CONO: Um Coletor Automatizado de Notícias sobre Corrupção em Santa Catarina

Ana Clara Stupp de Souza[1], **Carina F. Dorneles**[1]

[1]Instituto de Informática – Universidade Federal de Santa Catarina (UFSC)

ana.stupp@grad.ufsc.br, carina.dorneles@ufsc.br

Abstract. *In the information age, staying informed is essential, but the overwhelming volume of news makes it challenging to extract relevant data efficiently. The Public Ministry of Santa Catarina depends on strategic information to conduct investigations but faces difficulties due to information overload. To address this, the CONO tool was developed—a crawler built with JavaScript and Python to automatically collect and filter corruption-related news. This article presents the tool's development process, underlying technologies, operational workflow, and the results achieved in streamlining data collection for investigative purposes.*

Resumo. *Na era da informação, estar bem informado é essencial, mas o volume avassalador de notícias torna difícil extrair dados relevantes de forma eficiente. O Ministério Público de Santa Catarina depende de informações estratégicas para conduzir investigações, mas enfrenta dificuldades devido à sobrecarga de informações. Para resolver isso, foi desenvolvida a ferramenta CONO — um crawler construído com JavaScript e Python para coletar e filtrar automaticamente notícias relacionadas à corrupção. Este artigo apresenta o processo de desenvolvimento da ferramenta, as tecnologias subjacentes, o fluxo operacional e os resultados alcançados na otimização da coleta de dados para fins investigativos.*

1. Introdução

A crescente quantidade de informações disponíveis diariamente torna essencial o uso de ferramentas que possibilitem o monitoramento eficaz de notícias sobre temas específicos. Um coletor de dados focado [Chakrabarti 2009] permite a análise contínua da imprensa, a filtragem de informações relevantes e a integração com bases de dados para correlação e detecção de padrões. No contexto de órgãos de fiscalização como o Ministério Público, essas ferramentas podem acelerar a apuração de denúncias, reduzir custos operacionais, otimizar investigações e reforçar a transparência. A automação do processo aprimora a detecção e resposta a casos de corrupção, contribuindo para a proteção dos recursos públicos e o combate a fraudes de forma mais eficiente.

O excesso de notícias prejudica o trabalho de órgãos de fiscalização, que dependem de dados estratégicos para tomar decisões. Embora sites de notícias ofereçam filtros para temas políticos, nenhum em Santa Catarina possui um filtro específico para notícias relacionadas à corrupção. A busca pela palavra ”corrupção”nem sempre é eficaz, já que muitas notícias relevantes não contêm esse termo. Além disso, a busca manual em

múltiplos sites é ineficiente e onerosa. A partir dessa necessidade, surgiu a ideia de desenvolver uma ferramenta automatizada para coletar dados estratégicos para o Ministério Público de Santa Catarina.

Este trabalho tem como objetivo apresentar o CONO (Coletor de Notícias), uma ferramenta que possibilita o acesso a notícias relacionadas à corrupção. O CONO é uma ferramenta construída com JavaScript e Python, que coleta e filtra automaticamente as notícias e as disponibiliza através de uma interface Web. Os sites de notícias utilizados para a coleta foram os disponibilizados na categoria Jornais de Santa Catarina do site Guia de Midia[1], sendo visitados duas vezes ao dia em horários dispersos, buscando potencializar a coleta.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia de desenvolvimento, enquanto a Seção 4 apresenta os resultados obtidos. A Seção 5 traz discussões sobre os diferenciais da ferramenta e possíveis melhorias e a Seção 6 conclui o artigo e oferece possíveis direções para trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, são apresentados alguns trabalhos relacionados, que são divididos em três grupos principais: ferramentas de coleta de dados (*web crawlers*), aplicações de Processamento de Linguagem Natural (PLN) e monitoramento de notícias. Não foram encontrados na literatura trabalhos que apresentam propostas de *crawler* para coleta de notícias focadas em temas de corrupção. O trabalho de [Reips et al. 2023] apresenta o desenvolvimento de um *crawler* direcionado à coleta de notícias de forma genérica, não focada. A ferramenta, intitulada ENoW, foi desenvolvida em Python, utilizando as bibliotecas BeautifulSoup4, Selenium e Python Newspaper. Ela armazena os metadados das notícias, bem como seu texto na íntegra, em um banco de dados relacional MySQL. O trabalho apresentado em [Moraes 2019] trata da análise textual de notícias falsas, a fim de buscar padrões na escrita que possam ajudar a identificar as chamadas *Fake News*. Foram utilizadas as bibliotecas scikit-learn, NLTK e spaCy para processamento da linguagem natural e aprendizado de máquina, além de outras bibliotecas de classificação, indexação e visualização, resultando em um classificador, capaz de identificar a veracidade de uma notícia a partir da estrutura de seu texto. No trabalho de [de Paula 2022], é demonstrada uma abordagem para automatização de coleta de notícias utilizando a tecnologia de *web crawlers*, além da criação de um modelo de aprendizado de máquina para classificação das notícias por nível de relevância. Bases de dados foram criadas com o auxílio do *framework scrapy* e diversos classificadores foram testados, buscando o melhor desempenho possível na classificação de relevância.

3. CONO - (Coletor de Notícias)

O *crawler* foi desenvolvido em *JavaScript*, utilizando a biblioteca *Playwright* para acessar os sites de notícias e *Readability* e *JSDOM* para extrair o texto puro. A API em *Python* usa *SpaCy* para identificar nomes de pessoas e empresas nas notícias e *Flask* para comunicação com o frontend. O frontend, desenvolvido em *React*, utiliza *react-bootstrap*

[1] https://www.guiademidia.com.br/jornaisdesantacatarina.htm

para melhorar a visualização dos dados. A Figura 1 apresenta o *workflow* do CONO, que consiste de validações realizadas a fim de decidir pela coleta ou não de uma notícia.

Analisando a estrutura dos sites, verifica-se que o link das notícias sempre contém o seu título, então o mesmo é comparado com uma lista de palavras relacionadas à corrupção para detectar o contexto da notícia antes mesmo de acessá-la. Uma vez identificada alguma palavra no título da notícia, o *crawler* verifica se a URL consta na lista de URLs já coletadas, presente no arquivo *processed_urls.json*. Em caso negativo, a ferramenta segue o seu processo, acessando a página da notícia, coletando seu conteúdo interno e normalizando-o, removendo tudo que se refere à estrutura HTML. Com o texto íntegro da notícia, o *crawler* verifica se alguma cidade de Santa Catarina é mencionada. Se nenhuma cidade é mencionada, entende-se que a notícia não possui relevância para o estado, podendo-se tratar de algum caso de corrupção em esfera federal ou internacional, sendo então descartada. Com a detecção da menção de alguma cidade, a notícia é coletada e inserida no arquivo *corruption_articles.json* (conforme estrutura apresentada no *output* da figura), além de ser disponibilizada no *frontend* da ferramenta.

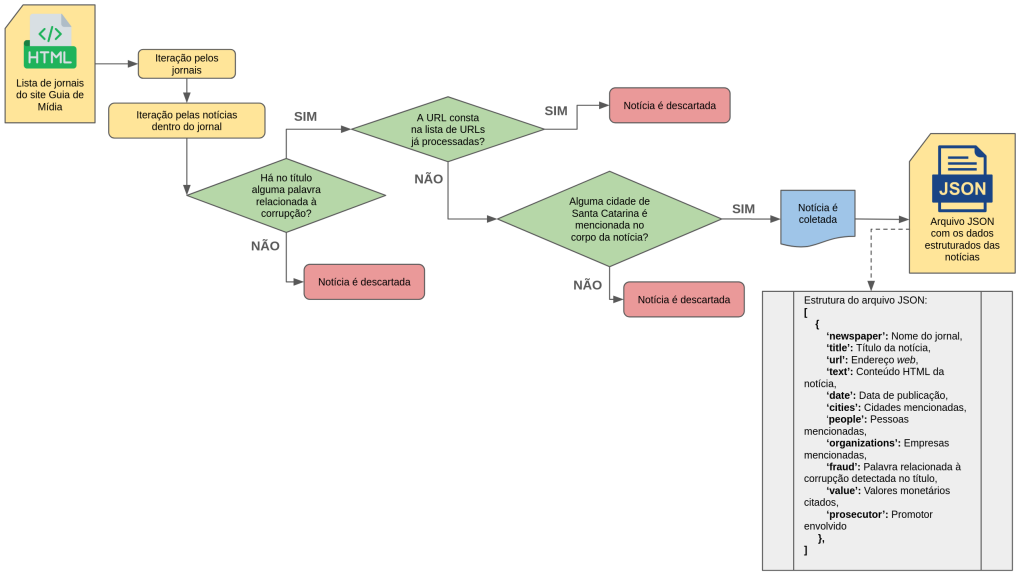

Figura 1. Fluxo de coleta do *crawler*

3.1. Detalhamento da coleta

As notícias são coletadas de forma estruturada, destacando informações que, estrategicamente, são mais importantes. O nome do jornal e o título da notícia são extraídos da URL e armazenados nos campos *newspaper* e *title*. A data é coletada através da busca por classes com nome date dentro do corpo da página (*.date, .published-date, .published* entre outras), e fica armazenada no campo *date*. O texto da notícia é então percorrido para identificar menções às cidades de Santa Catarina. Durante esse percurso, quando uma cidade é mencionada, ela é adicionada a uma lista, chamada *foundSCCities*. Essa lista é armazenada no campo *cities* do *JSON*. Para coletar o nome de pessoas e empresas mencionadas, o CONO envia o texto da notícia para a API, que realiza o processamento do texto com o *SpaCy*[2] e retorna os nomes identificados nas variáveis *people* e *organizations*. Contudo,

[2]https://spacy.io/

antes de armazenar o conteúdo das variáveis, é realizada uma verificação para remover nomes como Dionísio Cerqueira, que no contexto de Santa Catarina se trata de um nome de cidade. Em seguida, o conteúdo delas é armazenado nos campos *people* e *organizations* do *JSON*. O campo *fraud* armazena as palavras referentes à corrupção detectadas no título da notícia. Os campos *value* e *prosecutor* armazenam as menções a valores monetários e o nome do promotor envolvido, respectivamente. Ambas as informações são coletadas através de expressões regulares.

4. Discussão

Durante o desenvolvimento do CONO, diversos desafios foram enfrentados. Um deles foi a estruturação dos sites de notícias, que não seguia um padrão. As informações frequentemente estavam em classes diversas do HTML, impossibilitando o desenvolvimento e aplicação de uma coleta 'genérica'. Foi necessário analisar a fundo a estrutura de cada site e adaptar a coleta para as suas particularidades, sem prejudicar a coleta em outros sites de estrutura diferente. Outro desafio foi a coleta dos dados dentro do texto íntegro das notícias. A ideia inicial era desenvolver toda a ferramenta utilizando JavaScript e suas bibliotecas, mas nenhuma se mostrou eficiente na detecção de nomes na língua portuguesa, então foi necessário recorrer ao *SpaCy*. Para utilizar o *SpaCy* juntamente com o *JavaScript* foi necessário criar uma aplicação web em *Flask* para armazenamento e execução das requisições, o que gerou a necessidade de reestruturação do fluxo de funcionamento do CONO. Para além dos desafios, a ferramenta traz inovações como a aplicação do frontend em React. Com essa biblioteca foi possível trazer uma visualização limpa dos dados coletados, o que é um importante diferencial considerando o público-alvo do CONO.

5. Conclusão

Este artigo apresentou o CONO como uma ferramenta inovadora para o Ministério Público de Santa Catarina (MP-SC), facilitando o monitoramento de informações relevantes. Contudo, a baixa taxa de efetividade aponta para a necessidade de aprimoramento nos mecanismos de busca. O sistema é altamente adaptável, podendo ser expandido para outros estados e áreas, como o combate à lavagem de dinheiro. Melhorias futuras podem incluir o uso de técnicas avançadas de aprendizado profundo para aprimorar a precisão e a automação da análise. O CONO se destaca como um modelo promissor, com potencial para contínuo aprimoramento e novas aplicações.

Referências

Chakrabarti, S. (2009). *Focused Web Crawling*, pages 1147–1155. Springer US, Boston, MA.

de Paula, T. S. (2022). Classificação de notícias de fraude e corrupção em português para instauração de processo investigativo. Master's thesis, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rio de Janeiro.

Moraes, M. P. (2019). Mineração de dados aplicada à identificação de notícias falsas.

Reips, L., Musicante, M., Vargas-Solar, G., Pozo, A., and Hara, C. (2023). Enow - extrator de dados de notícias da web. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 78–83, Porto Alegre, RS, Brasil. SBC.