UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto n°17.228 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

DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA CONVERSÃO DE REPRESENTAÇÕES SIMBÓLICAS EM LIVROS DIDÁTICOS NO FORMATO OCR PDF PARA TXT.

Matheus Oliveira dos Santos[1]; Marcos Grilo Rosa[2];
1. Bolsista PIBIC/FABESB, Graduando em Licenciatura em Matemática, Universidade Estadual de Feira de Santana, e-mail: matheusods10@gmail.com
2. Orientador, Departamento de Exatas, Universidade Estadual de Feira de Santana, e-mail: grilo@uefs.br

PALAVRAS-CHAVE: Redes Semânticas; Livro Didático; Latex; Símbolos Matemáticos; Linguagem Python.

INTRODUÇÃO

Pesquisas com redes semânticas possuem aplicações na análise de discursos escritos (CALDEIRA et al., 2006) e de discursos orais (TEIXEIRA et al., 2010), no estudo da difusão do conhecimento a partir de títulos de artigos científicos (FADIGAS et al., 2009; PEREIRA et al., 2011), na identificação de temáticas relevantes em trabalhos de conclusão de curso (SANTOS e GRILO, 2020), dentre outras aplicações. Nestes trabalhos, cliques foram utilizadas para representar as sentenças/títulos. Uma clique é um conjunto de vértices mutuamente conectados. Por meio de um tratamento computacional desenvolvido por Caldeira et al. (2006), eliminam-se as palavras sem significado intrínseco (e.g. artigos, preposições). As palavras com significado intrínseco (e.g. substantivos, verbos) são escritas em uma forma canônica (e.g. substantivos são convertidos para o masculino singular, verbos são convertidos para o infinitivo). Logo, os vértices representam palavras com significado intrínseco. Uma aresta entre duas palavras é estabelecida se ambas estiverem presentes em uma mesma sentença/título. Dessa maneira, obtém-se redes semânticas de cliques.

A maneira de construção de redes semânticas de cliques utilizadas nos trabalhos de Caldeira et al. (2006), Fadigas et al. (2009), Teixeira et al. (2010), Pereira et al. (2011) e Santos e Grilo (2020) requer que as sentenças/títulos sejam coletadas a partir de uma fonte de dados e armazenados em um arquivo no formato TXT. Devido às especificidades das fontes de dados, não havendo uma ferramenta computacional específica e eficiente, esta etapa é feita de forma manual. As sentenças são copiadas para um arquivo TXT, em um processo que pode ter que ser repetido milhares de vezes a depender da fonte de dados. Em seguida, é preciso que uma limpeza seja realizada para que dados desnecessários para a construção da rede semântica sejam eliminados. Cada sentença deve ser armazenada em uma linha. Símbolos devem ser escritos por extenso a fim de que o programa computacional que fará a construção das redes possa fazer o reconhecimento de forma

correta. Essa etapa de coleta e mineração dos dados, se feita manualmente, demanda bastante tempo.

O projeto de pesquisa base deste plano de trabalho utiliza redes semânticas de cliques para analisar livros didáticos de Matemática para a Educação Básica. Para tanto, avaliaremos as relações entre conceitos matemáticos e suas representações por meio de expressões matemáticas, esquemas, tabelas e figuras. Neste sentido, será necessário construir redes semânticas a partir de livros didáticos de Matemática em formato PDF. A primeira etapa da metodologia de construção de redes semânticas a ser utilizada no Projeto requisita que as sentenças sejam armazenadas em arquivos no formato TXT. Regras serão aplicadas para converter representações (e.g. expressões matemáticas, figuras, esquemas, tabelas) e exercícios em sentenças. Essa necessidade de um tratamento especial para os materiais de pesquisa, demanda um trabalho nesse sentido, se feito manualmente.

Neste plano de trabalho, apresentamos um método que converte símbolos matemáticos presentes em livros didáticos de Matemática no formato PDF OCR em TXT. Para tanto, desenvolvemos uma ferramenta computacional denominada de TEX-2-NET que converte textos em linguagem LaTex para um formato reconhecível pelo NetPal, programa gerador de redes de semânticas de cliques utilizado no projeto de pesquisa base desse plano de trabalho.

METODOLOGIA

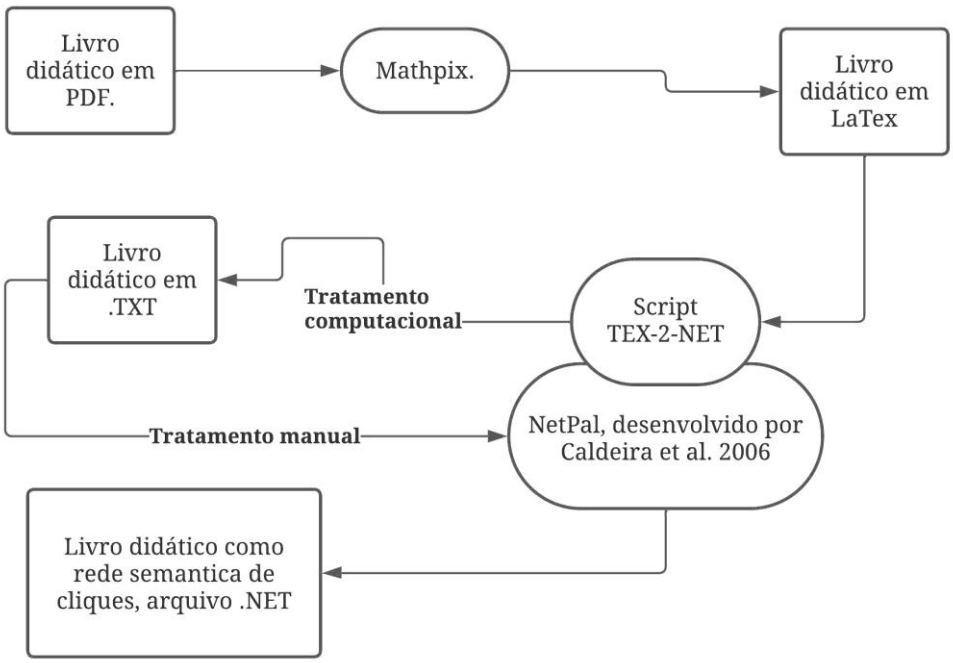
Para o desenvolvimento da ferramenta computacional, realizamos basicamente três passos: o levantamento e a leitura das documentações de códigos livres (PYTHON, 2023); a readequação do vocabulário de controle baseado na linguagem LaTeX; e o desenvolvimento do script em Python, que batizamos de TEX-2-NET. Inicialmente foi feito um levantamento de códigos livres que realizassem a transferência direta das informações contidas em arquivos no formato PDF OCR para TXT. No entanto, nenhum código, dentre os que encontramos, contemplava a conversão de símbolos matemáticos de um modo adequado à pesquisa. Dessa forma, ampliamos a pesquisa para o levantamento de ferramentas computacionais que reconhecessem símbolos matemáticos em arquivos no formato PDF OCR, não importando o formato do arquivo de saída nem a linguagem utilizada. Avaliamos o software Mathpix como uma solução adequada para a nossa pesquisa. O software é capaz de transformar um arquivo PDF em um documento em linguagem LaTeX, editor de textos consagrado na comunidade matemática.

Desta forma, conseguimos desenvolver uma ferramenta computacional na linguagem Python denominada de TEX-2-NET que realiza o tratamento do arquivo no formato Latex, resultante da coleta de informações presentes em livros didáticos de Matemática no formato PDF OCR. Além disso, o TEX-2-NET está integrado ao NetPal, automatizando o processo de geração de redes no formato .net. Para o desenvolvimento do TEX-2-NET, foi necessário reformular o vocabulário de controle que estava sendo elaborado de forma paralela a esta pesquisa. Com a introdução da linguagem Latex, o vocabulário passou a ser alimentado de forma muito mais adequada para os propósitos da pesquisa base desde plano de trabalho.

RESULTADOS E DISCUSSÃO

A ferramenta computacional desenvolvida, TEX-2-NET, automatiza vários processos que antes seriam bastante maçantes e demorados ou até mesmo impossíveis se feitos manualmente. Com o TEX-2-NET, diminui-se a chance de erros no processo de tratamento dos dados. A Figura 1 sintetiza o processo descrito na Seção Metodologia

Figura 1. Processo de construção de redes semânticas baseadas em livros didáticos de Matemática.



O desenvolvimento da ferramenta computacional TEX-2-NET foi direcionado para adequar um texto LaTeX em um arquivo TXT que permite a criação de uma rede semântica de cliques. Para isso, foram desenvolvidos na linguagem de programação Python métodos de substituição de palavras, manipulação da interface da linha de comando e uma integração com o software NetPal O tratamento manual, necessária para os objetivos da pesquisa base deste plano de trabalho, está presente na execução do TEX-2-NET possibilitando que o usuário corrija coisas que não foram percebidas pela automatização computacional.

CONSIDERAÇÕES FINAIS

Com o desenvolvimento do TEX-2-NET, vários procedimentos necessários para a geração de uma rede semântica de cliques baseada em livros didáticos de Matemática no formato PDF OCR foram automatizados. A incorporação da linguagem LaTeX através do software MathPix foi bastante adequada para a pesquisa base desse plano de trabalho. Pode-se considerar a linguagem LaTeX um modo consagrado na comunidade cientifica de produzir textos que utilizem muitos símbolos matemáticos. Desta forma, otimizamos o processo de geração das redes de forma considerável visto que eliminamos a etapa manual de copiar as informações contidas em arquivos PDF OCR em arquivos

de textos e simultaneamente, elaborar um vocabulário de controle para a substituição de símbolos matemáticos. Como trabalhos futuros, esperamos melhorar a qualidade do tratamento dos dados, aperfeiçoando o TEX-2-NET nesse processo de conversão de símbolos matemáticos em palavras.

REFERÊNCIAS

CALDEIRA, S. M. G. et al. The network of concepts in written texts. **The European Physical Journal B - Condensed Matter and Complex Systems**, v. 49, n. 4, p. 523–529, fev. 2006.

TEIXEIRA, G. M. et al. Complex Semantic Network. **International Journal Modern Physics C**, v. 21, n. 3, p. 333–347, 2010.

PEREIRA, H. B. B. et al. Semantic networks based on titles of scientific papers. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 6, p. 1192–1197, 15 mar. 2011.

SANTOS, V. C. DOS; GRILO, M. Identificação de temáticas de trabalhos de conclusão de curso por meio de redes semânticas. **Revista Paranaense de Educação Matemática; Vol. 9, No 20 (2020)**.

FADIGAS, I. DE S. et al. Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. **Educação Matemática Pesquisa : Revista do Programa de Estudos Pós-Graduados em Educação Matemática; v. 11, n. 1 (2009)**, 24 jan. 2010.

BRASIL. Programa Nacional do Livro e do Material Didático – PNLD. Ministério da Educação, s.d.. Disponível em: <https://www.fnde.gov.br/index.php/programas/programas-do-livro/pnld/escolha-pnld-2021-projetos>. Acesso em 01 mai. 2022.

BRASIL. Programa Nacional do Livro e do Material Didático – PNLD. Ministério da Educação, 2021. Disponível em:<http://portal.mec.gov.br/component/content/article?id=12391:pnld >. Acesso em 01 mai. 2022.

PYTHON SOFTWARE FOUNDATION. A Referência da Linguagem Python¶. Python, 2001-2023. Disponível em: <https://docs.python.org/pt-br/3/reference/index.html#reference-index>. Acesso em 31 ago. 2023.