UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2024

NOVOS AVANÇOS EM FERRAMENTAS COMPUTACIONAIS PARA INVESTIGAÇÕES EM HUMANIDADES DIGITAIS

Thiago Sena[1]; Angelo Loula[2]
1. Bolsista PROBIC/UEFS, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: thiagopinto.sena@gmail.com
2. Orientador, Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, e-mail: angelocl@uefs.br

PALAVRAS-CHAVE: humanidades digitais; colaboração científica; escansão computacional

INTRODUÇÃO

Humanidades Digitais (HD) referem-se a novas abordagens acadêmicas e institucionais que utilizam tecnologias computacionais para pesquisa, ensino e publicações colaborativas, englobando práticas convergentes que ampliam os conceitos tradicionais de conhecimento nas artes, humanidades e ciências sociais. As HD exploram as oportunidades e desafios que emergem da integração dos aspectos digitais com as humanidades, criando um novo campo de estudo (Burdick et al., 2012). Neste contexto, surgem ferramentas computacionais desenvolvidas para facilitar e otimizar o processamento e análise de dados da área das humanidades.

O currículo Lattes é fundamental para pesquisadores brasileiros, reunindo informações sobre suas produções acadêmicas e profissionais. Para atender a essa demanda, ferramentas como ScriptLattes (Mena-Chalco & Junior, 2009) e o LucyLattes (Tieppo, 2021) foram desenvolvidas para automatizar a extração e análise dessas informações. Enquanto o ScriptLattes tornou-se obsoleto, o LucyLattes trouxe melhorias, mas ainda apresentava limitações na detecção de redes de coautoria, problemas que foram abordados no projeto anterior (Sena, 2023). Com a necessidade de processar um grande volume de currículos, este projeto visa criar uma ferramenta para indexação e filtragem eficiente desses dados.

O projeto focou também na avanços em uma ferramenta computacional para a descoberta de padrões métricos em textos literários, com mais aprimoramentos o software MIVES (Mining Verse Structures) (Carvalho, 2017), desenvolvido para escansão de estruturas métricas em prosa de língua portuguesa. Embora ajustes já tenham sido feitos no projeto anterior, estudos preliminares indicaram a necessidade de melhorias adicionais. Neste projeto, foram realizadas, especialmente, correções e melhorias na visualização e exportação dos resultados da ferramenta MIVES (*output* do sistema).

METODOLOGIA

Para o desenvolvimento desta pesquisa, foram seguidas as etapas metodológicas em ordem cronológica: (1) Avaliação de demandas de aprimoramentos finais a serem feitos da ferramenta MIVES; (2) Ajustes e correções da ferramenta MIVES; (3) Análise prévia dos currículos de pesquisadores; (4) Início do desenvolvimento da ferramenta de filtragem e indexação de currículos; (5) *Profiling* e otimização da ferramenta.

Nas etapas 1 e 2, foram realizadas avaliações da ferramenta MIVES por meio da leitura e testagem de seus códigos, resultando em ajustes finais, com foco na interface de exportação e visualização dos dados. Em etapa subsequente, atividades foram dedicadas ao desenvolvimento de uma ferramenta de filtragem e indexação para um grande volume de currículos Lattes. O processo iniciou com um estudo preliminar dos possíveis filtros a serem aplicados, seguido pela análise estrutural dos currículos. Durante o desenvolvimento da ferramenta, foram identificados gargalos de desempenho, levando à necessidade de realizar *profiling* e otimizar o código.

RESULTADOS E/OU DISCUSSÃO

Na fase inicial do projeto, o foco principal foi a visualização e exportação dos resultados gerados pela ferramenta MIVES (*output* do sistema), o que resultou em diversas modificações para aprimorar a experiência do usuário. Um dos problemas corrigidos foi na tela final, onde os resultados não eram exibidos automaticamente após o processamento dos textos. Agora, ao finalizar a mineração, o usuário pode visualizar os resultados imediatamente, sem a necessidade de interações adicionais na última tela.

Além disso, ajustes significativos foram feitos na aba de estatísticas, que agora exibe corretamente informações relevantes, como o nome do arquivo processado, número de frases e estruturas identificadas, metros buscados e o tipo de busca realizada. Também foram corrigidos problemas no menu de importação e exportação de resultados, eliminando ações redundantes e resolvendo erros que afetavam a funcionalidade, tornando a interface mais clara e eficaz. A tela de resultados da ferramenta MIVES pode ser vista na Figura 1.

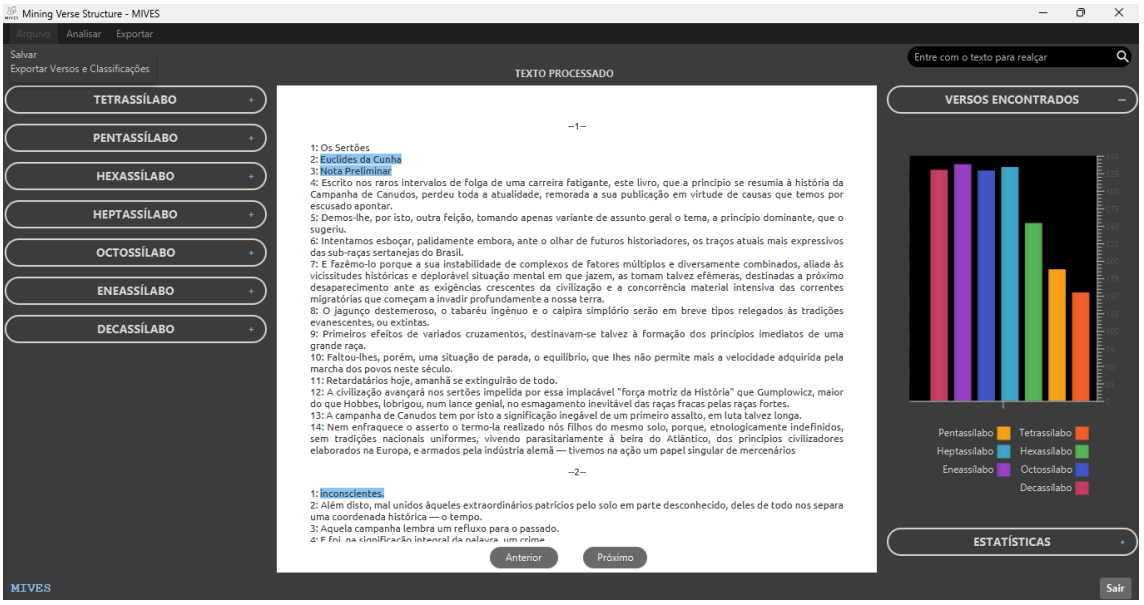

Figura 1: Tela de Resultados do MIVES ajustada

Além da análise e ajuste da ferramenta MIVES, este trabalho iniciou o desenvolvimento de uma ferramenta de indexação de um grande volume de currículos Lattes. O primeiro passo no desenvolvimento da nova ferramenta foi investigar como a Plataforma Lattes realiza a filtragem na busca de currículos dos pesquisadores. Foi necessário analisar também a estruturação dos arquivos XML que contêm os dados dos pesquisadores para identificar os filtros que poderiam ser implementados. A filtragem pode ser feita por critérios como titulação, nacionalidade, idiomas, atuação profissional, formação acadêmica, e dados sobre localização e natureza da atividade profissional.

O segundo passo envolveu a criação de um *script* em *Python* para processar os currículos em arquivos XML compactados em formato ZIP, extraindo informações e as salvando em um arquivo CSV. Esse processo revelou problemas relacionados ao consumo elevado de memória RAM devido ao tamanho dos dados. Inicialmente, o *script* enfrentou desafios significativos de memória, o que levou a interrupções no seu funcionamento.

Após identificar que o consumo de memória era causado por uma variável que aumentava com a leitura dos currículos, optou-se por reduzir o uso de memória e aumentar a frequência das operações de I/O, o que, no entanto, causou um gargalo de desempenho. Para melhorar a eficiência, foi feito um *profiling* do código, revelando que a maior parte do tempo era gasta na inicialização de objetos ZipFile (usado para leitura de arquivos ZIP) e BeautifulSoup (usado para leitura de arquivos XML) (Figura 2).

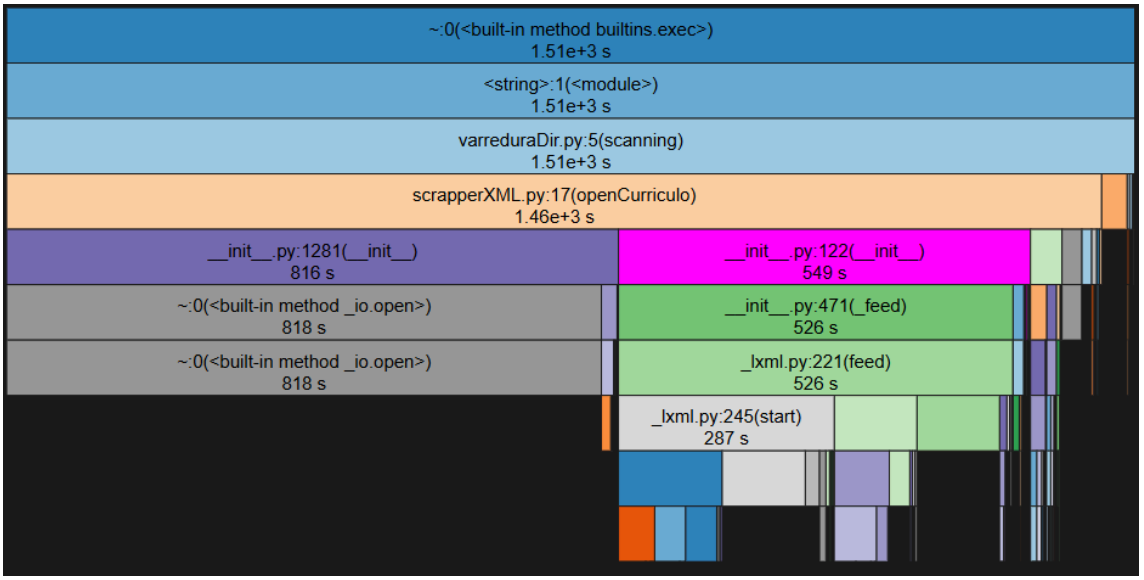

Figura 2: Gráfico de chamadas de funções do script original. Em rosa, é indicado o tempo acumulado da inicialização de objetos BeautifulSoup, e em lilás, inicialização de objetos ZipFile.

A substituição da BeautifulSoup pela biblioteca padrão ElementTree para *parsing* de XML resultou em uma redução de aproximadamente 40% no tempo total de execução do *script*. Apesar dessa melhoria, a inicialização dos objetos ZipFile ainda consumia tempo significativo, e ainda não foram encontradas alternativas melhores para a leitura de arquivos compactados.

Para otimizar ainda mais, testou-se a execução paralela de processos. Processando duas partições de dados simultaneamente, o tempo de execução foi reduzido a pouco mais da metade do tempo necessário para processar cada partição individualmente, indicando que um paralelismo de processos pode ajudar na otimização. A ferramenta de filtragem e indexação de currículos ainda está em fase de otimização de desempenho, e o trabalho continuará no próximo período de iniciação científica.

CONSIDERAÇÕES FINAIS

As Humanidades Digitais têm desempenhado um papel fundamental ao combinar métodos das ciências humanas com ferramentas computacionais, oferecendo novas abordagens para a análise e visualização de informações em diversas áreas do conhecimento. Nesse contexto, ferramentas como MIVES e os sistemas de extração de dados da Plataforma Lattes são essenciais para a evolução dos estudos sobre colaboração científica e padrões literários.

A ferramenta de filtragem e indexação de currículos, ainda em fase de desenvolvimento, será aprimorada no próximo trabalho de Iniciação Científica, garantindo a continuidade dos avanços no processamento de grandes volumes de dados acadêmicos.

REFERÊNCIAS

BURDICK, Anne; DRUCKER, Johanna; LUNENFELD, Peter; PRESNER, Todd;

SCHNAPP, Jeffrey. Digital Humanities. Cambridge: The MIT Press, 2012. ISBN 9780262312103. Disponível em: https://doi.org/10.7551/mitpress/9248.001.0001.
CARVALHO, R. S. (2017) MIVES: um sistema para identificação automática de padrões métricos de versificação em prosa literária brasileira. 120 f. Dissertação (Mestrado em Computação Aplicada)- Universidade Estadual de Feira de Santana, Feira de Santana.
MENA-CHALCO, J. P., & JÚNIOR, R. M. C. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, 15, 31-39.
RAFAEL TIEPPO (2021). LucyLattes script para a extração e compilação de dados do currículo Lattes. Disponível em https://github.com/rafatieppo/lucylattes
SENA; CONRADO, A. AVANÇOS EM FERRAMENTAS COMPUTACIONAIS PARA INVESTIGAÇÕES EM HUMANIDADES DIGITAIS. Anais do ... Seminário de Iniciação Científica/Anais Seminário de Iniciação Científica, n. 27, 18 jun. 2024.