

PESQUISA/INICIAÇÃO CIENTÍFICA - PESQUISA/ INICIAÇÃO CIENTÍFICA

O PROCESSAMENTO DE LINGUAGEM NATURAL (NLP) PARA AUXILIAR A ELABORAÇÃO DE UM DICIONÁRIO ELETRÔNICO

Guilherme Ferreira Mendes Vieira
(guilherme.vieira484@academico.ufgd.edu.br)

Bruno Oliveira Maroneze (brunomaroneze@ufgd.edu.br)

O Dicionário Histórico de Termos da Biologia é um dicionário eletrônico de acesso gratuito (disponível em <https://dicbio.fflch.usp.br/>) que tem como objetivo descrever a etimologia, a história e a evolução semântica de termos da Biologia identificados em um conjunto de textos (cópus) constituído por obras científicas produzidas no século XVIII. Esses textos, de relevância histórica e científica, incluem tratados de Botânica, Zoologia e Medicina, representando um período fundamental para a consolidação da linguagem científica em língua portuguesa. O projeto se insere na área das Humanidades Digitais, empregando recursos computacionais para a exploração automatizada de textos digitais, o armazenamento estruturado de dados linguísticos e a exibição de resultados em formato web. A arquitetura técnica envolve um fluxo integrado de coleta, processamento e disponibilização dos dados: os textos são digitalizados, processados e analisados, e os termos extraídos são apresentados ao público com informações linguísticas e históricas detalhadas. A pesquisa atual busca ampliar e otimizar esse processo, elaborando algoritmos mais eficientes que melhorem o código existente, permitam atualizações contínuas do banco de dados e viabilizem a extração automatizada de informações. Os objetivos específicos incluem: (1) atualizar e

documentar o código-fonte do site; (2) automatizar a identificação e etiquetagem de termos no corpus, acrescentando informações como classe gramatical, lematização e segmentação em sentenças; (3) aumentar a eficiência no processamento e integração de novos textos e termos. O sistema, originalmente escrito em R, foi migrado para Python com o framework Django, utilizando CSS e Bootstrap para a interface. A mudança possibilitou maior modularidade, reaproveitamento de ferramentas e colaboração entre desenvolvedores. Foram incorporadas bibliotecas como Pandas (manipulação de dados tabulares), spaCy (processamento linguístico) e lxml (manipulação de XML). O processo metodológico inclui a digitalização de textos históricos com o uso da plataforma Transkribus, que aplica reconhecimento automático por IA, seguida da conversão para XML com marcação manual de elementos relevantes. Adotou-se o padrão TEI (Text Encoding Initiative), o que garante interoperabilidade e precisão na representação dos dados textuais. Desenvolveu-se ainda um microsserviço para automação de marcação de termos, com extração de metadados e geração de arquivos CSV, conversão para HTML e armazenamento em banco de dados. Entre os resultados alcançados estão: (a) um novo site funcional, capaz de receber atualizações frequentes de livros e termos; (b) automações que aceleram o processamento e reduzem erros humanos; (c) maior acessibilidade e interatividade na consulta, com integração de recursos de busca e visualização. Conclui-se que as mudanças implementadas proporcionam escalabilidade, eficiência e dinamismo à plataforma, fortalecendo o dicionário como ferramenta de referência para pesquisas em Etimologia, História da Ciência, Biologia e Humanidades Digitais. AGRADECIMENTO: À FUNDECT pelo apoio financeiro essencial para a realização desta pesquisa.

Palavras-chave: etimologia; processamento de texto; humanidades digitais.