encontrado nas páginas solicitadas pelo usuário. São oferecidos modelos pré-construídos de *Web Scraping*, como modelos com a interface da *Amazon*, *Google*, *Google Maps*, *eBay* e *Twitter*.

O *FactExtract* realiza a extração de quinze *sites* específicos de notícias senegaleses. Ele detecta idiomas automaticamente e atende a dez deles. Os dados extraídos são processados, limpos e analisados. O resultado é armazenado em arquivos CSV e em um banco de dados. Além disso, o *FactExtract* não armazena dados duplicados e possui um recurso de automação de execução, que realiza verificações diárias nos artigos dos quinze *sites* envolvidos. A ferramenta foi desenvolvida na linguagem de programação *Python*, usando a biblioteca *Newspaper*. A análise constatou que as ferramentas atendem parcialmente às necessidades na versão gratuita, mas não são suficientes para atender os requisitos elencados.

3. A Ferramenta ENoW

Para atender às necessidades anteriormente mencionadas, foi desenvolvido o ENoW (Extrator de Dados de Notícias da *Web*). A sua arquitetura é apresentada na Figura 1. Os três principais módulos da arquitetura são: o cadastro de *sites* de noticias, o cadastro de projetos e de *strings* de busca e o coletor de notícias.

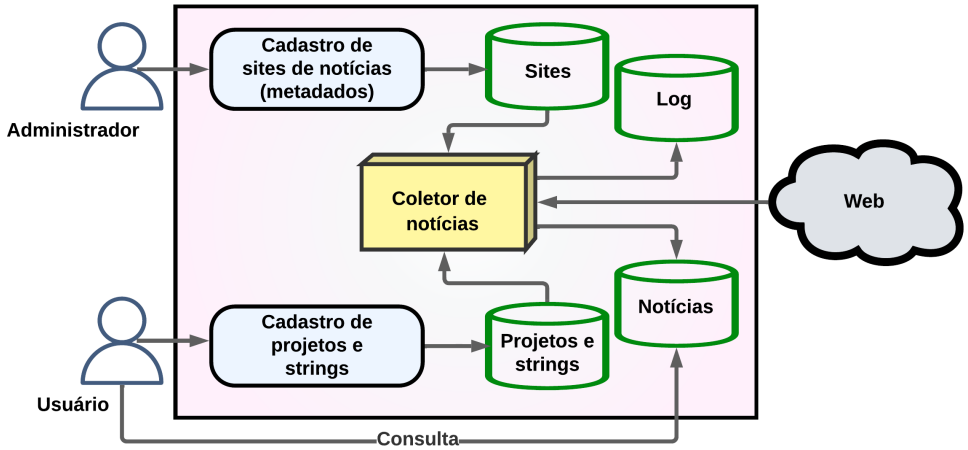

Figura 1. Arquitetura Geral do ENoW.

A função do módulo de cadastro de sites de notícias é alimentar a base de *Sites*, que contém informações sobre jornais *online*, que incluem dados sobre como extrair as informações de interesse. Esta base é composta pelas tabelas *sites* de notícias, campo, início da estrutura de notícias e estrutura geral de notícias. O módulo de cadastro de projetos é necessário porque o ENoW considera que podem existir diversos projetos, cada um associado a um conjunto de *strings* de busca e a um conjunto de jornais dos quais se deseja extrair as notícias. Estas informações são armazenadas na base de *Projetos e Strings*. Após o cadastro do projeto, o coletor de notícias pode ser iniciado. Os dados coletados na *Web* são armazenados em uma base de dados de *Notícias*. Os dados extraídos de cada notícia incluem o título, descrição (resumo), data, localização, imagem e texto completo. Cada entrada nesta base é também associada ao jornal do qual foi extraído e um ou mais projetos. Um histórico de coletas também é armazenado em uma base de *Logs*. Cada entrada nessa base faz associação às notícias coletadas, armazenando a data da coleta e a disponibilidade de cada página *online*.

O ENoW foi implementado utilizando módulos de extração de dados na linguagem de programação *Python*. Foram utilizadas as bibliotecas *BeautifulSoup*[4], *Selenium* [5] e *Python Newspaper* [Bansal et al. 2014]. A biblioteca *BeautifulSoup* analisa os elementos HTML de uma página e cria uma árvore de análise para extrair os dados. A biblioteca *Selenium* simula interações de um usuário com uma página da *Web*, permitindo a manipulação e inserção de dados. A biblioteca *Python Newspaper* extrai dados no estilo de jornal, incluindo textos e imagens, realizando uma limpeza prévia dos dados e removendo anúncios. O sistema gerenciador de banco de dados utilizado foi o *MySQL*. O ENoW foi implementado utilizando a interface do *framework Django*, visando melhorar a integração e visualização dos dados coletados.

[4]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[5]https://www.selenium.dev/documentation/

4. Demonstração da ferramenta

A demonstração do ENoW[6] contempla a execução dos três módulos que compõem a ferramenta. O cadatro de *sites* é executado pelo administrador da ferramenta e envolve a inserção de dados em quatro tabelas distintas no banco de dados. A Figura 2 apresenta um exemplo das inserções dos dados relacionados à página de notícias "Folha de São Paulo". Como pode ser observado na figura, a extração de campos das notícias requer que o administrador informe o seu caminho completo no documento HTML. Atualmente, o ENoW possui 42 jornais cadastrados.

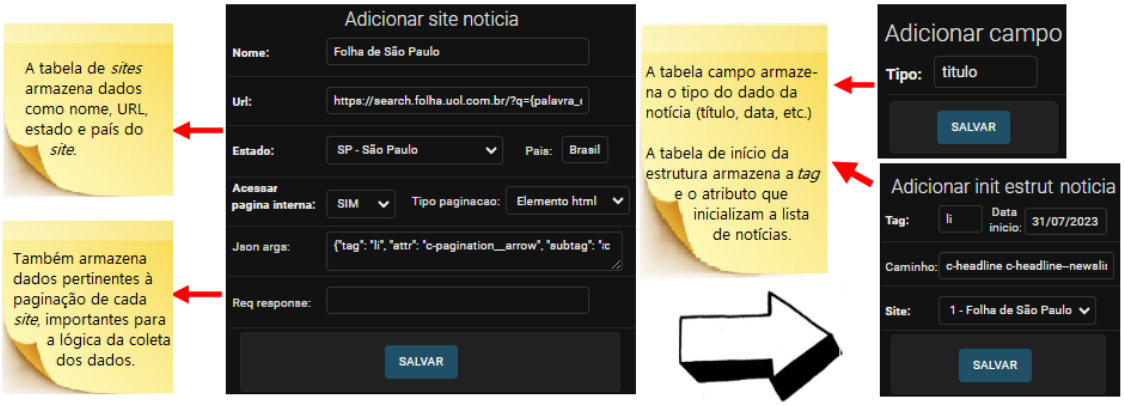

Figura 2. Fluxo da inserção de dados de uma página de notícias *online*.

O cadastro de projetos envolve a escolha dos jornais de interesse e dos strings de busca, como ilustrado na Figura 3. Este módulo é utilizado pelo *usuário final* e a demostração vai permitir a criação de novos projetos, bem como a adição de novos strings de busca a projetos já criados.

Figura 3. Inserção de projeto, de *strings* de busca e escolha das páginas de notícias *online*.

Na sequência, a coleta pode ser realizada e os dados coletados são armazenados na base de dados de notícias. A Figura 4 apresenta a execução da coleta, via interface do *framework*, e os dados coletados, via banco de dados.

[6]Vídeo disponível em https://youtu.be/PQgK5A7XUSM

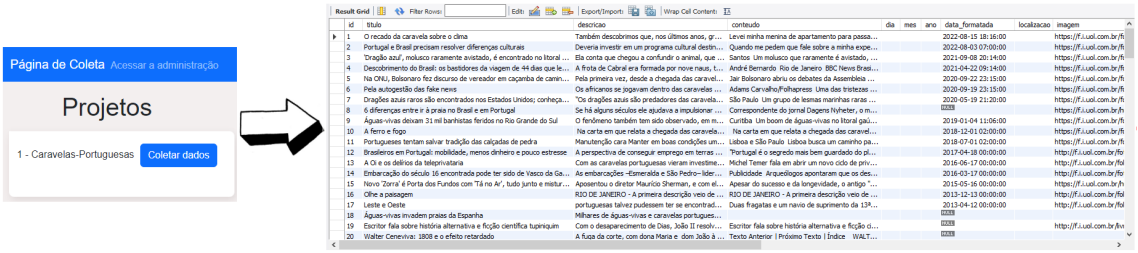

Figura 4. Realização da coleta e visualização dos dados coletados.

5. Conclusão e Trabalhos Futuros

Este artigo apresentou a ferramenta ENoW, um extrator de dados de notícias da *Web*. Conforme a demonstração, o ENoW está operacional para a realização das coletas, armazenando tanto as informações de *sites* de jornais bem como os dados coletados em uma base de dados relacional. O sistema mantém a proveniência das notícias, relacionando-as aos *sites* dos quais foram extraídas e mantendo o histórico das coletas. Atualmente o agendamento das coletas está em desenvolvimento.

Como trabalho futuro, planeja-se trabalhar o gerenciamento da capacidade de armazenamento do ENoW e realizar a curadoria dos dados coletados para fornecer uma base organizada e limpa ao tema de interesse do usuário. Isso envolve padronizar a representação das datas e localização nas notícias e fornecer apenas informações relevantes para as palavras-chave pesquisadas pelo usuário, eliminando dados não pertinentes.

Referências

Bansal, A., Chaudhury, S., Roy, S. D., and Srivastava, J. (2014). Newspaper article extraction using hierarchical fixed point model. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 257–261. IEEE.

Franceschini, R., Rosi, A., Catani, F., and Casagli, N. (2022). Exploring a landslide inventory created by automated web data mining: the case of italy. *Landslides*, 19(4).

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, 16:263–274.

Krotov, V., Johnson, L., and Silva, L. (2020). Tutorial: Legality and ethics of web scraping. *Communications of the Association for I nformation Systems*.

Park, E., Park, J., and Hu, M. (2021). Tourism demand forecasting with online news data mining. *Annals of Tourism Research*, 90:103273.

Salem, H. and Mazzara, M. (2020). Pattern matching-based scraping of news websites. In *Journal of Physics: Conference Series*, page 012011. IOP Publishing.

Sarr, E. N., Ousmane, S., and Diallo, A. (2018). 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. IEEE.