REVISTA DO CCEI
ISSN: 1415-2061 2356-6635
Revista Técnico-Científica
EDIURCAMP
urcamp

Desenvolvimento de uma ferramenta computacional para análise de dados não estruturados

Development of a computational tool for analysis of unstructured data

Paulo Henrique Seixas Leite[1]; Vera Lúcia Duarte Ferreira[2]; Fernando Luis Dias[3]
Aden Rodrigues Pereira[4]

Resumo: O presente artigo apresenta a primeira versão de ferramenta computacional para análise de dados não estruturados, desenvolvida em linguagem Python e embasada em técnicas de mineração de texto. A ferramenta proposta tem como foco central a extração da frequência de palavras, bem como a determinação da matriz de termo de ocorrências de um corpus textual. A aplicação do experimento valeu-se de um corpus textual do gênero notícias composto por 18 textos cedidas pelo projeto de pesquisa "A intermediação da Linguística de Corpus na análise e interpretação de dados quali-quantitativos dos gêneros discursivo-textuais nos processos de Ensino, Aprendizagem e Letramento de Línguas". Como resultados são apresentados gráficos com a frequência de palavras, nuvem de palavras, bem como um dendrograma mostrando a similaridade entre textos do gênero notícias produzido a partir da matriz de ocorrência saída da ferramenta computacional. Os resultados mostraram que alguns usuários da língua portuguesa atingiram as competências desejáveis para produção textual, com escritas fortemente padronizadas em relação aos verbos dicendi que caracterizam o gênero notícia.

Palavras-chave: Dados Não Estruturados; Ferramenta Computacional, Similaridade de Corpus Textual.

Abstract: *This article presents the first version of a computational tool for analyzing unstructured data, developed in Python language and based on text mining techniques. The proposed tool has as its central focus the extraction of the frequency of words, as well as the determination of the matrix of occurrences of the term from a textual corpus. The application of the experiment started from a textual corpus of the news genre composed of 18 texts foreseen in the research project "The intermediation of Corpus Linguistics in the analysis and interpretation of qualitative-quantitative data of discursive-textual genres in the Teaching, Learning and Literacy processes Linguistic. "As a result, graphs are presented with the frequency of words, a word cloud, as well as a dendrogram that shows the similarity between the texts of the news genre produced from the output of the occurrence matrix of the computational tool. The results showed that some users of the Portuguese language achieved the desirable competencies for textual production, with strongly standardized writings in relation to the dicendi verbs that characterize the news genre.*

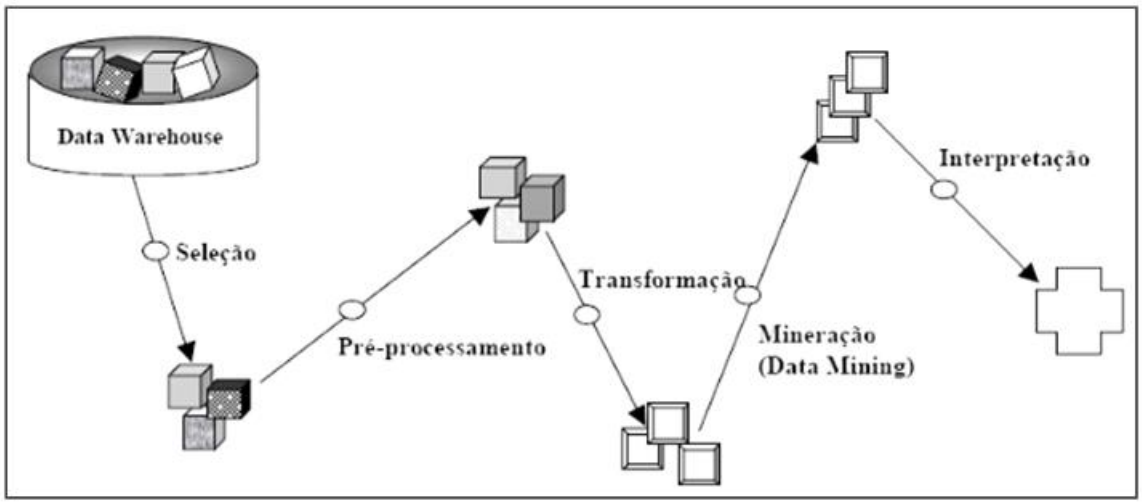
Keywords*: Unstructured Data; Computational Tool, Textual Corpus Similarity*.

1 Bacharel em Sistemas de Informação, Centro Universitário da Região da Campanha – Urcamp Bagé {paulohenriquesleite@gmail.com}
2 Doutora em Modelagem Computacional, Professora da Universidade Federal do Pampa {veraferreira@unipampa.edu.br}
3 Doutor em Modelagem Computacional, Professor da Universidade Federal do Pampa {fernandodias@unipampa.edu.br}
4 Doutora em Estudos da tradução, Professora da Universidade Federal do Pampa {adenpereira@unipampa.edu.br}

1 INTRODUÇÃO

O termo Mineração de Dados (do inglês, data mining), denominada o processo de analisar diferentes dados, consistindo em um processo de descoberta do conhecimento a partir de uma base de dados e transformá-los em informações importantes como, por exemplo, perfis de consumidores ou colaboradores que conseguem determinar vários fatores como preços e indicadores econômicos. Outro processo chamado KDD (do inglês, "*Knowledge Discovery in Databases*") é definido como busca de conhecimentos em Banco de dados, sendo relacionado à área que tem como objetivo a descoberta de informações novas dentro do contexto da análise de grandes quantidades de dados (SOUZA 2016; CUNHA, 2018). A Figura 1 abaixo, demonstra o processo de funcionamento do KDD.

Figura 1: Etapas do KDD



Fonte: BOENTE et.al, 2007.

Por sua vez, a mineração de texto contempla um conjunto de métodos para organizar, analisar e encontrar informações em corpus textual, ou seja, é um processo que tem como foco central a extração de padrões de informações em dados textuais não estruturados e semiestruturados (MAIA e SOUZA, 2010).

Nesse sentido, a ferramenta de busca de documentos é muito importante e poderosa para os pesquisadores nas áreas do conhecimento, corroborando com Scarpa (2017) quando enfatiza que existem inúmeros documentos disponíveis e encontrar aqueles que são relevantes pode contribuir para o desenvolvimento de diversos estudos. Diante disso, as aplicações da linguística computacional são de extrema importância para análise de corpus textuais com

diferentes recursos na área da linguística, bem como a evidente necessidade de softwares que permitam realizar tais análises (FIORIO *et. al*, 2019).

Segundo Da Silva e Silva (2019), a linguagem *Python* surgiu na cidade de Amsterdã, na capital da Holanda sendo um dos desenvolvedores da linguagem Guido Van Hossum que trabalhava no CWI (Instituto de Pesquisa Nacional para Matemática e Ciência da Computação) em um sistema chamado amoeba. Como esse programa apresentava várias falhas e era desenvolvido em linguagem C, Guido resolveu desenvolver outra linguagem para resolver os problemas que outra linguagem apresentava. Logo o holandês batizou a linguagem como *Python* devido ao seu programa favorito que era o *Monty Python 's Flying Circus*.

Entretanto, várias outras linguagens são utilizadas na programação, cada uma com sua contribuição à linguagem em PHP que é voltada para as aplicações web; já o *Java* é voltado para o desenvolvimento em desktop, sendo que o *Python* não possui objetivo. A linguagem oferece suporte a desktop e desenvolvimento web, aplicação mobile, geoprocessamento, processamento de *Data Science* e científico pois trabalha com grande número de informações e utiliza diversas bibliotecas. A Linguagem em *Python* é encontrada no cotidiano de muitos usuários estando presente em várias ferramentas digitais tais como nos buscadores de pesquisa como o *Google* no processamento de pesquisa de dados e em plataformas de streaming como *Youtube* e *Netflix* e entre outras grandes empresas (SILVA E SILVA, 2019).

As informações científicas estão crescendo em grande escala de maneira que novas demandas vão surgindo, essas informações são apresentadas em formatos como PDF ou HTML. Para aqueles que buscam por informações de características científicas como pesquisadores, é exigido um tempo maior no processo de leitura textual e, com isso, a necessidade de desenvolvimento de técnicas capazes de extrair conteúdo de forma rápida e objetiva (FERREIRA E CORREA, 2021)

O *software* Sobek está disponível tanto na versão on-line quanto na versão para *download*, essa ferramenta utiliza o processo de organização textual, inserção do conteúdo textual e geração do grafo dos termos mais relevantes extraídos dos documentos. A ferramenta foi desenvolvida pelo programa de Pós-graduação em informática na Educação, da Universidade Federal do Rio Grande do Sul – UFRGS, o destaque do *software* é já ter sido utilizado em textos

em português e, com isso, gerar como resultado um gráfico de mapa mental com uma representação visual do texto (MEDEIROS et. al, 2019).

A seguir, é apresentada a metodologia utilizada na elaboração e aplicação da ferramenta computacional, bem como na construção do dendrograma hierárquico. Apresenta-se também os resultados, as análises e as discussões. Para ilustrar esses resultados, são mostradas a interface de usuário da referida ferramenta, assim como as considerações sobre a importância da análise de similitude do gênero notícia.

2 METODOLOGIA

O suporte metodológico utilizado no presente trabalho iniciou com a definição dos descritores, tendo em vista a identificação, em corpus textual, o gênero notícias, a partir de verbos dicendi. Na sequência foram escolhidas uma métrica e um critério de agrupamento de pares de grupos de palavras.

2.1 A FERRAMENTA

Neste trabalho adotou-se como corpus, textos do gênero notícias, e padronização de descritores a partir da ocorrência de verbos dicendi, conforme PEREIRA (2016).

A contagem da ocorrência das palavras foi feita a partir de algoritmo elaborado em linguagem *Python*, cuja saída fornecida é a matriz de ocorrência.

2.2 A MÉTRICA E O CRITÉRIO DE AGRUPAMENTO

Para medir a similaridade foi utilizado o índice *Jaccard* (DRIEGER, 2013; MATUI, 2020), acompanhada do critério de agrupamento fornecido pelo método *average-link.*

Por se tratar de um processo dependente do conceito de proximidade (dissimilaridade ou similaridade), da estratégia de clusterização e do padrão de descritores, formar *clusters*, segundo Everitt *et al.* (2001) de dados demanda ~~os~~ realizar as seguintes tarefas e critérios:

I. Padronização do descritor pelo especialista.

II. Escolha da métrica.

III. Escolha de um critério de para agrupamento de pares de grupos:

Critério *single-link*: utiliza o critério de vizinhos mais próximos, adotando a medida de proximidade entre dois documentos.

Critério *complete-link*: adota como critério a maior distância entre par de documentos.

Critério *average-link*: adota como critério a média das distâncias dentro do grupo, considerando cada par formado por um documento de cada grupo.

Muito embora não exista um consenso sobre o conceito de *cluster*, segundo Aldenderfer; Blashfield (1984), a presença de propriedades como densidade, variância, dimensão, forma e separação, são essenciais para a criação dos diversos exemplos de *clusters*.

Para escolher os descritores representativos no conglomerado, utilizou-se a técnica de escolha das palavras identificadoras da predominância do gênero textual notícia.

Visando ressaltar a influência da similaridade na formação dos *clusters,* foi utilizada a métrica Jaccard.

O software utilizado foi *Visual Studio Community 2019, v. 16.8.2,* para o desenvolvimento do algoritmo de análise dos dados, utilizando a linguagem em *Python*. A interface *web* foi desenvolvida em linguagem PHP 7.4 e a aplicação foi realizada sob o sistema operacional *Linux Ubuntu 20.04 lts*.

2.3 O CORPUS TEXTUAL

O corpus utilizado no presente artigo constitui-se de textos do gênero notícias coletadas por Pereira (2016) que em sua tese analisou a frequência de expressões multipalavras em um córpus de mais de 1 milhão de palavras através do uso do programa *WordSmith Tools 6.0*.

Para o presente artigo, foram utilizados apenas 18 textos desse gênero a título de amostra, desta vez buscando-se os verbos dicendi que se apresentam como caraterística das notícias, já que sempre existe o repórter que procura relatar o que foi dito pelos atores envolvidos na narrativa dos fatos foco central de reportagens jornalísticas seja através de discurso direto ou indireto.

Importante destacar que a coleta deu-se através de sites de notícias veiculados na imprensa, bem como partiu de um instrumento de *Google Forms* aplicado a professores em formação e formados na área de Letras cujo principal objetivo era interpretar e, posteriormente, produzir uma notícia.

Essa foca não somente em detectar se os usuários apresentam domínio e conseguem distinguir as características do gênero notícia, bem como se apresentam competência em produzir uma notícia destacando os aspectos mais relevantes desse gênero em suas produções, já que, de acordo com Marcuschi (2003), essa é uma das competências que se espera de um usuário da língua, ou seja, o domínio de diversos gêneros textuais e suas principais características, haja vista a diversidade textual e discursiva em que os falantes da língua estão imersos diariamente.

Dessa forma, procedeu-se à coleta dos dados que foram rodados no referido programa para, posteriormente, selecionar os verbos *dicendi* de maior frequência, dando destaque na análise para os contextos linguísticos em que se apresentavam.

Em geral esses verbos funcionam como elos entre o que foi declarado ou citado pelo interlocutor com quem interage o produtor da notícia para dar maior realismo a esse tipo de texto, já que aquilo que é dito costuma dar maior veracidade ao fato narrado.

Assim, após rodar os dados no programa *WordSmith Tools 6.0,* observou-se que os verbos de maior recorrência foram os seguintes: “falou”, “relatou”, “declarou”, “disse”, “informou” e “concluiu”.

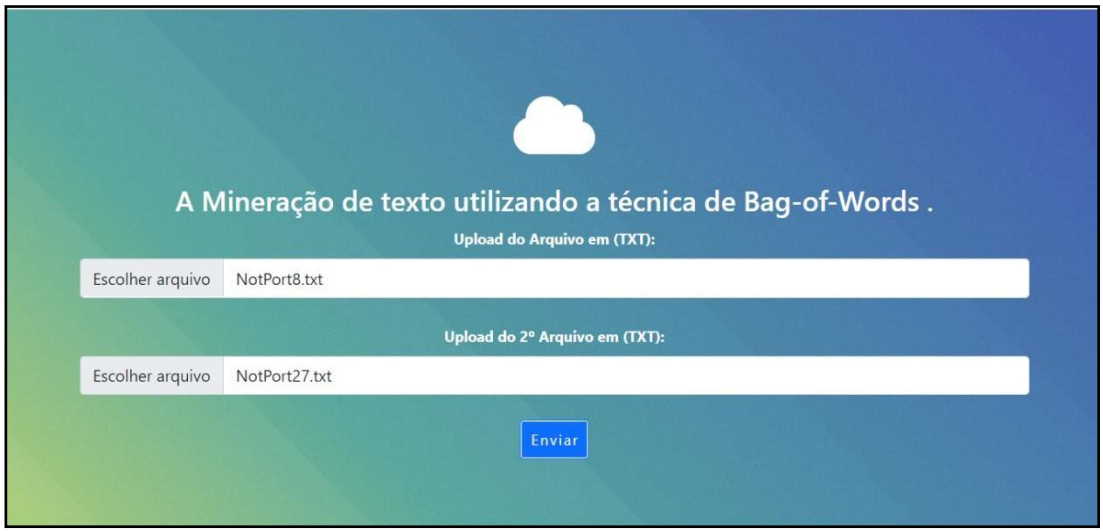
Assim, no projeto intitulado “A intermediação da Linguística de Corpus na análise e interpretação de dados quali-quantitativos dos gêneros discursivo-textuais nos processos de Ensino, Aprendizagem e Letramento de Línguas” as pesquisadoras, a partir de corpus coletado de textos dos gêneros notícias e biografias, partiu-se da variável verbos dicendi, como forma de detecção da utilização dessa característica para identificação dos gêneros em questão.

A partir de então, foi possível verificar em quais contextos esses verbos ocorriam, corroborando a hipótese de que eles poderiam ser caracterizadores do gênero notícia, partindo-se das declarações, diretas ou indiretas, como recursos utilizados pelos repórteres para levar a reportagem ao público-alvo.

3 DESENVOLVIMENTO DA PESQUISA

Como resultados, apresenta-se na Figura 4 o *layout* da interface de usuário desenvolvido como ferramenta de mineração de texto, bem como as análises realizadas nesta pesquisa piloto.

Figura 4: Página inicial da ferramenta de busca



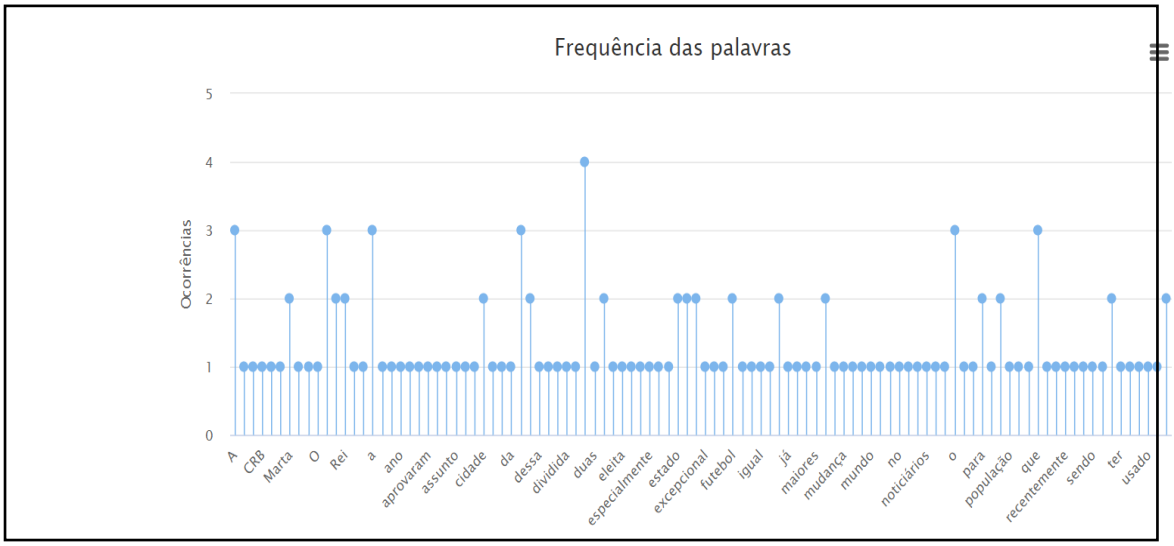
Fonte: Os autores

A Figura 4 reflete o menu principal do minerador de texto, onde é possível realizar o upload de dois arquivos por vez.

Em um estudo semelhante de Klemann et. al, 2012, descreve uma ferramenta chamada Sobek que foi desenvolvida pela Universidade Federal do Rio Grande do Sul para busca textual, esta pode ser executada em computadores com diferentes sistemas operacionais *Linux*, *Windows* ou *Mac OS*, podendo ser utilizada sem maiores restrições, contudo não está disponível online.

A Figura 5 apresenta a representação gráfica da frequência de palavras utilizando um fragmento textual do corpus organizado por Pereira (2016).

Figura 5: Representaçao gráfica da frequencia de palavras



Fonte: Os autores

Assim, a Figura 5 foi elaborada para descrever a frequência das palavras em um gráfico de linhas onde o eixo” y “ identifica as ocorrências e o eixo “x” quanto à frequência das palavras. Já na Figura 6 é apresentada a nuvem de palavras escalonadas em ordem frequência.

Dentre os assuntos que foram selecionados, as palavras que apresentaram maior destaque são “Pelé”, “Marta” e "Futebol", configurando assim uma notícia relacionada ao esporte. Um estudo de Gil (2016), semelhante e utilizando um comparativo entre ferramentas de análise de texto através da semântica das palavras usando como exemplo o software TagCrowd com o mesmo objetivo de gerar nuvem de palavras e destacando-as.

Figura 6: Nuvem de palavras por frequência

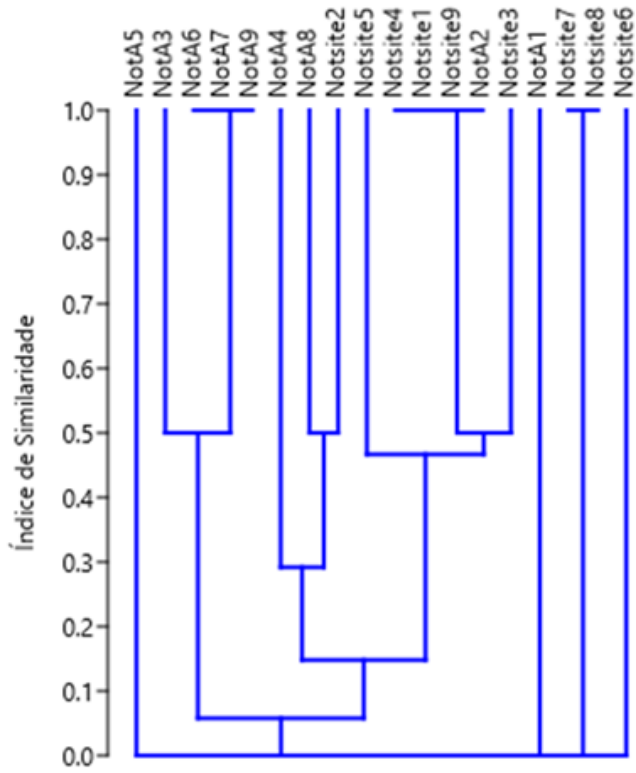
Fonte: Os autores

4 RESULTADOS E DISCUSSÃO

No intuito de verificar a similaridade entre as notícias utilizadas por Pereira (2016) e as notícias produzidas por alunos, foi realizada a análise de 18 textos desse gênero, sendo 9 notícias de sites e 9 notícias dos alunos, respectivamente, denotadas por Notsite1, …, Notsite9, e NotA1, …, NotA9. via agrupamento hierárquico. Vale ressaltar que, o agrupamento foi obtido via o *software* de código aberto PAST[5] tomando como entrada a matriz de ocorrência dos verbos dicendi "falar", Relatar, Declarar, Dizer, Informar, Concluir, Encontrar, Afirmar, Comentar".

[5] https://www.nhm.uio.no/english/research/infrastructure/past/

Figura 7: Dendrograma de Similaridade das Notícias



Fonte: Os autores

No dendrograma apresentado na Figura 7, as folhas (as notícias de sites e produzidas por alunos) retratam os elementos completamente isolados. Percebe-se que as notícias que menos se assemelham NotA1 e Notsite6 e (índice Jaccard 0,4), bem como as que apresentam maior similaridade Notsite 1, Notsite 4, Notsite 9 com a produção textual do aluno NotA2 e Notsite 2 e NotA8 (índice Jaccard 1,0). Corroborando com a abordagem gráfica hierárquica. Drieger (2013) enfatiza a relevância de analisar visual de texto. A formação de grupos no dendrograma mostrou a familiaridade entre a produção textual dos discentes e a produção textual do gênero notícia coletadas em sites pelos profissionais da área de jornalismo. Vale ressaltar que três grupos foram bem definidos destaque de agrupamentos (NotA6, NotA7, NotA9). Em síntese, as similaridades mostradas no agrupamento hierárquico apontaram que os usuários da língua NotA2, NotA8 atingiram o desenvolvimento de competências desejáveis para caracterização e produção textual do gênero notícia. Abdullah, Ali e Makttof (2019) destacam a importância da avaliação do

grau de similaridade entre textos, bem como o potencial desafiador na elaboração de um esquema lógico de cálculo para análise lexical na área de processamento de linguagem natural.

5 CONSIDERAÇÕES FINAIS

Em contextos educacionais nos quais a qualidade da produção textual é um indicativo da aprendizagem, faz-se necessário avaliar a similaridade entre a produção escrita e alguma tomada como padrão. Para tal propósito, tem se tornado habitual o uso de ferramentas de mineração de textos.

Neste trabalho, foi apresentada uma interface gráfica para mineração de textos, tomando-se por base textos do gênero notícias disponibilizadas em sites, bem como textos produzidos por alunos conforme pesquisa de Pereira (2016; 2021).

A interface gráfica produz como saídas uma matriz de ocorrências e a representação em nuvem de palavras. Além disso, possui as características de estar disponibilizada em modo *online*, ser utilizável em qualquer plataforma e apresentar compatibilidade com qualquer sistema operacional.

Por se tratar de uma fase de protótipo, estudos comparativos de desempenho relativos a outros algoritmos serão objeto de estudos futuros.

6 REFERÊNCIAS

ABDULLAH, S. M ; ALI, S.M, MAKTTOF, A.B. **Modifying Jaccard Coeffcient for Texts Similarity**. Revista de Ciências Humanas y Sociales, Año 35, Nº Especial 19 (2019):2899-2921p.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis**. Beverly Hills, CA: Sage, 1984. 88 p.

BOENTE, Alfredo Nazareno Pereira; ROSA, José Luiz Dos Anjos. Utilização de Ferramentas de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa. **Seget, http://www. aedb. br/seget/artigos07/1219_Artigo% 20SEGET**, v. 202007, 2007.

DA SILVA, Rogério Oliveira; SILVA, Igor Rodrigues Sousa. Linguagem de Programação Python. **TECNOLOGIAS EM PROJEÇÃO**, v. 10, n. 1, p. 55-71, 2019.

DE MEDEIROS, Wagner Oliveira; PINHO, Fabio Assis; CORREA, Renato Fernandes. APLICAÇÃO DE SOFTWARE DE MINERAÇÃO DE TEXTO NA REPRESENTAÇÃO DA INFORMAÇÃO DE OBRAS ARTÍSTICO-PICTÓRICAS. **Logeion: Filosofia da Informação**, v. 6, n. 1, p. 146-170, 2019.

DRIEGER, P. Semantic Network Analysis as a Method for Visual Text Analytics. Procedia - Social and Behavioral Sciences, v. 79, p. 04-17, 2013.

EVERITT, B.; LANDAU, S.; LEESE, M. Cluster Analysis. A Hodder Arnold Publication. **Willey, London**, 2001.

FERREIRA, Márcio Henrique Wanderley; CORREA, Renato Fernandes. Mineração de textos científicos: análise de artigos de periódicos científicos brasileiros da área de Ciência da Informação. **Em Questão**, v. 27, n. 1, p. 237-262, 2021.

FIORIO, Rosaine et al. Linguisticun: Uma Ferramenta de Auxílio ao Ensino da Língua Portuguesa e à Linguística Computacional. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2019. p. 11.

GIL, Carmem Zeli Vargas; SEFFNER, Fernando. Dois monólogos não fazem um diálogo: jovens e ensino médio. **Educação & Realidade**, v. 41, p. 175-192, 2016.

KLEMANN, Miriam; REATEGUI, Eliseo; RAPKIEWICZ, Clevi. Análise de ferramentas de mineração de textos para apoio a produção textual. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2012.

MAIA, Luiz Cláudio; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da informação**, v. 15, p. 154-172, 2010.

MARCUSCHI, Luiz Antônio. Gêneros textuais: definição e funcionalidade. In: **Gêneros textuais e ensino**. 2. ed. Ângela Paiva Dionísio, Ana Rachel Machado, Maria Auxiliadora Bezerra (Orgs). São Paulo: Parábola Editorial, 2003.

MATUI, Natália da Conceição. Mapeamento semântico do conceito de inovação para a ciência da informação: uma análise gramático-sistêmico funcional. 2020.. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de São Carlos, 2020.

PEREIRA, Aden R. Análise contrastiva de verbos dicendi em textos jornalísticos de corpus paralelo português-espanhol à luz da Linguística de Corpus. In: NADIN, Odair FERREIRA, Anise A. G. D.; FARGETI, Cristina M. (orgs.) Léxico e suas interfaces: descrição, reflexão e ensino. São Paulo/: Cultura Acadêmica, 2016. pp. 177-197.

_____; JURGINA, Daniele. A intermediação da Linguística de Corpus na análise e interpretação de dados quali-quantitativos dos gêneros discursivo-textuais nos processos de Ensino, Aprendizagem e Letramento de Línguas. Revista EALQ 2021.(no prelo)

SCARPA, Alice Duarte. **Técnicas de processamento de linguagem natural aplicadas às Ciências Sociais**. 2017. Tese de Doutorado.

SOUZA, Adriano; FORTES, Reinaldo; LIMA, Joubert. OLAP Textual com Múltiplas Hierarquias de Tópicos e Rankings Segmentados. In: **Anais do XIII Simpósio Brasileiro de Sistemas de Informação**. SBC, 2017. p. 480-487.