Construção de um banco de dados de espécies arbóreas nativas brasileiras, suas relações ecológicas com a fauna e a flora e importância econômica, através de mineração de texto

Palavras-Chave: Revisão Sistemática, Mineração de Texto, Automação

Autores:

Gustavo Henrique Lima Medeiros de Sousa (Autor) - UNICAMP

Lucas Augusto Cursino (Coautor) - BioForests

Igor Fernando Mandello (Coautor) - BioForests/UNICAMP

Marcio Luiz Magrini (Coautor) - BioForests/UNICAMP

Prof. Dr. Jurandir Zullo Junior (orientador) - UNICAMP - BI0S/FAPESP - CNPq

INTRODUÇÃO:

A conservação da biodiversidade e a promoção do uso sustentável dos recursos naturais tornaram-se temas centrais nos debates ambientais, científicos e econômicos, refletindo a crescente preocupação com o desenvolvimento sustentável. Nesse contexto, as espécies arbóreas nativas dos mais diversos biomas brasileiros desempenham papel fundamental na manutenção dos ecossistemas e na prestação de serviços ambientais essenciais, como a captura de carbono, o equilíbrio hidrológico e a sustentação de comunidades da fauna e flora. Apesar de sua relevância ecológica, econômica e científica, as informações disponíveis sobre as espécies arbóreas nativas da Mata Atlântica encontram-se dispersas em diferentes bases de dados, dificultando a realização de uma análise sistemática, integrada e abrangente. Essa fragmentação do conhecimento limita o avanço de estratégias eficazes de conservação, uso sustentável dessas espécies e restauração ecológica, especialmente em um contexto de severa perda de cobertura vegetal, como

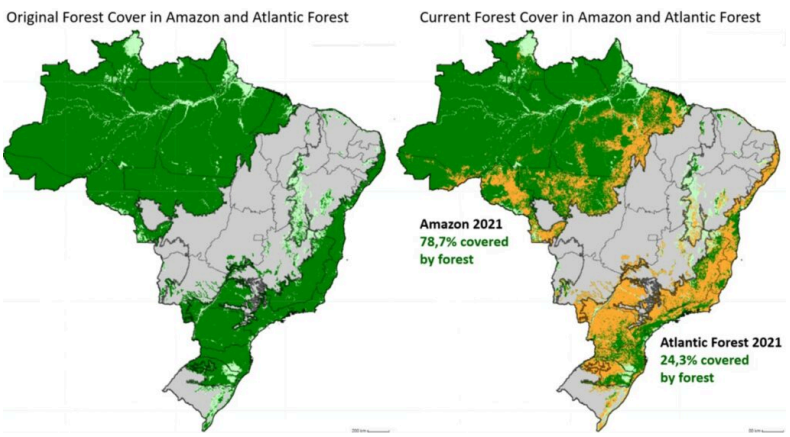

Figura 1. Mapa comparativo do desmatamento da Mata Atlântica FONTE: SOS MATA ATLÂNTICA

ilustrado na Figura 1, que compara a área originalmente florestada com a atual, reforçando a urgência de ações que subsidiem sua preservação e recuperação.

Diante disso, o presente projeto, realizado junto à greentech **Bioforests**[1], empresa sediada no prédio núcleo do Parque Científico e Tecnológico da UNICAMP, tem como objetivo o design e a construção de um banco de dados voltado à coleta e organização de informações sobre espécies arbóreas nativas da Mata Atlântica, suas interações ecológicas com outros organismos e seus impactos socioeconômicos em contextos locais e regionais. Para garantir eficiência e abrangência na obtenção dos dados, adotou-se o uso de técnicas de mineração de texto ("text mining"), capazes de automatizar a extração de conhecimento a partir de grandes volumes de literatura científica.

METODOLOGIA:

A construção do banco de dados e o processo de mineração de texto foram realizados a partir da combinação de fontes confiáveis[2][3][4] e diversas ferramentas computacionais[5][6][7][8][9]. Como base, utilizou-se o GHG Protocol Florestas e SAFs[10][11] e a extração textual de livros presentes na coleção Espécies Arbóreas Brasileiras[12], da EMBRAPA Florestas, visando garantir confiabilidade e riqueza de informações sobre aspectos ecológicos e usos econômicos das espécies, essenciais para análises mais amplas de suas interações com fauna, flora e cadeias produtivas.

Para o pré-processamento e análise dos dados, foi utilizado um conjunto de ferramentas em **Python**[5], incluindo as bibliotecas **NLTK**[9], para o processamento de linguagem natural, **Pandas**[6] e **NumPy**[8], para manipulação de dados. Além disso, em conjunto com a biblioteca de linguagem natural NLTK, foi utilizado auxílio de **Expressões Regulares (RegEx)**[5], ferramenta contida dentro da própria linguagem Python, para extração de padrões textuais específicos presente na coleção de livros, visto que todos os exemplares possuem uma estrutura semelhante e padronizada, facilitando a análise bibliográfica. Ainda, o uso do ambiente **Jupyter Notebook**[13] facilitou a organização dos scripts e análises exploratórias. Adicionalmente, ferramentas como *Matplotlib, Seaborn e Scikit-learn* foram utilizadas para visualização gráfica e interpretação de dados, para uma análise mais elaborada e uma melhor compreensão das dificuldades.

Como finalização dos métodos utilizados, a ferramenta de banco de dados utilizada foi o **PostgreSQL**[7] que, juntamente com a linguagem *SQL,* permitiu a manipulação de dados dentro do próprio banco, ajudando a modelagem de estruturas dentro do banco.

RESULTADOS E DISCUSSÃO:

Inicialmente, foram utilizados os dados do GHG Protocol Florestas e SAFs, que trata-se de uma abordagem metodológica voltada para a quantificação e gestão das emissões de gases de efeito estufa (GEE) em atividades vinculadas a florestas e sistemas agroflorestais, e que oferece diretrizes para mensurar tanto as capturas quanto às emissões de carbono, incentivando práticas sustentáveis com potencial econômico. O protocolo também disponibiliza informações estruturadas sobre 238 espécies arbóreas, incluindo densidade da madeira, incremento anual e participação da biomassa,

permitindo o cálculo de estimativas de carbono (C) e gás carbônico (CO_2) por hectare em diferentes partes da árvore (raiz, parte aérea e serrapilheira). Paralelamente, foi iniciada uma revisão bibliográfica sobre conceitos de aprendizado de máquina, com o intuito de embasar uma futura aplicação de modelos de mineração de texto sobre a literatura científica especializada.

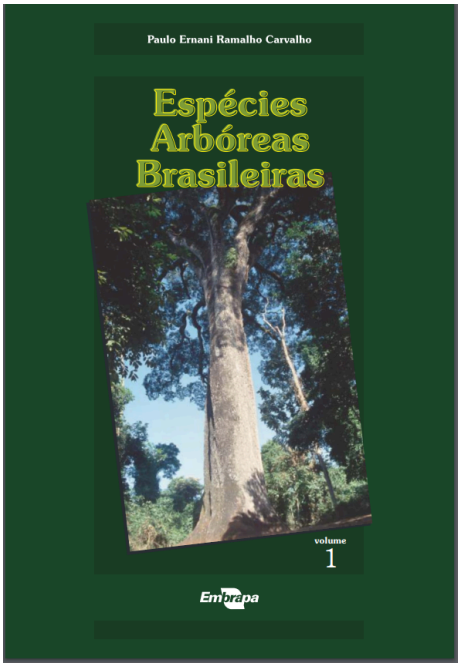

Figura 2. Primeiro livro da coleção *Espécies Arbóreas Brasileiras*

Neste sentido, a extração de dados textuais da coleção *Espécies Arbóreas Brasileiras* (figura 2) é de grande importância pois fornece informações confiáveis, padronizadas e cientificamente consolidadas sobre espécies nativas, incluindo aspectos ecológicos e usos econômicos. A utilização dessa fonte qualificada permite enriquecer significativamente o banco de dados do projeto, garantindo maior precisão nas análises e contribuindo para o entendimento das interações ecológicas e dos impactos socioeconômicos dessas espécies, especialmente no contexto da agricultura sustentável voltada a pequenos e médios produtores rurais.

As informações que foram escolhidas para serem retiradas da coleção se baseiam em características biológicas de cada espécie, como nome popular, nome científico, densidade da madeira e ritmo de crescimento; interação entre fauna e flora, como dispersão de frutos/sementes e grupos sucessionais; e também informações de interesse econômico, como produção de celulose, papel e produção de madeira. Além disso, é possível destacar características voltadas ao reflorestamento que cada espécie possui, dado fundamental para projetos futuros que usarão este banco de dados como base. Para que fosse possível retirar essas informações, utilizou-se um script em Python (Jupyter Notebook) com a função de extrair todo o texto presente no livros, utilizando principalmente *Expressões Regulares* com auxílio de *NLTK* (apenas na parte onde continham as espécies). Esse texto foi então transformado em *DataFrames,* utilizando a biblioteca *Pandas,* que foram exportadas como arquivos .csv, para facilitar o transporte de todos os dados para o banco de dados no software *PostgreSQL.*

Todos os testes iniciais e primeiras funções foram construídos com base no primeiro exemplar da coleção da Embrapa Florestas, onde foi possível extrair todo o conteúdo do primeiro livro e através dos testes aplicados posteriormente demonstra-se que esse mesmo script pode ser aplicado nos demais volumes da coleção. Permitindo portanto que todo o conteúdo proposto no objetivo do projeto seja extraído, focando principalmente no impacto econômico das espécies nativas da Mata Atlântica.

CONCLUSÕES:

O desenvolvimento deste projeto permitiu a criação estruturada e funcional de um banco de dados robusto sobre espécies arbóreas nativas da Mata Atlântica, integrando informações ecológicas, econômicas e botânicas provenientes de fontes confiáveis. A utilização de técnicas de mineração de

texto, aliadas ao uso de ferramentas computacionais avançadas em Python e ao sistema de gerenciamento de banco de dados PostgreSQL, viabilizou a extração, organização e armazenamento de dados relevantes de forma automatizada, padronizada e escalável.

Esse banco de dados representa uma importante contribuição para estudos voltados à conservação da biodiversidade, ao uso sustentável dos recursos florestais e à restauração ecológica. Ele fornece uma base para análises ambientais e econômicas mais precisas, que poderão subsidiar projetos de reflorestamento e iniciativas voltadas à agricultura de base florestal, especialmente em apoio a pequenos e médios produtores rurais.

BIBLIOGRAFIA

[1] **Greentech Bioforests**. Parque Científico Tecnológico da UNICAMP. Agência de Inovação da UNICAMP. Disponível em: <https://parque.inova.unicamp.br/portfolio/bioforests/> Acesso em: 02 ago. 2025.

[2] Metzger, J. P., & Casatti, L.. (2006). **Do diagnóstico à conservação da biodiversidade: o estado da arte do programa BIOTA/FAPESP**. Biota Neotropica, 6(2). doi: 10.1590/S1676-06032006000200002

[3] Antons, David & Grünwald, Eduard & Cichy, Patrick & Salge, Oliver. (2020). **The application of text mining methods in innovation research: current state, evolution patterns, and development priorities.** R&D Management. doi: 50. 10.1111/radm.12408.

[4] Marsi, Erwin & Ozturk, Pinar. (2016). **Text Mining of Related Events from Natural Science Literature.** doi: 10.1007/978-3-319-53637-8_4.

[5] Python 3 Documentation. **Python.** Disponível em <https://docs.python.org/3/>. Acesso em: 01 ago. 2025.

[6] PANDAS. **pandas documentation — pandas 1.0.1 documentation**. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: 01 ago. 2025.

[7] THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostgreSQL: The world's most advanced open source database**. Disponível em: <https://www.postgresql.org/>. Acesso em: 01 ago. 2025.

[8] NUMPY. **NumPy**. Disponível em: <https://numpy.org/>. Acesso em: 01 ago. 2025.

[9] Bird, Steven, Edward Loper and Ewan Klein (2009). **Natural Language Processing with Python**. O'Reilly Media Inc.

[10] **GHG Protocolo Florestas e SAFs**. Disponível em: <https://www.wribrasil.org.br/publicacoes/ferramenta-de-calculo-para-balanco-de-emissoes-de-gases-de-efeito-estufa-em-florestas-e>. Acesso em: 01 ago. 2025.

[11] DELGADO ASSAD, E. et al. **GREENHOUSE GAS PROTOCOL CALCULATION TOOL FOR FORESTRY IN BRAZIL.** [s.l: s.n.]. Disponível em: <https://www.wribrasil.org.br/sites/default/files/ghg_protocolo-florestas-technicalnote.pdf>. Acesso em: 01 ago. 2025.

[12] CARVALHO, P. E. R. **Espécies arbóreas brasileiras. Brasília, DF: Embrapa Informação Tecnológica;** Colombo: Embrapa Florestas, 2006. Disponível em: <https://www.embrapa.br/florestas/publicacoes/especies-arboreas-brasileiras>. Acesso em: 01 ago. 2025.

[13] JUPYTER NOTEBOOK. **Jupyter Notebook**. Disponível em: <https://docs.jupyter.org/en/latest/>. Acesso em: 02 ago. 2025