ESTUDO PILOTO

Processamento de linguagem natural aplicado às anotações elaboradas com o DLNotes

Luis Gustavo Saibro da SILVA
Universidade Federal de Santa Catarina (UFSC)

Elder RIZZON SANTOS
Universidade Federal de Santa Catarina (UFSC)

OPEN ACCESS

EDITADO POR
- Roberlei Alves Bertucci (UTFPR)
- Emanoel Cesar Pires de Assis (UEMA)
- Rebeca Schumacher Eder Fuão (UiO)

AVALIADO POR
- Ana Patrícia Sá Martins (UEMA)
- Maurini de Souza (UFTPR)

SOBRE OS AUTORES
- Luis Gustavo Saibro da Silva Escrita – rascunho original – e Escrita – análise e edição.
- Elder Rizzon Santos Supervisão e Revisão da escrita.

DATAS
- Recebido: 16/10/2023
- Aceito: 12/03/2024
- Publicado: 11/05/2023

COMO CITAR
Silva, L. G. S.; Rizzon Santos, E. (2024). Processamento de linguagem natural aplicado às anotações elaboradas com o DLNotes. *Revista da Abralin*, v. 23, n. 2, p. 261-279, 2024.

RESUMO

O DLNotes é uma ferramenta para anotações em obras digitais utilizada durante o processo de ensino e aprendizagem de disciplinas de Literatura. A utilização do DLNotes nos últimos anos resultou em um grande acervo de anotações, o qual pode ser processado e explorado computacionalmente com intuito de produzir e descobrir conhecimento a partir dos dados. Neste trabalho foram adotadas técnicas de processamento de linguagem natural para o pré-processamento e produção de um conjunto de dados que facilite a extração de conhecimento. Além dos dados do DLNotes, foram integrados dados de atividades realizadas via Moodle. Como prova de conceito, o conjunto de dados foi utilizado na previsão da avaliação de atividades com base nas anotações dos alunos. Esta aplicação visa dar celeridade no feedback ao aluno e apoiar o professor na etapa de avaliação. A contribuição principal deste trabalho é a abordagem adotada para construção do conjunto de dados e o relato dos resultados preliminares na previsão das avaliações.

ABSTRACT

The DLNotes is an annotation tool tailored to digital texts adopted during the teaching and learning process of Literature courses. The adoption of the DLNotes system in the last years resulted in a large annotation collection suitable for computational processing and discovery aimed at

producing more knowledge. Natural language processing techniques were adopted in this project to develop a dataset allowing the extraction of knowledge. In addition to the data from DLNotes external data from the Moodle learning system is also aggregated in the proposed dataset. The resulting dataset was applied to the prediction of the teacher´s evaluation of activities based on the student´s annotation. This prediction model was developed as proof of concept of the dataset. Furthermore, the prediction is aimed at speeding up the student feedback and supporting the teacher during the evaluation process. Finally, the main contribution of this work is the adopted approach to construct the dataset and the preliminary results report from the evaluation prediction.

PALAVRAS-CHAVE
Processamento de linguagem natural. Extração de conhecimento. Modelos baseados em dados. Predição.

KEYWORDS
Natural language processing. Knowledge extraction. Data driven models. Prediction.

RESUMO PARA NÃO ESPECIALISTAS

Considerando-se a grande disponibilidade de obras digitais na Internet, torna-se relevante a adoção de ferramentas que facilitem o estudo e análise dessas obras. Uma ferramenta que possibilita essa interação é o DLNotes o qual permite que as obras sejam anotadas tal como fizéssemos uma anotação em um livro ou artigo físico. O DLNotes é utilizado por professores e alunos no processo de ensino e aprendizagem de disciplinas de Literatura. Tendo em vista essa adoção, já foi produzida uma grande quantidade de anotações as quais estão armazenadas de maneira digital. O presente trabalho propõe um sistema computacional capaz de explorar estas anotações de maneira a produzir mais conhecimento a partir delas. A principal finalidade desse sistema é prever a nota de um estudante com base nas anotações feitas por ele em uma atividade. Esta aplicação visa dar celeridade no feedback ao aluno e apoiar o professor na etapa de avaliação. A contribuição principal deste trabalho é a abordagem adotada para o desenvolvimento do sistema de predição e a análise dos resultados obtidos.

Introdução

A quantidade de informação originada de textos livres aumenta a cada minuto e, portanto, abordagens automáticas para extração de conhecimento destes textos são relevantes para diminuir a sobrecarga de informação (Edmunds; Morris, 2000) e possibilitar o uso direto da informação por aplicações computacionais ou usuários finais.

Neste cenário, o Processamento de Linguagem Natural (PLN) (Chowdhary, 2020) agrupa abordagens e ferramentas para a realização de tarefas envolvendo a análise e interpretação textual por meios computacionais (desde a análise léxica até a produção textual). Esta área vem se tornando cada vez mais predominante entre os tópicos mais procurados em Sistemas de Informação, pois possui uma imensa gama de aplicações dada a natureza das informações disponíveis em páginas na Internet.

O PLN, apesar de lidar fundamentalmente com símbolos, apresenta melhores desempenhos nas técnicas baseadas em modelos (Bishop, 2013) do que nas puramente simbólicas (Russel; Norvig, 2020), por exemplo, um modelo pode ser sintetizado como um algoritmo (uma sequência de instruções) ou um conjunto de funções matemáticas capazes de se autoajustarem (aprenderem) com base em exemplos, a saber, as informações disponíveis.

Ademais, o desenvolvimento e a aplicação de modelos não são exclusividade do PLN e ultimamente são destaque no estado da arte em "Aprendizagem de Máquina" devido à atual disponibilidade de dados e de poder de processamento. Considerando que um modelo é construído com base nos dados de exemplo, notamos que quanto mais exemplos, mais casos o modelo será capaz de lidar e mais conhecimento implícito será representado (Abel; Rama Fiorini, 2013).

Ainda sobre o que chamamos de modelo, sua construção possui diversas etapas, sendo as mais importantes o treinamento e a avaliação, as quais são realizadas com subconjuntos dos dados distintos de modo a avaliá-lo com outros dados não vistos durante o treinamento (Wang *et al.* 2022). Os dados de entrada podem vir de diversas fontes, como textos de *blogs*, comentários na Internet, em livros etc. e os dados utilizados neste trabalho foram extraídos da ferramenta DLNotes2 (Willrich; Mittmann; Fileto, 2019), que é utilizada para fazer anotações em trechos de livros lidos por alunos da Graduação em disciplinas de Literatura.

A ferramenta DLNotes permite a inserção de anotações livres e anotações semânticas em trechos dos livros ou obras digitalizadas e disponibilizadas na plataforma da ferramenta. Além disso, o DLNotes tem sido utilizado em disciplinas de Literatura nas quais o professor responsável escolhe algumas obras para os estudantes analisarem utilizando as duas formas de anotação disponíveis: as livres e as semânticas.

As anotações livres possibilitam relacionar um trecho da obra a algum conteúdo qualquer - produzido ou não pelo estudante. Já as anotações semânticas entregam ao estudante conceitos definidos em ontologias para que ele realize associações. Segundo Gruber (1993), ontologias são um conjunto de primitivas representacionais que modelam um sistema de conhecimento em que são definidos os elementos e os relacionamentos presentes em um domínio, bem como regras e restrições para caracterizar e diferenciar as entidades.

Ainda sobre as anotações semânticas, elas podem ser utilizadas no contexto das ontologias para criar conjuntos de representação de conhecimento. Diante das anotações, no contexto de uma atividade, por exemplo, o professor responsável pela disciplina faz a avaliação de ambos os tipos de anotações atribuindo uma nota para cada uma delas e, assim, são justamente essas anotações e as avaliações que serão a entrada principal para o modelo de PLN.

A construção desse tipo de modelo tem início com a preparação dos dados, que também é chamada de pré-processamento. Esta etapa envolve a limpeza dos dados, a sua integração com outras fontes e a sua transformação e, logo depois disso, eles podem pertencer a três categorias diferentes, a saber, não estruturados, estruturados e semiestruturados (Nayak; Kanive, 2016). No contexto deste nosso trabalho, as anotações são não estruturadas (no caso das livres) ou estruturadas (no caso das semânticas).

A criação de um modelo é um processo interativo e adaptativo que envolve o teste de diversas abordagens até que se atinja um objetivo específico, em geral determinado por uma taxa de erro. No presente estudo, o conhecimento representado pelo modelo é utilizado para estimar a nota do estudante com base nas suas anotações. Tendo em vista que a ferramenta de anotações pode ser utilizada em diferentes contextos, as avaliações dos professores não estão integradas no banco de dados do DLNotes.

Desta maneira, os dados das avaliações têm sua origem no ambiente de aprendizagem utilizado por um professor de uma turma de graduação em Literatura. Por consequência, a etapa final da previsão da nota é considerada uma prova de conceito quanto ao uso do modelo, uma vez que a quantidade de notas disponíveis é pequena com relação à quantidade de anotações. Mesmo com essa limitação, é possível analisar o resultado do desenvolvimento do modelo e da previsão, descrevendo todo o processo para a extração e aplicação do conhecimento.

Isto posto, salientamos que este artigo está organizado da seguinte forma: na Seção 1 são apresentados os principais trabalhos relacionados a esta pesquisa, notadamente a ferramenta de anotações DLNotes; a Seção 2 descreve as etapas relacionadas à construção do conjunto de dados; na Seção 3 são apresentados os detalhes referentes ao desenvolvimento dos modelos a partir do conjunto de dados; por fim, na Seção 4 apresentamos as considerações finais desta pesquisa.

1. Trabalhos Relacionados

Mittmann *et al.* (2013) descrevem o DLNotes2 como uma ferramenta que possibilita estudantes e professores fazerem suas anotações em conteúdos de atividades educacionais. Por meio dela, o professor define uma atividade de anotações associada à um conjunto de conteúdos, os quais, atualmente, são obras literárias ou textos em geral. Apesar de existirem inúmeras ferramentas para anotações (Sun *et al.*, 2020), o DLNotes diferencia-se ao possibilitar anotações semânticas, as quais associam elementos de uma ontologia à alguma parte do conteúdo (Willrich *et al.*, 2019).

É importante ressaltarmos que em uma ontologia são definidos os conceitos (denominados classes ou categorias), as relações e os indivíduos (exemplares) de um domínio. No contexto do DLNotes,

a ontologia representa os principais elementos a serem identificados e analisados nas obras, por exemplo, os conceitos de autor, personagem, estilo literário e local e as relações de parentesco, amizade e "morar em". Assim, o uso destes elementos para anotar o conteúdo possibilita a construção automática de uma base de conhecimento que pode ser analisada e modificada graficamente, tanto pelo professor quanto pelo estudante que a criou.

Deste modo, a utilização de ontologias nas anotações possibilita três aspectos: fazer uma representação mais robusta entre as entidades e seus relacionamentos; permitir a troca de conhecimento entre diversos sistemas por meio de uma conceitualização compartilhada; e, por último, possibilitar o uso de tecnologias da *Web Semântica* (Berners-Lee; Hendler; Lassila, 2023), facilitando o processamento computacional e, por consequência, a execução de inferências para a descoberta de mais conhecimento a partir do que se está representado.

Ainda quanto à construção do conhecimento, existe a possibilidade de desenvolver e ampliar as ontologias de domínio de acordo com o problema estudado. Neste caso, os estudantes colaboram na criação de bases de conhecimento por meio de exemplares dos conceitos gerados a partir de suas anotações semânticas.

Além disso, Mittmann *et al.* (2013, p. 535) concluem que a ferramenta "oferece à aprendizagem eletrônica recursos para apoiar o ciclo de produção e agregação de conhecimento", auxiliando na construção e na associação do conteúdo estudado por meio das anotações livres e semânticas.

A base de conhecimento, portanto, resultante do uso do DLNotes, facilita o compartilhamento de informações e a descoberta de novos conhecimentos por meio de deduções lógicas, uma vez que a linguagem utilizada para construção das ontologias, a OWL[1] (*Web Ontology Language*), foi especialmente desenvolvida para essa finalidade e utiliza o paradigma da Inteligência Artificial Simbólica[2].

Para esta nossa pesquisa, as técnicas de PLN utilizadas visam ampliar a gama de conhecimentos extraída das anotações e das demais interações dos estudantes com os conteúdos. As anotações semânticas já entregam agregações de conhecimento, entretanto, isso não ocorre com as anotações livres. Tendo em vista que a maior parte das técnicas e implementações atuais de PLN estão fundamentadas em modelos baseados em dados, e não em lógica, é possível aplicar essa abordagem para ampliar o escopo de utilização do DLNotes.

Nesse sentido, descrevemos a aplicação de PLN para explorar especialmente as anotações livres e apresentamos, também, uma proposta de representação adicional para as anotações semânticas. Ademais, é possível modelar e associar informações relacionadas ao uso da ferramenta ainda que não façam parte da sua base de dados, tais como avaliações, interações e pesquisas. Inicialmente, como prova de conceito, são adicionadas ao modelo informações referentes às avaliações de uma atividade envolvendo anotações livres.

---

[1] A OWL, desenvolvida pela W3C (*World Wide Web Consortium*), estabelece uma linguagem padronizada para a definição de ontologias de maneira a facilitar o compartilhamento de conhecimento e o seu uso automático por aplicativos computacionais.

[2] Esse paradigma utiliza representações e formas de raciocínio fundamentadas em lógica para a modelagem de conhecimento.

A partir daqui, nas próximas seções são descritas as etapas desenvolvidas para construção de um conjunto de dados adequado para a posterior utilização de técnicas de aprendizagem de máquina visando a produção de um modelo baseado em dados para a previsão de avaliações de anotações.

2. Construção do conjunto de dados

A organização e a construção do conjunto de dados, bem como o posterior desenvolvimento da prova de conceito utilizaram a metodologia CRISP-DM (Wirth; Hipp, 2000) para a execução de projetos de mineração de dados e ciência de dados. Em resumo, essa metodologia propõe que sejam realizadas ações de maneira cíclica para o entendimento do domínio, a compreensão dos dados disponíveis, a preparação computacional dos dados, a produção de modelos, a avaliação dos modelos e, por último, a implantação das aplicações que fazem uso dos modelos. A Figura 1 ilustra a execução de um ciclo da metodologia no contexto da presente pesquisa.

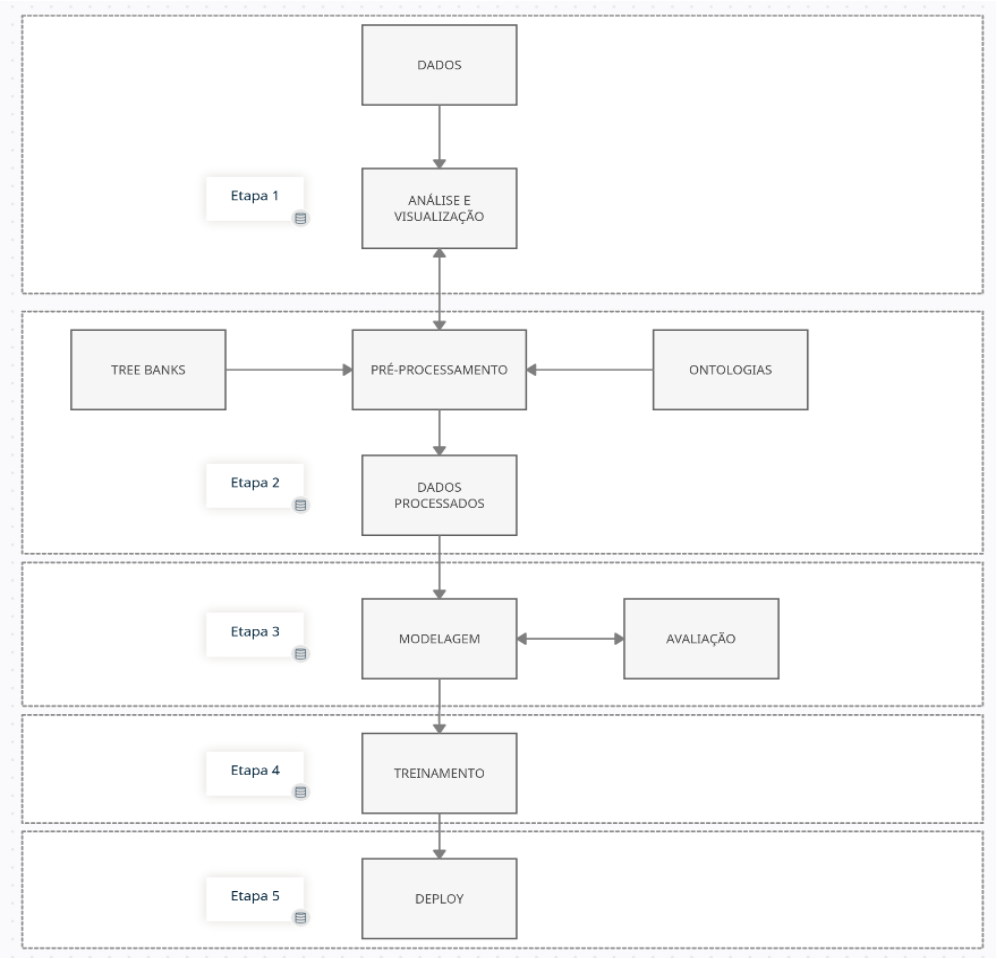

FIGURA 1 – Etapas adotadas segundo a metodologia CRISP-DM.
Fonte: Os autores (2024).

Seguindo a ordem apresentada na Figura 1, acima, na primeira etapa é abordada a compreensão do domínio e dos dados disponíveis. Conforme apresentado nas seções anteriores, a fonte principal dos dados foi o banco de dados relacional da ferramenta DLNotes2. Os dados principais foram retirados das tabelas de anotações livres, ontologias, anotações semânticas e atividades e, depois, foram unidos através da linguagem SQL (*Structured Query Language*) e convertidos para um arquivo separado por vírgulas, de maneira a facilitar o seu processamento por ferramentas de análise de dados. A Tabela 1, abaixo, apresenta a estrutura do conjunto de dados das anotações livres.

Dado	**Descrição**	**Exemplo**
exerpt	Intervalo selecionado para a anotação livre	A cada canto um grande conselheiro, // Que nos quer governar a cabana, e vinha, // Não sabem governar sua cozinha, // E podem governar o mundo inteiro.
content	Conteúdo da anotação	Nesta estrofe o poeta nos apresenta a realidade das fofocas rotineiras, em cada canto alguém querendo tomar conta da vida alheia sem ao menos tomar conta da sua.
title	Título da atividade proposta pelo professor da disciplina	Poemas de Gregório de Matos

TABELA 1 – Estrutura dos dados referentes às anotações livres.
Fonte: Os autores (2024).

O conjunto de dados de anotações livres contém 707 anotações, sendo 58 delas *links* para referências externas e a Figura 2, abaixo, mostra as 10 primeiras entradas desse conjunto.

```
free_annotations.head(10)
```

	excerpt	content	title
0	pertendo	<p>\n\tpertendo: pertender, pertendente, prete...	Poemas de Gregório de Matos
1	A salvação pertendo em tais abraços,	<p>\n\tPretende encontrar a salvaç&atil...	Poemas de Gregório de Matos
2	Delinqüido vos tenho, e ofendido, // Ofendido...	<p>\n\tO poeta assume, neste ponto, que ofende...	Poemas de Gregório de Matos
3	barrete,	<p>\n\tPeça do vestuário que se ...	Poemas de Gregório de Matos
4	Mancebo sem dinheiro, bom barrete, // Medíocr...	<p>\n\tDescreve, aqui, a indumentária d...	Poemas de Gregório de Matos
5	Presumir de dançar, cantar falsete, // Jogo d...	<p>\n\tDescreve, nesta estrofe, o comportament...	Poemas de Gregório de Matos
6	A putinha aldeã achada em feira, // Eterno mu...	<p>\n\tAqui, o poeta se refere ao comportament...	Poemas de Gregório de Matos
7	ser Quixote com as damas,	<p>\n\tAo dizer Ser Quixote com as da...	Poemas de Gregório de Matos
8	Pouco estudo: isto é ser estudante.	<p>\n\tAponta que, na vida de um estudante, o ...	Poemas de Gregório de Matos
9	A cada canto um grande conselheiro, // Que no...	<p>\n\tNesta estrofe o poeta nos apresenta a r...	Poemas de Gregório de Matos

Figura 2 – Primeiros 10 registros do conjunto de dados de anotações livres.
Fonte: Os autores (2024).

Nesta Figura 2, notamos que o conteúdo da coluna *content* apresenta a marcação de parágrafo <p> do HTML e possui uma formatação não usual. Após feita uma análise com a biblioteca chardet[3], foi observado que se trata da codificação em *Unicode Transformation Format*[4]. Já na coluna *excerpt*, observamos que existem diversos caracteres que podem afetar o modelo caso não sejam removidos, tais como pontuações e *stopwords* (quebras de linha, parênteses etc.).

Assim, a análise do conteúdo dos dados disponíveis indica aspectos que precisam ser adequados para facilitar o posterior processamento e análise textual. As adequações no conteúdo deste conjunto são descritas na próxima seção, a qual trata do pré-processamento dos dados. Com relação às ontologias e às anotações semânticas, foram extraídas apenas as informações necessárias para a representação de uma anotação desse tipo. Assim sendo, o conjunto contém o trecho anotado, a identificação do que está sendo anotado e a qual classe da ontologia ele pertence. A Tabela 2, abaixo, ilustra a estrutura desse conjunto de dados:

Dado	Descrição	Exemplo
excerpt	Intervalo selecionado para a anotação semântica	Luisiana,
identifier	Entidade de uma classe da ontologia	Lusiana
label	Classe da ontologia	Pessoa

TABELA 2 – Estrutura dos dados referentes às anotações semânticas.
Fonte: Os autores (2024).

[3] *The Universal Character Encoding Detector* - biblioteca de código aberto. Disponível em: https://pypi.org/project/chardet/ . Acesso em: 18 jan. 2024.

[4] Um tipo de codificação binária

Além dos dados obtidos do DLNotes2, foi elaborado um conjunto de informações contendo as notas atribuídas pelo professor para cada atividade de uma turma. Essa turma tinha somente 25 alunos e foram realizadas 4 atividades por aluno, totalizando 100 notas.

Em geral, bons modelos baseados em dados demandam uma quantidade muito maior deles para ter mais acerto nas previsões e, por isso, o modelo resultante desta pesquisa é considerado apenas uma prova de conceito, tendo em vista a utilização de uma pequena parcela dos dados de avaliações.

2.1 Pré-processamento e enriquecimento dos dados com PLN

Devido ao escopo deste trabalho, optamos por não avaliar as referências externas disponíveis no conjunto de anotações livres, haja vista que seria necessário desenvolver um *web scraper*[5] (Zhao, 2017) para obter o conteúdo das páginas. Deste modo, as anotações que continham *links* (referências externas) foram retiradas do conjunto para não prejudicar o treinamento dos modelos. Em seguida, foi realizada a filtragem do conjunto de dados para considerar e manter somente as anotações nas quais estavam disponíveis as notas atribuídas pelo professor da disciplina. Após esta filtragem, o conjunto de dados continha apenas 385 registros.

Na etapa de análise do conjunto de dados foi identificado um problema com a coluna *content*. Os dados contidos nela estão dentro de uma marcação HTML e, por esse motivo, é necessário processá-los para obter somente o conteúdo de dentro da marcação. Para tanto, foi utilizada a biblioteca BeautifulSoup[6], por meio da qual é possível "desencapsular" o conteúdo contido em hipertextos. Assim, foi criada uma coluna para os dados "desencapsulados", a saber, a *clean_content* e a Figura 3, a seguir, ilustra esse resultado.

clean_content	vectorized_content	vectorized_excerpt
A poesia de Lord Byron inspira a segunda geraç...	[1.8283124, 1.9294281, -0.5927496, -0.18052478...	[-2.5138092, 1.4757074, -1.2350407, 1.2805302,...
A personagem, em vários momentos da obra me fe...	[1.3405192, 1.3653994, 0.01631254, -0.06349088...	[3.8362758, 0.5435313, -0.7067365, 1.5985585, ...
Essa passagem me fez lembrar da obra "Romeu e ...	[1.481269, 1.0512458, 0.90511817, -0.14004879,...	[1.1825184, 0.9921109, -0.3147725, -0.14165698...
Achei curioso o fato de que o autor se refere ...	[0.77085954, 0.64076585, -0.06110506, -0.55214...	[-1.7706503, 1.6898797, -2.399301, -0.7312531,...
Aurélia com as duas características mais preza...	[1.8284634, 1.0633658, -0.6671939, 0.8883805, ...	[1.2686689, 1.0331404, 0.60578376, -0.9982712,...

FIGURA 3 – Conjunto de dados de anotações após a criação de vetores de palavras.
Fonte: Os autores (2024).

Considerando que o propósito desta etapa é preparar os dados para desenvolver modelos baseados neles, é necessário converter as informações textuais em representações numéricas porque

[5] Programas específicos para coletar dados de páginas *web*.

[6] Biblioteca de código aberto para extração de informações em páginas *web*. Disponível em: https://pypi.org/project/beautifulsoup4/ . Acesso em: 18 jan. 2024.

tais representações são indispensáveis uma vez que, computacionalmente, todo texto e seus caracteres são representados numericamente. Entretanto, a simples conversão de uma palavra para um número desconsidera o seu significado, o qual pode ser determinado pela sentença em que ela se encontra. Ademais, mesmo considerando o significado em uma sentença, acaba-se por excluir o uso dessa palavra em outros contextos.

Esse problema é abordado pelo PLN com a introdução dos vetores de palavras (*word embedings*) (Levy; Goldberg, 2014), os quais são construídos com grandes volumes de dados e, em geral, disponibilizados em grandes modelos de linguagens (*Large Language Models - LLM*).

Resumidamente, os vetores de palavras construídos com esse tipo de modelo de linguagem representam uma palavra considerando-se todos os contextos em que ela é encontrada. Tendo em vista que a representação parte dos dados disponíveis, é possível perceber certa limitação desse tipo de abordagem, uma vez que ela será tão boa quanto os textos utilizados. Neste trabalho, adotamos um dos modelos de linguagem para o Português disponibilizado pela biblioteca *Spacy*[7], o qual foi desenvolvido com base em textos de notícias.

Além da implementação fornecida pelo *Spacy*, para fins de verificação e comparação, também foi utilizada a construção de vetores de palavras por meio do BERT[8] (Tenney; Das; Pavlick, 2019).

Portanto, mediante o uso de vetores de palavras, é possível produzir uma representação dos conteúdos das anotações com informações contextuais e em um formato compatível com as abordagens de aprendizagem de máquina. Por conseguinte, as colunas *exerpt* e *content* foram vetorizadas e armazenadas em duas novas colunas contendo os vetores gerados pelo processo de vetorização da biblioteca *Spacy*. Esses conteúdos foram inseridos nas colunas *vectorized_excerpt* e *vectorized_content* e podem ser observados na Figura 3, já apresentada acima.

Outro passo importante no pré-processamento é conseguir identificar na estrutura do texto digitado pelo aluno as informações sintáticas contidas neste texto. Para isto, foi utilizada a técnica de etiquetagem morfossintática (*part of speech tagging*), também da biblioteca *Spacy*. Para todos os textos foram obtidas essas informações e elas foram resumidas em uma contagem simples de ocorrência de cada tipo de etiqueta (substantivo, advérbio, pronome etc.) contido no texto. A Figura 4, a seguir, ilustra a adição dessas informações no conjunto de dados.

---

[7] Biblioteca de código aberto para PLN. Disponível em: https://spacy.io/ . Acesso em: 18 jan. 2024.

[8] Uma arquitetura de rede neural utilizada para a construção de LLMs (*Large Language Models*, um modelo de inteligência artificial) a qual manteve-se como estado-da-arte até o desenvolvimento das abordagens via GPT (*Generative Pre-Training Transformer*).

vectorized_content	vectorized_excerpt	DET	NOUN	ADP	...	CCONJ	PRON	SPACE	ADV	SCONJ	AUX	NUM	SYM	X	INTJ
[1.8283124, 1.9294281, -0.5927496, -0.18052478...	[-2.5138092, 1.4757074, -1.2350407, 1.2805302,...	2	2	1	...	3	1	1	1	3	1	1	1	1	1
[1.3405192, 1.3653994, 0.01631254, -0.06349088...	[3.8362758, 0.5435313, -0.7067365, 1.5985585, ...	2	2	1	...	3	1	1	1	3	1	1	1	1	1
[1.481269, 1.0512458, 0.90511817, -0.14004879,...	[1.1825184, 0.9921109, -0.3147725, -0.14165698...	2	2	1	...	3	1	1	1	3	1	1	1	1	1
[0.77085954, 0.64076585, -0.06110506, -0.55214...	[-1.7706503, 1.6898797, -2.399301, -0.7312531,...	2	2	1	...	3	1	1	1	3	1	1	1	1	1
[1.8284634, 1.0633658, -0.6671939, 0.8883805, ...	[1.2686689, 1.0331404, 0.60578376, -0.9982712,...	2	2	1	...	3	1	1	1	3	1	1	1	1	1

FIGURA 4 – Conjunto de dados de anotações após a etiquetagem morfossintática.
Fonte: Os autores (2024).

Por último, então, aplicou-se a técnica de reconhecimento de entidades nomeadas para identificar nomes de personagens, de autores, cidades, entre outros. Novamente, adicionamos uma coluna contendo os valores vetorizados para cada uma das entidades reconhecidas nos textos, porém foi levantada a possibilidade de introdução de um viés no modelo, visto que nem todos os textos continham essas entidades. A Figura 5, abaixo, ilustra o conjunto de dados após a aplicação dessa técnica.

clean_content	vectorized_content	vectorized_excerpt	DET	NOUN	ADP	...	PRON	SPACE	ADV	SCONJ	AUX	NUM	SYM	X	INTJ	NER
A poesia de Lord Byron inspira a segunda geraç...	[1.8283124, 1.9294281, -0.5927496, -0.18052478...	[-2.5138092, 1.4757074, -1.2350407, 1.2805302,...	2	2	1	...	1	1	1	3	1	1	1	1	1	[[1.2726889, 2.8953836, -0.7336545, 1.0329155,...
A personagem, em vários momentos da obra me fe...	[1.3405192, 1.3653994, 0.01631254, -0.06349088...	[3.8362758, 0.5435313, -0.7067365, 1.5985585, ...	2	2	1	...	1	1	1	3	1	1	1	1	1	[[2.4862916, 2.107809, 1.8457295, -0.040821433...
Essa passagem me fez lembrar da obra "Romeu e ...	[1.481269, 1.0512458, 0.90511817, -0.14004879,...	[1.1825184, 0.9921109, -0.3147725, -0.14165698...	2	2	1	...	1	1	1	3	1	1	1	1	1	[[1.3252956, 1.9256511, -1.6719532, -2.177884,...
Achei curioso o fato de que o autor se refere ...	[0.77085954, 0.64076585, -0.06110506, -0.55214...	[-1.7706503, 1.6898797, -2.399301, -0.7312531,...	2	2	1	...	1	1	1	3	1	1	1	1	1	[[1.465732, 0.2867495, -3.225106, 1.086696, 2....
Aurélia com as duas características mais preza...	[1.8284634, 1.0633658, -0.6671939, 0.8883805, ...	[1.2686689, 1.0331404, 0.60578376, -0.9982712,...	2	2	1	...	1	1	1	3	1	1	1	1	1	[[0.77389693, 4.9895525, -2.745476, 6.034116, ...

FIGURA 5 – Conjunto de dados de anotações após o reconhecimento de entidades nomeadas.
Fonte: Os autores (2024).

Depois da preparação do conjunto de anotações livres, verificamos quais conteúdos de anotações continham entidades representadas no conjunto de dados das ontologias. Assim, caso fossem encontradas entidades que pertencem ao conjunto de dados das ontologias, elas seriam colocadas em uma nova coluna, no entanto, não foram encontradas entidades das ontologias disponíveis nas anotações consideradas. Vale ressaltarmos que isso se deve, provavelmente, à utilização dos dados filtrados com relação às avaliações disponíveis, o que diminui consideravelmente a quantidade deles.

Assim sendo, encerrada a preparação do conjunto de dados, concluíram-se as etapas 1 e 2 do método CRISP-DM, cujo principal produto é o conjunto de dados pronto para a execução dos métodos para produção dos modelos baseados em dados.

A seguir, trataremos da execução das etapas 3 e 4 do método adotado tratando, essencialmente, da construção e da avaliação do modelo para a predição de avaliações.

3. Modelo para previsão de avaliações

A construção de modelos baseados em dados é amplamente estudada pela área de Aprendizagem de Máquina (Carbonell; Michalski; Mitchell, 1983) e, como o próprio nome sugere, são propostas abordagens que aprendam o que fazer ao invés de serem programadas para fazer algo. Neste caso, a aprendizagem supervisionada é concretizada por meio de tentativa e erro, ou seja, o computador faz uma tentativa para alcançar um objetivo e verifica quão bom foi o resultado comparando-o com a resposta esperada (fornecida pelo conjunto de dados).

No contexto desta nossa discussão, a resposta é a nota do estudante, a qual o modelo precisa descobrir utilizando as demais colunas do conjunto de dados (trecho anotado, vetores de palavras, etiquetas morfossintáticas, entre outras, conforme vimos na Figura 5, apresentada acima).

Nas primeiras tentativas, o modelo calcula uma resposta com base em pesos atribuídos aleatoriamente às colunas dos dados, ou seja, atribui-se uma importância a cada uma das características conhecidas, porém, como ainda não se sabe qual é essa importância, ela inicia com um valor aleatório. Assim, em geral, quanto maior o peso, maior a importância e vice-versa. Por exemplo, o modelo pode atribuir o peso 2,0 para o vetor de palavras, 1,5 para a etiqueta substantivo e −2,0 para a etiqueta advérbio. Tendo em vista que na etapa de preparação dos dados todas as anotações foram convertidas para números, o resultado do modelo será a multiplicação entre peso e o respectivo valor da coluna (Hopfield, 1988).

Nesse sentido, uma vez que os pesos foram atribuídos aleatoriamente, o modelo inicia errando em quase todas as situações. Os ajustes nos pesos são obtidos mediante os métodos de aprendizagem de máquina, os quais utilizam-se de métodos numéricos e estatísticos (James *et al.*, 2021) para propor os melhores ajustes para cada um dos pesos e, além disso, a escolha sobre qual método de aprendizagem utilizar depende, especialmente, do que se deseja obter como resultado.

Neste trabalho, desejamos obter uma nota, ou seja, um número. Contudo, outra possibilidade seria obter uma categorização, por exemplo, a partir de uma foto, identificando o sujeito nela. Neste caso, a saída não seria um número, mas, sim, uma categoria ou classe, ou seja, o sujeito poderia ser uma pessoa, um gato, um cachorro etc.

As duas principais possibilidades de saída do modelo (um número ou uma classe) determinam duas grandes categorias de métodos de aprendizagem: regressão e classificação. Assim, os métodos de regressão originam-se da estatística (regressão linear) (James *et al.*, 2021) e são utilizados para a obtenção de valores numéricos a partir dos dados de entrada. Já os métodos de classificação têm sua origem na aprendizagem supervisionada, cuja técnica mais proeminente é mediante o uso de redes neurais artificiais (Hopfield, 1988).

No que se refere à previsão das avaliações foram construídos dois modelos mediante técnicas distintas, o que possibilita uma análise comparativa entre eles. Sendo assim, o primeiro modelo foi construído por um método baseado em regressão linear e, o segundo, pelo uso de máquinas de vetores de suporte (Lorena; De Carvalho, 2007). Ambos os métodos são amplamente conhecidos e estão disponíveis na maior parte das bibliotecas de aprendizagem de máquina.

Além disso, para facilitar a reprodutibilidade dos experimentos avaliativos, foram utilizados os métodos disponibilizados pela biblioteca *scikit-learn*[9], que também facilita a padronização das avaliações dos modelos, calculando-as automaticamente. A métrica padrão para os modelos criados é o R^2 (coeficiente de determinação) por se tratar de modelos de regressão. Nessa métrica, quanto mais próximo de 1 for o valor de R^2, melhor é o modelo.

Então, considerando que o conjunto de dados elaborado apresenta duas abordagens para vetorização de palavras (*Spacy* ou BERT), as duas abordagens são avaliadas separadamente com ambos os modelos de regressão. Além disso, avaliamos também o impacto do uso da última coluna do conjunto de dados, o reconhecimento de entidades nomeadas. Essas possibilidades foram combinadas para treinar um total de 8 modelos para previsão de notas. Estes modelos foram avaliados com dados não utilizados durante o treinamento.

Para melhor entendimento e visualização, elaboramos a Tabela 3, a seguir, que apresenta um sumário do desempenho dos modelos.

Modelo	Vetor de palavras	Entidades Nomeadas	Método de aprendizagem	Treinamento	Teste
1	Spacy	Não	RL	0,0155	00227
2	Spacy	Não	SVM	-0,1443	0,1695
3	Spacy	Sim	RL	0,0120	0,0113
4	Spacy	Sim	SVM	-0,1049	-0,0238
5	BERT	Não	RL	0,0381	0,0679
6	BERT	Não	SVM	0,9994	-0,0138
7	BERT	Sim	RL	0,0411	0,0647
8	BERT	Sim	SVM	0,9994	-0,0021

TABELA 3 – Desempenho dos modelos desenvolvidos.
Fonte: Os autores (2024).

A coluna “vetor de palavras” descreve qual abordagem para obtenção dessa informação foi utilizada; em seguida, a coluna “entidades nomeadas” indica se essa informação foi utilizada durante o treinamento; a coluna “método de aprendizagem” informa qual o método utilizado (regressão linear,

[9] Biblioteca de código aberta para análise de dados e aprendizagem de máquina. Disponível em: https://scikit-learn.org/stable/ . Acesso em: 18 jan. 2024.

indicado por RL ou máquinas de vetores suporte, indicado por SVM); já as duas últimas colunas, "treinamento" e "teste", referem-se ao desempenho do respectivo método de aprendizagem treinado com as configurações dadas pelas duas primeiras colunas.

Além disso, há a coluna "treinamento", que apresenta a nota obtida ao fim do ajuste dos pesos, enquanto a coluna "teste" apresenta a nota do modelo executado em dados novos e que não foram utilizados durante o treinamento. Para ambas as colunas, quanto mais próximo de 1 for a nota, melhor o modelo.

Assim, a partir dos resultados obtidos observamos que as abordagens com o BERT apresentaram melhor desempenho durante o treinamento com os modelos 6 e 8 alcançando valores muito próximos de 1. Entretanto, o desempenho desses modelos foi ruim na etapa de teste, indicando que a previsão das notas está pior do que se fosse utilizada uma média simples. Essa situação aponta o sobreajuste dos pesos, ou seja, o modelo está ajustado demais ao conjunto de treino e não é capaz de lidar com situações diferentes.

Ademais, o modelo que apresentou o melhor desempenho foi o modelo 7, embora salientemos que os valores obtidos estão muito próximos do 0, o que indica um resultado ruim na previsão do modelo. Para seguir com a última etapa do método CRISP-DM, que é a implantação do modelo para uso e obtenção de *feedback* diretamente pelos professores e estudantes, seria necessário no mínimo valores de 0,9 para o treinamento e 0,8 para o teste.

Por conseguinte, e conforme proposto no delineamento deste artigo, os modelos aqui apresentados são considerados provas de conceito, uma vez que a quantidade de dados avaliativos disponíveis foi baixa (385 registros) considerando-se a quantidade de características (colunas) utilizadas para o treinamento. Ainda assim, apesar do resultado quantitativo abaixo do esperado, cumpre-se o objetivo de apresentar uma forma para utilização de PLN para produção de um conjunto de dados apto a ser utilizado por métodos de aprendizagem de máquina.

Considerações finais

Nas seções anteriores deste nosso artigo, buscamos descrever todas as etapas realizadas para a construção de um modelo para previsão de avaliações fundamentado em dados textuais de anotações de alunos. A etapa mais trabalhosa do desenvolvimento desse tipo de modelo de Inteligência Artificial está, quase sempre, na preparação dos dados e, no caso do modelo aqui apresentado, essa etapa merece destaque uma vez que as técnicas de PLN foram utilizadas como meio para extração e representação do conhecimento, e não como um fim.

Dessa maneira, o conjunto de dados desenvolvido pode ser utilizado para aplicações além da previsão de notas dos estudantes. No contexto do DLNotes, consideramos relevantes aplicações tais como: a sugestão de conteúdos adicionais, sumarização de conteúdos, assistentes para facilitar a correção de atividades e análise de estilos literários.

Além do mais, as representações vetoriais dos conteúdos das anotações foram obtidas de duas maneiras distintas, o que possibilitou também uma comparação de desempenho entre elas. Uma

destas abordagens utilizou o BERT, considerado o estado-da-arte na maioria das tarefas de PLN. Hoje, essa abordagem ainda se destaca, ainda que já esteja ultrapassada, o que abre algumas possibilidades interessantes para a continuidade dessa pesquisa avaliando-se os modelos mais recentes para produção de vetores de palavras, reconhecimento de entidades nomeadas e demais tarefas realizadas pelo PLN.

Outro aspecto importante a se tratar nesta nossa pesquisa é sobre a pequena quantidade dos dados de avaliação, o que impossibilita o uso dos modelos desenvolvidos devido ao baixo desempenho obtido. Tendo em vista que o mecanismo para a construção das representações computacionais encontra-se pronto, o principal trabalho futuro para aprimorar o desempenho dos modelos é considerar dados avaliativos de mais turmas. Idealmente, o desenvolvimento de uma integração com o Moodle permite a obtenção automática das avaliações e, por consequência, a utilização de dados avaliativos de todas as turmas que fazem uso do DLNotes.

Além disto, independentemente do modelo para previsão de notas, o processo para a construção das representações vetoriais possibilita novos caminhos para a evolução da ferramenta DLNotes por meio de abordagens que facilitem a integração de diferentes fontes de dados, tais como as propostas em Giebler; Gröger; Hoos (2019) e Sawadogo e Darmont (2020).

Por fim, voltamos a ressaltar que a abordagem aqui apresentada pode ser aplicada para o desenvolvimento de aplicações para muitos contextos educacionais ao apoiar o ensino de redação, leitura crítica e outros tópicos de interesse da comunidade de Letras, Linguística, Literatura e Pedagogia.

Informações complementares

Avaliação e resposta dos autores

Avaliação: https://doi.org/10.25189/rabralin.v23i2.2199.R

Editores

Roberlei Alves Bertucci
Afiliação: Universidade Tecnológica Federal do Paraná
ORCID: https://orcid.org/0000-0003-4014-5610

Emanoel Cesar Pires de Assis
Afiliação: Universidade Estadual do Maranhão
ORCID: https://orcid.org/0000-0001-7377-8540

Rebeca Schumacher Eder Fuão
Afiliação: Universidade de Oslo
ORCID: https://orcid.org/0000-0002-7658-7704

RODADAS DE AVALIAÇÃO

Avaliador 1: Ana Patrícia Sá Martins
Afiliação: Universidade Estadual do Maranhão
ORCID: https://orcid.org/0000-0002-5716-1580

Avaliador 2: Maurini de Souza
Afiliação: Universidade Tecnológica Federal do Paraná
ORCID: https://orcid.org/0000-0001-8914-2133

AVALIADOR 1

O artigo submetido apresenta uma relevante discussão no desenvolvimento de metodologias no processamento de linguagem natural aplicado às anotações elaboradas com o DLNotes, contribuindo, assim, para o diálogo e aproximação das pesquisas relacionadas à produção de softwares e o texto literário com as possiblidades didáticas no ensino-aprendizagem de Literatura. Ademais, o texto é claro e objetivo, com a explicitação das etapas realizadas na geração e no tratamento dos dados utilizados na referida proposta. Entretanto, para publicação, a revisão textual é necessária, observando-se não apenas o uso do padrão culto da língua portuguesa, mas também a utilização de alguns recursos linguísticos de modo a favorecer a textualidade. Recomendamos que se observe principalmente o emprego de conectores ligando enunciados, o uso de repertório lexical diversificado com vistas a: (a) que se evite a repetição excessiva de determinada expressão num mesmo parágrafo ou no todo do texto; e (b) que se valorize o registro coloquial no texto escrito com finalidade de exposição acadêmica. Ressalta-se que algumas palavras foram destacadas em amarelo visando facilitar a identificação de algumas dessas repetições excessivas.

AVALIADOR 2

Usar modo 'justificado' nos parágrafos, conforme regras de submissão. Cuidar com excesso ou falta de espaço entre palavras no texto. As notas de rodapé devem ter a mesma fonte do texto (Times) e tamanho 10.

Os parágrafos poderiam passar por uma revisão de escrita, de modo a torná-la mais fluente. A falta de elementos de coesão, por exemplo, torna o texto bastante duro para o público em geral.

Considerando o público da revista, uma explicação maior sobre o DLNotes e a formação de ontologias seria bem-vinda. Prints explicativos, por exemplo, seriam ótimos recursos.

Nas considerações finais, são apresentados aspectos limitantes. Sugiro aos autores que possam incluir como poderiam (ou farão, no futuro) superar esses aspectos. Outro aspecto que considero uma lacuna é uma possível comparação do DLNotes com outras ferramentas de anotação (e a verificação se alguma comparação com elas poderia ser possível). Acredito ser outro fator interessante a ser acrescido à última seção.

Considerando tais aspectos, considero que o artigo pode ser aceito para a publicação após uma revisão.

Conflito de Interesse

Os autores não têm conflitos de interesse a declarar.

Protocolo e Pré-Registro de Pesquisa

Avaliando os roteiros propostos pela Equator Network, consideramos que nenhum deles se mostra relevante para a pesquisa em tela. Também informamos que a pesquisa desenvolvida não foi pré-registrada em repositório institucional independente.

Declaração de Disponibilidade de Dados

Os dados, códigos e materiais que suportam os resultados deste estudo estão disponíveis abertamente no repositório GitHub através de https://github.com/GustavoSaibro/TCC

Fontes de financiamento

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Ampara à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC).

REFERÊNCIAS

ABEL, Mara; RAMA FIORINI, Sandro. Uma Revisão Da Engenharia Do Conhecimento: Evolução, Paradigmas E Aplicações. **International Journal of Knowledge Engineering and Management,** v. 2, n. 2, p. 1, 2013.

BERNERS-LEE, Tim; HENDLER, James A.; LASSILA, O. The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. *In*: HENDLER, James (org.). **Linking the World's Information:** Essays on Tim Berners-Lee's Invention of the World Wide Web. [s.l.]: ACM eBooks, 2023. p. 91-103.

BISHOP, Christopher M. Model-based machine learning. **Philosophical Transactions of the Royal Society A**: Mathematical, Physical and Engineering Sciences, v. 371, n. 1984, 2013.

CARBONELL, Jaime G.; MICHALSKI, Ryszard S.; MITCHELL, Tom M. An overview of Machine Learning. *In*: CARBONELL, Jaime G.; MICHALSKI, Ryszard S.; MITCHELL, Tom M. (org.). **Machine Learning - An Artificial Intelligence Approach**, Volume I. [s.l.]: Morgan Kaufmann, 1983. p. 3-23.

CHOWDHARY, K. R. Natural Language Processing. **Fundamentals of Artificial Intelligence**, New Delhi, p. 603-649, 2020.

EDMUNDS, Angela; MORRIS, Anne. The problem of information overload in business organisations: a review of the literature. **International Journal of Information Management**, v. 20, n. 1, p. 17-28, 2000.

GIEBLER, Corinna; GRÖGER, Christoph; HOOS, Eva; *et al*. Leveraging the Data Lake: Current State and Challenges. *In*: ORDONEZ, C.; SONG, I.Y.; ANDERST-KOTSIS, G. (org.). **Big Data Analytics and Knowledge Discovery**, v. 11708, p. 179-188, 2019.

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199–220, 1993. Disponível em: https://dl.acm.org/citation.cfm?id=173747. Acesso em: 15 jan. 2024.

HOPFIELD, J.J. Artificial neural networks. *IEEE* **Circuits and Devices Magazine**, v. 4, n. 5, p. 3-10, 1988.

JAMES, Gareth; *et al*. **An Introduction to Statistical Learning.** New York, NY: Springer US, 2021.

LEVY, Omer; GOLDBERG, Yoav. Dependency-Based Word Embeddings. Annual Meeting of the Association for Computational Linguistics (ShortPapers), 52., [s.l.], 2014. **Proceedings** [...]. [s.l], 2014. p. 302-308.

LORENA, Ana Carolina; DE CARVALHO, André C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 14, n. 2, p. 43-67, 2007.

MITTMANN, Adiel; WILLRICH, Roberto; FILETO, Renato; SANTOS, Alckmar Luiz; ASSIS, Emanoel C. Pires; SANDOVAL, Isabela Melin Borges. DLNotes2: Anotações Digitais como Apoio ao Ensino. SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 24., 2013. **Anais** [...]. Campinas: SBIE, 2013. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/2531. Acesso em: 18 jan. 2024.

NAYAK, Arjun Srinivas; KANIVE, Ananthu P. Survey on Pre-Processing Techniques for Text Mining. **International Journal of Engineering and Computer Science**, v. 5, n. 5, 2016.

RUSSEL, Stuart; NORVIG, Peter. **Artificial intelligence:** a Modern approach**.** 4. ed. Upper Saddle River, New Jersey: Prentice Hall, 2020.

SAWADOGO, Pegdwendé; DARMONT, Jérôme. On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, v. 56, n. 1, p. 97-120, 2020.

SUN, Chunmei. *et al*. Trends and issues of social annotation in education: A systematic review from 2000 to 2020. **Journal of Computer Assisted Learning**, v. 39, n. 2, p. 329-350, 2022.

TENNEY, Ian; DAS, Dipanjan; PAVLICK, Ellie. BERT Rediscovers the Classical NLP Pipeline. Annual Meeting of the Association for Computational Linguistics, 57., 2019. **Proceedings [...].** Florence: ACL, 2019. p. 4593-4601.

WANG, Xuezhi; WANG, Haohan; YANG, Diyi. Measure and Improve Robustness in NLP Models: A Survey. *NAACL* - **Annual Conference of the North American Chapter of the Association for Computational Linguistics 2022,** Seattle, 2022. Disponível em: https://arxiv.org/abs/2112.08313. Acesso em: 15 jan. 2024.

WILLRICH, Roberto; MITTMANN, Adiel; FILETO, Renato; SANTOS, Alckmar Luiz dos. Capture and visualization of text understanding through semantic annotations and semantic networks for teaching and learning. **Journal of Information Science**, v. 46, n. 4, p. 528-543, 2019.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 4., 2000. **Proceedings** [...]. San Diego: 2000. p. 29-39.

ZHAO, Bo. Web Scraping. **Encyclopedia of Big Data.** [s.l.: s.n.], p. 1-3, 2017.