DATALUTA – Banco de Dados da Luta pela Terra: Automatização do Processo de Dados

Autores: SILVA, T. F. G.[1]; MARTINEZ, F. H. V.[2]

[1]Grupo PET-Computação, UFMS, Campus Cidade Universitária; [2]Tutor(a) do Grupo PET-Computação, UFMS, Campus Cidade Universitária
E-mail: thiago_fernandes@ufms.br, pet-comp.facom@ufms.br

RESUMO: O presente trabalho tem como objetivo automatizar o processo de coleta, extração e tratamento de informações do projeto DATALUTA – Banco de Dados da Luta pela Terra, criado pela UNESP. O estudo propõe o desenvolvimento de um fluxo automatizado capaz de identificar, extrair e organizar notícias relacionadas à luta pela terra, otimizando a geração de dados que dependem de leituras e análises manuais. Para a extração de informações estruturadas, empregou-se a técnica de *web scraping* por meio da biblioteca **BeautifulSoup**, enquanto a análise de conteúdo e identificação de entidades nomeadas foram realizadas por meio de técnicas de Processamento de Linguagem Natural (PLN), utilizando a biblioteca **SpaCy**. Posteriormente, as informações foram tratadas e estruturadas com o auxílio das bibliotecas **NumPy** e **Pandas**, resultando na criação de um *dataset* destinado ao treinamento de modelos de Inteligência Artificial. A proposta contribui para aprimorar a eficiência e a precisão da coleta de dados do DATALUTA de forma automatizada, fortalecendo a produção de conhecimento sobre os movimentos socioterritoriais no Brasil.

Palavras-chave: Mineração de dados; Movimentos socioterritoriais; Rede DATALUTA; Inteligência Artificial.

DATALUTA – Database of the Struggle for Land: Automation of Data Processing

ABSTRACT: This study aims to automate the process of collecting, extracting, and processing information from the DATALUTA project - Database on the Struggle for Land, created by UNESP. The research proposes the development of an automated workflow capable of identifying, extracting, and organizing news content related to the struggle for land, optimizing data generation that currently depends on manual reading and analysis. For extracting structured information, the *web scraping* technique was employed using the **BeautifulSoup** library, while content analysis and named entity recognition were performed through Natural Language Processing (NLP) techniques with the **SpaCy** library. Subsequently, the information was processed and structured using the **NumPy** and **Pandas** libraries, resulting in a dataset intended for training Artificial Intelligence models. This approach contributes to enhancing the efficiency and accuracy of DATALUTA's data collection in an automated manner, strengthening knowledge production on socioterritorial movements in Brazil.

Keywords: Data mining; Socioterritorial movements; Rede DATALUTA; Artificial intelligence

1 INTRODUÇÃO

O DATALUTA - Banco de Dados da Luta pela Terra, foi criado em 1998 pelo Núcleo de Estudos, Pesquisas e Projetos de Reforma Agrária (NERA), da Universidade Estadual Paulista (UNESP), com o objetivo de reunir, organizar e relatar dados da luta pela terra e da reforma agrária no Brasil. Em 2005, foi estabelecida a Rede DATALUTA com a incorporação de novos grupos de pesquisa. Hoje mais de 23 grupos instituídos no Brasil, além de países nas Américas e no Reino Unido fazem parte dessa rede (Fernandes, B.; Pereira L, 2025).

O projeto trabalha com diversas frentes como o DATALUTA Agrário, Águas, Floresta, e Urbano (Filho, J.; Perez, P.; Torres, P., 2025). Atualmente a Universidade de Brasília tem o objetivo de documentar e estruturar notícias de diversas fontes sobre o trabalho do DATALUTA Floresta e o processo é realizado por meio das seguintes etapas:

1. Receber notícias sobre a luta pela terra com o Google Alerts;
2. Ler, compreender e analisar cada notícia;
3. Preencher um formulário da plataforma JotForm sobre a notícia analisada;
4. Alocar as informações da notícia em uma Planilha Google.

Entretanto, essa atividade é realizada manualmente e revela-se muito exaustiva e inviável, pois as notícias obtidas na etapa 1 atingem centenas de ocorrências. O presente trabalho visa automatizar esse processo.

1.1 OBJETIVO GERAL

Este trabalho tem como objetivo geral relatar e registrar as atividades envolvidas na programação da automatização do preenchimento do formulário para registro de notícias relacionadas à luta pela terra.

1.2 OBJETIVO ESPECÍFICO

O trabalho tem por objetivos específicos:

1. Explorar a maneira mais eficiente de mineração de dados em portais de notícias;
2. Desenvolver um programa que automatize parte do processo de preenchimento de formulários com campos específicos;
3. Estudar técnicas de Inteligência Artificial (IA) para identificar ações características em notícias.

2 METODOLOGIA

Inicialmente, para a coleta de dados das notícias, foi idealizada realizar uma raspagem de dados (*web scraping*), portanto, notou-se necessário utilizar a biblioteca da linguagem Python chamada **BeatifulSoup**, que permite analisar documentos em HTML e XML, sendo conveniente para o trabalho. Com ela sucedeu a possibilidade de extrair dados mais simples, como a data, título e fonte da notícia.

Posteriormente, tornou-se essencial trabalhar com Processamento de Linguagem Natural (PLN), principalmente com Reconhecimento de Entidades Nomeadas (NER). Para isso, integramos a tecnologia do **SpaCy**, uma biblioteca de código aberto, que possibilitou capturar dados mais complexos, como o município onde ocorreu o evento relatado na notícia. A partir dessa informação, tornou-se viável acessar dados complementares, como a macrorregião, o estado e o código IBGE correspondentes. Foi possível, por meio das bibliotecas **Pandas** e **NumPy**, desenvolver um *dataset* destinado ao armazenamento e à organização das informações das ações matrizes e derivadas.

2.1 WEB SCRAPING

A técnica de *web scraping* foi idealizada para essa atividade pela sua eficácia e agilidade na extração e no processamento de informações provenientes no HTML das páginas *web* (Khder M. 2021). Optou-se pela utilização da biblioteca **BeautifulSoup** pois não há necessidade de uma grande quantidade de extrações. Essa biblioteca se destaca pela sua praticidade, eficiência e desempenho na extração de dados em ambientes de menor complexidade (DikilitaŞ *et al.,* 2021).

Nesta etapa, o foco é a extração de metadados e informações explícitas, como o título da notícia, a fonte (portal de origem) e a data de publicação. Essas informações, por estarem geralmente presentes em *tags* padronizadas no código HTML, permitem uma execução de extração ampla e eficiente por meio da técnica de *web scraping*.

Tal abordagem permite a obtenção de informações de forma eficiente e sistemática, garantindo assim uma maior consistência na padronização dos dados, minimizando erros associados à coleta manual.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Para a extração de informações mais complexas e contextuais, que necessitam de análise textual, ou seja, não são acessíveis via análise da estrutura HTML, foi utilizado PLN, que estuda

como o computador analisa, reconhece ou gera textos em linguagens humanas (Vieira, R.; Lopes, L. 2010).

Dentre as técnicas em PLN, destaca-se o Reconhecimento de Entidades Nomeadas (*Named Entity Recognition – NER*), aplicado para identificar e classificar automaticamente entidades relevantes no texto (Sharnagat, R. 2014).

A tecnologia selecionada para esta tarefa foi a biblioteca de código aberto **SpaCy**, escolhida por sua capacidade de lidar com um grande volume de textos e por oferecer suporte em múltiplos idiomas, incluindo o português brasileiro. A biblioteca realiza análises linguísticas avançadas para a identificação de entidades de interesse no corpo textual e disponibiliza um conjunto abrangente de operações de PLN, como tokenizador, analisador e etiquetador (KUMAR et al., 2023).

O principal objetivo da aplicação do NER nesta fase é a identificação de localidades geográficas mencionadas nas notícias, especificamente o município onde o evento ocorreu. Uma vez que o município é identificado e extraído do texto, o sistema procede com o enriquecimento desses dados. São realizadas consultas a bases de dados complementares para obter informações associadas, como a macrorregião, o estado e o código IBGE correspondente à localidade identificada.

2.3 TREINAMENTO DE MODELO

Por fim, utilizando-se das bibliotecas **NumPy** e **Pandas**, foi possível realizar a filtragem das planilhas do DATALUTA armazenadas nas planilhas Google, a fim de identificar e separar as ocorrências de ações matrizes e ações derivadas. A partir desse processo, foi possível construir um *dataset* estruturado, destinado ao treinamento de um modelo de Inteligência Artificial.

3 RESULTADOS E DISCUSSÃO

A aplicação conjunta das tecnologias citadas permitiu a extração automatizada de um conjunto de dez informações principais: título, fonte, data, macrorregião, estado, município, código IBGE, bioma, ação derivada e ação matriz. É importante ressaltar que os campos derivados da análise de PLN, como a localização e as ações, ainda não alcançaram 100% de precisão, indicando a necessidade de refinamento contínuo dos modelos e do código.

Como trabalhos futuros, a metodologia prevê a expansão da capacidade de extração para campos de maior complexidade semântica. Isso inclui a identificação de tipos de movimentos

sociais (como movimentos indígenas), palavras-chave e a classificação das notícias de acordo com os Objetivos de Desenvolvimento Sustentável (ODS) da ONU. Este avanço exigirá o estudo e a aplicação de técnicas mais sofisticadas de Inteligência Artificial para a análise e classificação de textos, assim ampliando o potencial analítico e a contribuição científica do trabalho.

4 CONCLUSÕES

Este trabalho demonstra que a automatização da coleta e registro de notícias sobre a luta pela terra é viável, reduzindo o esforço manual e permitindo a análise de grandes volumes de dados. Além disso, é evidenciado pelo projeto o potencial das tecnologias aplicadas para a automatização e construção de uma base de dados direcionada para a compreensão dos movimentos socioterritoriais e conflitos agrários no Brasil. Embora algumas informações mais complexas ainda apresentem taxas de erro significativas, o uso de técnicas de raspagem de dados e PLN mostra-se promissor. Futuras melhorias poderão aumentar a precisão e ampliar os campos de informação, tornando o DATALUTA uma ferramenta mais eficiente para pesquisadores da área e para o fortalecimento de iniciativas de pesquisa sobre os movimentos socioterritoriais e o uso da terra.

AGRADECIMENTOS

Este trabalho contou com o apoio do Fundo Nacional de Desenvolvimento da Educação (FNDE), cujo incentivo foi essencial para o desenvolvimento desta pesquisa. Agradecemos também ao professor Marco Aurélio Stefanes, professor da Faculdade de Computação da UFMS, pela orientação, contribuições teóricas e acompanhamento ao longo do processo, que foram fundamentais para o crescimento e consolidação dos resultados apresentados. O apoio conjunto do FNDE e do professor Marco reforça a importância da integração entre pesquisa, educação e inovação para o avanço científico e social.

REFERÊNCIAS

DIKILITAŞ, Yılmaz et al. Performance analysis for web scraping tools: case studies on beautifulsoup, scrapy, htmlunit and jsoup. In: International Conference on Emerging Trends and Applications in Artificial Intelligence. Cham: Springer Nature Switzerland, 2023. p. 471-480. Disponível em: https://www.researchgate.net/publication/380189817_Performance_Analysis_for_Web_Scraping_Tools_Case_Studies_on_Beautifulsoup_Scrapy_Htmlunit_and_Jsoup. Acesso em: 2 de out. 2025

KHDER, Moaiad Ahmad. Web scraping or web crawling: State of art, techniques, approaches and application. International Journal of Advances in Soft Computing & Its Applications, v. 13, n. 3, 2021. Disponível em: https://www.i csrs.org/Volumes/ijasca/2021.3.11.pdf. Acesso em: 04 out. 2025

KUMAR, Murari et al. An algorithm for automatic text annotation for named entity recognition using Spacy framework. ICAR, Delhi, India, Tech. Rep, 2023. Disponível em: https://doi.org/10.21203/rs.3.rs-2930333/v1. Acesso em: 4 de out. 2025.

MANÇANO FERNANDES, B.; PEREIRA, L. I. Movimientos socioterritoriales y acaparamiento del territorio en Brasil. Punto Sur, n. 12, p. 2-20, 6 jun. 2025. Disponível em: http://revistascientificas2.filo.uba.ar/index.php/RPS/article/view/17203. Acesso em: 4 out. 2025.

SHARNAGAT, Rahul. Named entity recognition: A literature survey. Center For Indian Language Technology, v. 1, p. 1, 2014. Disponível em: https://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf. Acesso em: 4 out. 2025.

SOBREIRO FILHO, J.; CEPERO RUA PEREZ, P.; VITOR LUNA TORRES, P. Entre as sementes, as raízes e os frutos da pesquisa, a natureza territorial das lutas das florestas: aportes teórico-metodológicos do DATALUTA Floresta . Punto Sur, n. 12, p. 175-196, 5 jun. 2025. Disponível em: https://doi.org/10.34096/ps.n12.14530. Acesso em: 02 out. 2025

VIEIRA, Renata; LOPES, Lucelene. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. Em corpora, p. 183, 2010. Acesso em: 03 de out. 2025.