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

XXVII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2023

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

As áreas das ciências humanas estão incorporando cada vez mais ferramentas computacionais para facilitar a pesquisa, análise, síntese e visualização de informações. Isso resulta na formação de um campo interdisciplinar, multidisciplinar e transdisciplinar conhecido como Humanidades Digitais (HD) (Loula , 2021).

Um tema de interesse é a identificação e análise das redes de colaboração científica nas áreas das humanidades. Segundo Balancieri et al. (2005), a colaboração proporciona uma fonte de apoio para potencializar a produção científica e seus resultados. Para essa finalidade, foram avaliadas e adequadas ferramentas que auxiliam na identificação de coautoria em artigos científicos, a partir da extração de dados da plataforma Lattes.

A análise computacional de títulos de publicações é um tema complementar de estudo das HD, já que o título desempenha um papel essencial em um artigo, pois é o principal indicador do assunto a ser abordado e uma forma de atrair o leitor (Jamali & Nikzad, 2011). Para esta análise, algumas ferramentas computacionais foram desenvolvidas, capazes de extrair informações de títulos, como quantidade de palavras e caracteres.

Outro foco de investigação nas Humanidades Digitais é a Linguística Computacional, que intersecta estudos de pesquisadores das áreas de Matemática, Computação e Linguística (Tramunt & Batista, 2017). Uma ferramenta para análise computacional de obras literárias, é o MIVES (Carvalho, 2017), um software desenvolvido para escansão computacional de estruturas métricas de versificação em prosa de língua portuguesa. Neste trabalho foi realizada a avaliação, desenvolvimento de melhorias e ajuste de erros da ferramenta MIVES.

METODOLOGIA

Para o desenvolvimento desta pesquisa, foram seguidas as etapas metodológicas em ordem cronológica: (1) Avaliação das ferramentas de extração de dados da Plataforma Lattes; (2) Ajuste das ferramentas de extração de dados da Plataforma Lattes; (3) Desenvolvimento

de algoritmos computacionais de análise de títulos; (4) Avaliação do software MIVES; (5) Aperfeiçoamento e ajuste de erros do MIVES.

Nas etapas 1 e 4, foram feitas avaliações das ferramentas computacionais já existentes por meio de leitura e testagem de seus códigos. Após estas avaliações, as etapas 2 e 5 puderam ser feitas, aperfeiçoando e consertando os erros encontrados dos softwares analisados. Já a etapa 3 se tratou do desenvolvimento de algoritmos na linguagem Python, com auxílio da ferramenta Google Colab.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A primeira ferramenta analisada foi o ScriptLattes (Mena-Chalco & Júnior, 2009), pioneira na extração e sumarização de informações de currículos Lattes. No entanto, o ScriptLattes deixou de receber atualizações desde 2015, quando o CNPq estabeleceu um mecanismo de segurança, captcha, que evita a extração de informação por sites não autorizados (Corrêa et al, 2017). Além disso, o script foi desenvolvido em uma versão mais antiga do Python, com dependências modificadas ou descontinuadas.

Foi analisada, em seguida, a ferramenta LucyLattes (Tieppo, 2021). Realizados os testes de uso desta ferramenta a partir de um conjunto piloto de currículos, foi observada uma limitação no algoritmo de detecção de rede de coautoria, que realizava busca do nome do coautor na relação de nomes de pesquisadores com currículos disponíveis. Esse critério não se mostrou válido, porque erros de digitação ou ausência de um padrão de escrita dos nomes dos autores na Plataforma Lattes poderiam gerar inconsistências (Kang et al, 2019).

Para melhoria da identificação de co-autoria, foi desenvolvido no Lucylattes o método proposto pelo ScriptLattes, que faz a detecção de produções similares entre autores por meio da comparação entre seus títulos (Mena-Chalco & Júnior, 2013). Quando diferentes currículos possuem publicações com títulos similares, então é estabelecida a co-autoria entre pesquisadores destes currículos. Esta similaridade é baseada na distância de Levenshtein (Navarro, 2001), que avalia a quantidade de operações de edição necessárias para igualá-las. Assim, modificamos o algoritmo do LucyLattes para identificar títulos com similaridade mínima de 90%, considerando a distância de Levenshtein, além de considerar uma filtragem por ano da publicação, evitando processamento desnecessário.

Para além do estudo e ajuste do LucyLattes, foi desenvolvido um algoritmo em Python de detecção de *Trend Words* (palavras de tendência). Este algoritmo recebe um conjunto de títulos de publicações e divide cada título em tokens. A tokenização dos títulos é feita com a substituição de caracteres não-alfanuméricos pelo caractere espaço em branco, seguida da separação de tokens que estejam separados por espaços. Com isso, o algoritmo é capaz de identificar palavras e suas frequências em títulos. Outra funcionalidade desenvolvida foi a contagem de caracteres especiais nos títulos. O algoritmo também foi escrito em Python e calcula a ocorrência de caracteres especiais em títulos de publicações.

A última funcionalidade de análise de títulos desenvolvida foi um script em Python que carrega uma lista de títulos e gera um grafo direcionado de palavras. Primeiro é realizada a decodificação dos títulos, devido a marcações HTML. Após esta etapa, foi necessária a detecção automática da língua de cada título, para que as etapas seguintes de lematização e exclusão de *Stop Words* fossem realizadas, caso o título estivesse na língua portuguesa.

Lematização é o processo de obtenção do lema (palavra base), excluindo todas as flexões possíveis de uma palavra.

Além da análise de co-autoria e títulos de publicações, este trabalho também realizou a avaliação do software MIVES. Esta ferramenta, que foi desenvolvida para escansão computacional de estruturas métricas de versificação em prosa de língua portuguesa, apresentava demandas de aperfeiçoamento em sua interface gráfica, bem como ajustes de alguns erros.

A interface do MIVES solicita, inicialmente, que o usuário forneça informações sobre o texto a ser processado e sobre a parametrização do processamento. Entretanto, como essa interface inicial permitia que o usuário avançasse, mesmo sem que tivesse finalizado as escolhas necessárias, foi realizado o ajuste de desabilitar o botão de avançar até que fosse realizada a ação apropriada.. Outro ajuste feito foi o carregamento do livro, restringindo a seleção a apenas arquivos de texto, impedindo a escolha de vídeos, arquivos PDF ou outros. Alguns outros ajustes serão tratadas a posteriori no novo período de iniciação científica, dedicado à continuidade desse trabalho..

CONSIDERAÇÕES FINAIS

Este trabalho científico explorou a avaliação e o desenvolvimento de métodos computacionais para as Humanidades Digitais. Duas ferramentas de extração de dados de publicações foram analisadas, sendo selecionada e ajustada a ferramenta do LucyLattes. Foi identificado que o LucyLattes possuía uma limitação no algoritmo de detecção de co-autoria, que foi ajustado com base no método proposto pelo ScriptLattes.

Além disso, para análise de títulos de publicações, três funcionalidades foram desenvolvidas: uma para cálculo de frequência de palavras, obtendo assim as palavras mais recorrentes para um conjunto de títulos; uma para cálculo de ocorrências de caracteres especiais em títulos, como vírgula, ponto de interrogação e dois pontos; e uma para detecção e visualização de rede co-ocorrência de palavras de títulos.

E por fim, foi avaliado o software, MIVES, capaz de automatizar a escansão de estruturas métricas de versificação em prosa de língua portuguesa. Para além do estudo desta ferramenta, houve a necessidade de ajuste da sua interface gráfica, bem como a correção de alguns erros encontrados.

De maneira geral, todos os objetivos propostos foram concluídos, mas haverá continuação de mais aperfeiçoamentos no MIVES, que em nossa avaliação possui mais demandas para avanços. Apesar disto, a pesquisa se mostrou extremamente valiosa para a comunidade acadêmica, principalmente para a comunidade de pesquisadores da área de Humanidades.

REFERÊNCIAS

BALANCIERI, R., BOVO, A. B., KEM, V. M., PACHECO, R. C. D. S., & BARCIA, R. M. (2005). A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. Ciência da informação, 34, 64-77.

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.

CORRÊA, T. S., & SUZUKI, M. B., & CINTRA, P. R. O fim do scriptLattes? Uma análise de suas funcionalidades, alternativas para o presente e perspectivas para o futuro. Revista do EDICC - ISSN 2317-3815, v. 3, 2017.

JAMALI, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. Scientometrics, 88(2), 653–661.

KANG, I. S., NA, S. H., LEE, S., JUNG, H., KIM, P., SUNG, W. K., & LEE, J. H. (2009). On co-authorship for author disambiguation. Information Processing & Management, 45(1), 84-97.

LOULA, A. (2021). Humanidades Digitais: análise e síntese computacionais de texto verbal em língua portuguesa. Projeto de Pesquisa, Universidade Estadual de Feira de Santana, Bahia.

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.

MENA-CHALCO, J. P., & JÚNIOR, R. M. C. (2013). Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João, 109-128.

NAVARRO, G. (2001). A guided tour to approximate string matching. ACM Computing Surveys. 33 (1): 31–88.

TIEPPO, Rafael (2021). LucyLattes script para a extração e compilação de dados do currículo Lattes. Disponível em https://github.com/rafatieppo/lucylattes.

TRAMUNT IBAÑOS, A., & BATISTA PAIL, D. (2017). Fundamentos linguísticos e computação. EDIPUCRS