UNIVERSIDADE DE BRASÍLIA (UnB)
FACULDADE DE CIÊNCIA DA INFORMAÇÃO (FCI)
PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (PPGCINF)

NATHALY CRISTINE LEITE ROCHA

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS DIGITAIS: UMA PROPOSTA A PARTIR DE MARCAÇÕES DE LEITORES

Brasília
2023

NATHALY CRISTINE LEITE ROCHA

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS DIGITAIS: UMA PROPOSTA A PARTIR DE MARCAÇÕES DE LEITORES

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação (PPGCINF), da Faculdade de Ciência da Informação (FCI), da Universidade de Brasília (UnB) como requisito para obtenção do título de Mestra em Ciência da Informação.

Área de concentração: Gestão da Informação

Linha de pesquisa: Comunicação e mediação da informação

Orientador: Prof. Dr. Dalton Lopes Martins

Brasília
2023

Dados Internacionais de Catalogação na Publicação (CIP)

Rocha, Nathaly Cristine Leite
R672i Indexação automática de documentos digitais: uma proposta a partir de marcações de leitores / Nathaly Cristine Leite Rocha; orientador Dalton Lopes Martins. -- Brasília, 2023.
99 p.: il. Color.

Dissertação(Mestrado em Ciência da Informação) -- Universidade de Brasília, 2023.

1. indexação automática. 2. destaques em texto. 3. marcações em texto. I. Lopes Martins, Dalton , orient. II. Título.

025.4(0.034.1)

UNIVERSIDADE DE BRASÍLIA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Ata Nº: 46

Aos vinte e um dias do mês de dezembro do ano de dois mil e vinte e três, instalou-se a banca examinadora de Dissertação de Mestrado da aluna **Nathaly Cristine Leite Rocha**, matrícula 21/0007206. A banca examinadora foi composta pelos professores Dra. Cynthia Roncaglio / membro interno / PPGCINF/UnB, Dr. Tiago Emmanuel Nunes Braga / Membro externo / IBICT, Dr. João de Melo Maricato / PPGCINF/UnB, Suplente e Dr. Dalton Lopes Martins / orientador/presidente / PPGCINF/UnB. A discente apresentou o trabalho intitulado "Indexação automática de documentos digitais: uma proposta a partir de marcações de leitores ".

Concluída a exposição, procedeu-se a arguição do(a) candidato(a), e após as considerações dos examinadores o resultado da avaliação do trabalho foi:

( ) Pela aprovação do trabalho;

(X) Pela aprovação do trabalho, com revisão de forma, indicando o prazo de até 30 dias para apresentação definitiva do trabalho revisado;

( ) Pela reformulação do trabalho, indicando o prazo de **(Nº DE MESES)** para nova versão;

( ) Pela reprovação do trabalho, conforme as normas vigentes na Universidade de Brasília.

Conforme os Artigos 34, 39 e 40 da Resolução 0080/2021 - CEPE, o(a) candidato(a) não terá o título se não cumprir as exigências acima.

Dr. Dalton Lopes Martins, PPGCINF/UnB

(Presidente/orientador)

Dra. Cynthia Roncaglio, PPGCINF/UnB

(Membro interno)

Dr. Tiago Emmanuel Nunes Braga, IBICT

(Membro externo)

Dr. João de Melo Maricato, PPGCINF/UnB

(Suplente)

Nathaly Cristine Leite Rocha

(Mestranda)

Documento assinado eletronicamente por **Dalton Lopes Martins**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 16/01/2024, às 15:55, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Cynthia Roncaglio**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 17/01/2024, às 11:36, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Tiago Emmanuel Nunes Braga**, **Usuário Externo**, em 24/01/2024, às 05:38, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Nathaly Cristine Leite Rocha**, **Usuário Externo**, em 27/01/2024, às 11:11, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Clovis Carvalho Britto**, **Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 31/01/2024, às 10:40, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **10622221** e o código CRC **FDD783D5**.

Dedico este trabalho a todos aqueles que se fizeram presentes em minha vida durante os tempos de desenvolvimento da pesquisa. Vocês estão em meu coração e em cada frase desta dissertação.

AGRADECIMENTOS

Carrego comigo os sonhos de todas as mulheres que vieram antes de mim. Em tudo que faço e em tudo que sou, devo ser grata àquelas que me abriram espaços para escolhas. Não poderia agradecer outro alguém primeiramente senão as três mulheres mais importantes da minha vida: Glória, minha avó, Valquíria, minha mãe, Nayara, minha irmã. Vocês são minha inspiração para tudo que existe nessa vida, me espelho e honro vocês em todos os meus caminhos.

Guilherme, Catarina e Clarice, agradeço pela presença curiosa e animada nos meus dias. Conviver com crianças tão amáveis e inventivas como vocês me mantém curiosa e esperançosa em um mundo que por vezes é desestimulante.

Resgato agora palavras de Saramago, que quando foi questionado se seria capaz de viver sem sua Pilar responde que sim, evidentemente, mas que não saberia como. Uso estas para agradecer ao Lucas por trazer tanto amor à minha vida, por acreditar em mim e por me fazer acreditar também. Sem você eu viveria, mas, sinceramente, não sei como.

Aos amigos que caminharam comigo durante todo esse percurso da pós-graduação, eu não seria capaz de entregar resultados sem o apoio de vocês. Sou grata por ter tantos para incluir aqui em meus agradecimentos:

À Larissa, minha grande guia nesse trajeto, sempre atenta para me ouvir e sempre competente e amorosa para aconselhar e apoiar. Obrigada pela nossa tão frutífera amizade e parceria;

À Jessica, que me acompanha em momentos de descontração e também de angústias rotineiramente há alguns anos, obrigada por ser essa companhia que me faz tão bem. Você é um exemplo de coisas boas para mim;

À Denise, que viveu comigo intensamente toda a pós-graduação, você é um grande presente que vou levar dessa fase da vida. Obrigada por dividir comigo os mesmos neurônios;

Ao Carlos Henrique, que desde a graduação embarca comigo nas aventuras do desconhecido e sempre me apoia com muita paciência, contribuindo com conhecimentos que eu não conseguiria adquirir sozinha. Você é parte importante deste trabalho, obrigada por tudo;

À toda a equipe da COTEC: os queridos Milton, Ingrid, Diego, Fernanda e Fernando, agradeço pela companhia em tantos almoços divertidos, com sobremesa e café para dar conta da rotina de pesquisa tão intensa dos últimos anos;

As minhas amigas virtuais Filhas de Vênus, cada uma em um canto do país, por todo apoio diário e por vibrarem comigo a cada avanço que dei nesta empreitada;

Ao Dr. Tiago Braga, grande incentivador não só deste trabalho, mas também de toda minha caminhada profissional e acadêmica. Sou grata pela confiança e parceria de sempre;

Ao Bruce, meu gato, companheiro de todos os momentos de escrita e leituras.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pela concessão da bolsa de estudos que viabilizou o desenvolvimento desta pesquisa e à Universidade de Brasília (UnB), em especial, a todos da Faculdade de Ciência da Informação (FCI), meu lugar durante a graduação e mestrado;

Ao meu orientador, Prof. Dalton Lopes Martins, agradeço a atenção e acompanhamento durante a pesquisa;

Ao Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), agradeço por ser um espaço que me incentiva e me apresenta tantos caminhos e oportunidades para seguir estudando e descobrindo coisas novas.

Por fim, agradeço aos profissionais de saúde mental que me acompanharam ao longo do mestrado. Vocês foram essenciais para o meu desenvolvimento como pessoa e para o desenvolvimento deste trabalho.

"*Al fin y al cabo, somos lo que hacemos para cambiar lo que somos.*"

(Eduardo Galeano, *El Libro de los Abrazos*)

RESUMO

Lidar com a representação, organização e recuperação da informação em contexto digital representa um desafio ao mesmo tempo que possibilita a exploração de diferentes modos de atender as necessidades informacionais tão intrínsecas a todos os indivíduos em suas atividades cotidianas. A interação de leitores com textos transformou-se significativamente na era digital. A marcação de textos em meio digital tornou-se uma prática comum, permitindo aos leitores destacar trechos relevantes, fazer anotações e criar marcadores virtuais. Ferramentas como destaque de texto, sublinhado e anotações digitais proporcionam uma experiência de leitura interativa e personalizada. Além disso, a marcação em meio digital facilita a organização e revisão posterior, contribuindo para uma compreensão mais aprofundada dos conteúdos em um ambiente dinâmico e tecnologicamente avançado, também se caracterizando como potencial fonte de registros para organização e recuperação da informação, em específico no escopo deste trabalho, para indexação. Isto posto, o objetivo deste estudo é investigar de que maneira as anotações e outros registros feitos por usuários/leitores em documentos digitais podem ser usados para indexação automática de documentos digitais. Com a proposta de criar um fluxo de trabalho para indexação automática a partir de trechos grifados, o estudo se vale de métodos mistos utilizando aspectos tanto qualitativos como quantitativos para atender os objetivos geral e específicos. A coleta de dados se deu por questionário direcionado a pesquisadores da Ciência da Informação, criando um corpus de textos para analisar. Aplicou-se códigos computacionais escritos com a linguagem Python e o apoio das bibliotecas PyMuPDF, SciKit Learn e *Natural Language Toolkit* (NLTK) para extração de trechos, pré-processamento de dados e cálculos de frequência para determinação de termos indexadores. Como resultados, apresenta-se uma análise das estratégias de marcações dos respondentes da pesquisa aproximando-as de conceitos da Organização da Informação, mostrando convergências entre ambas. O processo de indexação apresentado como proposta foi considerado satisfatório no objetivo de gerar um conjunto de termos indexadores para o documento do *corpus*. Sendo assim, foi disponibilizado tanto o fluxo de trabalho como os códigos utilizados no processo.

Palavras chave: indexação automática; destaques em texto; marcações em texto; documentos digitais.

ABSTRACT

Dealing with the representation, organization, and retrieval of information in a digital context poses a challenge while enabling the exploration of different ways to meet the information needs so intrinsic to individuals in their daily activities. The interaction of readers with texts has undergone significant transformations in the digital age. Text markup in digital environments has become a common practice, allowing readers to highlight relevant passages, make annotations, and create virtual bookmarks. Tools such as text highlighting, underlining, and digital annotations provide an interactive and personalized reading experience. Furthermore, digital text markup facilitates organization and subsequent review, contributing to a deeper understanding of content in a dynamic and technologically advanced environment, also serving as a potential source of records for information organization and retrieval, specifically within the scope of this work, for indexing. With that said, the objective of this study is to investigate how annotations and other user/reader records in digital documents can be used for the automatic indexing of digital documents. Proposing to create a workflow for automatic indexing from highlighted passages, the study employs mixed methods using both qualitative and quantitative aspects to address the general and specific objectives. Data collection was done through a form directed at Information Science researchers, creating a corpus of texts for analysis. Computational codes were applied using the Python language and the support of the PyMuPDF, SciKit Learn, and Natural Language Toolkit (NLTK) libraries for extracting passages, data preprocessing, and frequency calculations to determine indexing terms. As results, an analysis of the marking strategies of the research respondents is presented, aligning them with concepts of Information Organization and demonstrating convergences between the two. The proposed indexing process was considered satisfactory in generating a set of indexing terms for the corpus document. Therefore, both the workflow and the codes used in the process have been made available.

Keywords: automatic indexing; text highlights; text markings; digital documents.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BERT - *Bidirectional Encoder Representations from Transformers*

CLIP - *Contrastive Language-Image Pre-training*

CI - Ciência da Informação

CNN - Redes Neurais Convolucionais (*Convolutional Neural Network*s)

COVID-19 - *Corona Virus Disease* 2019

DOCx - Abreviação de *document* (formato de arquivo)

DQN - *Deep Q-networks*

GPT - *Generative Pre-trained Transformer*

kNN - *K-nearest neighbor*

LDA - *Latent Dirichlet Allocation*

LSTM - *Long Short-Term Memory*

MAUI - acrônimo para *Multi-purpose Automatic Topic Indexing.*

ML - *Machine Learning*

MMT - *Multimodal Transformer*

NB - Naive Bayes

NLTK - *Natural Language Toolkit*

NMF - *Non-Negative Matrix Factorization*

OI - Organização da Informação

PDF - *Portable Document Format*

RNN - Redes Neurais Recorrentes (*Recurrent Neural Networks*)

SISA - Sistema de Indexação Semi Automática

SMS - *Short Message Service*

SVM - *Support Vector Machine*

TICs - Tecnologias de Informação e Comunicação

UNESCO - *United Nations Educational, Scientific and Cultural Organization*

SUMÁRIO

1 INTRODUÇÃO

A relação da sociedade com a produção e consumo de informações vem se alterando ao longo do tempo. Foi convencionado chamar a configuração dessa relação atualmente de sociedade da informação, a qual segundo Gouveia e Gaio (2004) é uma sociedade que predominantemente utiliza as tecnologias de informação e comunicação para a troca de informações em formato digital e que suporta a interação entre indivíduos e organizações com recurso a práticas e métodos em construção permanente. A internet ganha papel de destaque nas relações informacionais e nos processos de armazenamento, processamento e comunicação da informação, visto que "indivíduos e organizações podem produzir e consumir informação de um modo quase instantâneo e a qualquer hora e em qualquer lugar" (Silva; Gouveia, 2020).

A internet e a *web* mundial têm fundamental importância nessa sociedade da informação, já que permitem a conexão de bilhões de pessoas em todo o mundo por meio de dispositivos tecnológicos. A pandemia da Covid-19, que iniciou em 2020, intensificou o uso da internet como ferramenta para estudo, trabalho, registro e utilização de informações. Atualmente, o mundo conta com 5,3 bilhões de usuários da rede mundial de computadores, segundo pesquisa publicada pela ONU (2022)[1]. Além da rede mundial, também os *gadgets* de acesso à informação mudaram o paradigma da leitura em papel para o digital. No Brasil, a Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros - TIC Domicílios (2020) mostra que o acesso a tablets com conexão à internet compreende 8% da população, e os que acessam por notebook representam 30%. Outra pesquisa, esta do *Pew Research Center* (2022), mostra que apesar dos livros físicos continuarem sendo a preferência da maioria dos leitores adultos, o percentual daqueles que leem em formato digital já é de 30%.

No âmbito da literatura científica também é perceptível o crescimento da quantidade de documentos disponíveis em formato digital. A comunidade acadêmica pode acessar milhares de publicações, livros e periódicos indexados em bases de dados diversas. Para que esse acesso aconteça, destaca-se o trabalho de indexação, apontado por Gil Leiva (2009) como primordial para a recuperação da informação científica. Com a crescente produção de conhecimento atual, torna-se necessário o uso de ferramentas tecnológicas que proporcionem mais agilidade aos processos de documentação. Se tratando da indexação, Lancaster (2004) e Moreiro González (2004) apontam que as automatizações das tarefas envolvidas corroboram para um importante ganho de tempo dos profissionais indexadores que atuam em unidades de informação.

[1] Disponível em: https://news.un.org/pt/story/2022/09/1801381#:~:text=Ao%20todo%2C%20existem%205%2C3,da%20pandemia%20de%20Covid%2D19.

A influência do meio digital é percebida não só pelo acesso rápido e facilitado aos documentos em bases de dados, como também é notória a mudança na forma de interação dos usuários da informação com os textos neste formato. A leitura de documentos digitais se dá por meio de telas de computador ou outros dispositivos como tablets, celulares smartphone e e-readers, que são aparelhos eletrônicos desenvolvidos especificamente para leitura de documentos em formato digital. Esses dispositivos apresentam funções de interação que antes não eram experimentadas em mídias analógicas, a exemplo do uso de *links* e *hiperlinks*, o que permite um enriquecimento informacional na medida em que complementa o texto com outros textos que estão acessíveis ao leitor com um clique na tela do dispositivo utilizado.

Outra forma de interação do leitor com os documentos digitais é o manuseio da informação sem que se prejudique a integridade do texto: em um documento lido em uma tela, é possível riscar, grifar, anotar, marcar páginas sem que se perca o acesso ao texto no formato original disponibilizado. Essa possibilidade de interação pode ser investigada por diversos aspectos, no presente estudo considera-se um aspecto relevante para a indexação, mencionado por Glushko (2016). Ao elencar elementos que podem ser considerados em um processo de organização da informação, o autor cita os destaques feitos por um leitor ao grifar ou destacar[2] um texto como um recurso organizacional, chamado pelo autor de recurso de interação.

É sabido que antes das mídias de leitura digital os leitores também produziam grifos nos textos e interagiam de variadas formas com a leitura, entretanto, o meio digital proporciona um aspecto novo sobre essas interações: o registro e armazenamento das anotações e destaques em formatos que possibilitam a exportação e utilização em outros ambientes para além do documento lido. É possível que o leitor utilize o *e-reader Kindle*, da empresa *Amazon*, e realize anotações e destaques, por exemplo, que mais tarde poderão ser acessados em seu computador ou celular. A leitura se expande para além do processo de interpretação de códigos, mas também se apresenta parte de um processo em que o leitor gera novas informações também em meio digital.

Além dos documentos em meio digital, outra grande fonte informacional nativa do meio digital e, sobretudo, propiciada pela internet, são os dados registrados em grande quantidade a todo momento. Podemos citar alguns exemplos mensuráveis: existem cerca de 1 trilhão de páginas da web; uma hora de vídeo é carregada no *YouTube* a cada segundo, totalizando 10 anos de conteúdo todos os dias; os genomas de milhares de pessoas, cada um com um comprimento de 3,8 × 109 pares de bases, foram sequenciados por vários laboratórios; o

[2] No texto original o autor utiliza o termo highlight, em tradução nossa entendemos equivalência semântica em português com os termos grifar ou destacar.

Walmart processa mais de 1 milhão de transações por hora e possui bancos de dados com mais de 2,5 *petabytes* (2,5 × 10^*15*) de informações (Cukier, 2010). Esses dados podem ser trabalhados e manipulados para diversos fins e de diversas maneiras, e para elucidar a importância deste tipo de registro é válido mencionar a pirâmide de produção de conhecimento, proposta por Ackoff em 1980, que mostra os dados na base da hierarquia dados-informação-conhecimento-sabedoria, reforçando que a partir dos dados é possível produzir conhecimento.

A quantidade de dados produzidos pela humanidade impôs a necessidade de desenvolvimento de métodos e ferramentas para processamento e utilização da informação contida nas imensas bases de dados disponíveis. É junto com a *big data*, definidoa por De Mauro, Greco e Grimaldi (2016) como a conjuntura informacional caracterizada por altos volumes, velocidade e variedade de dados, que se nota o desenvolvimento de técnicas computacionais para o aproveitamento e exploração do potencial informacional presente nos dados disponíveis atualmente. Neste contexto está a aprendizagem de máquina, definida por Samuel (1959) como um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para tal.

A aprendizagem de máquina, campo que está inserido na grande área de estudo de inteligência artificial, tem promovido a automatização e melhoramento de inúmeros processos na vida cotidiana, na pesquisa e na execução de tarefas diversas. Também está inserido neste contexto o *deep learning*, campo de estudo onde o objetivo é aproximar as ações e análises de máquinas ao pensamento humano. Estas três áreas – aprendizagem de máquina, inteligência artificial e *deep learning* – caminham juntas e por muitas vezes se mostram convergentes em estudos e pesquisas, criando conceitualmente uma esfera onde no centro está o *deep learning*, envolvido pela camada da aprendizagem de máquina que por sua vez está envolvida pela camada da inteligência artificial (Damasceno; Vasconcelos, 2018).

O cenário de intensa transformação digital supramencionado apresenta impactos também nas epistemologias e campos de pesquisa. Nesse contexto, emerge a área denominada humanidades digitais: uma área transversal às humanidades tradicionais e à ciência da computação. Cambridge aponta que as humanidades digitais fazem uso de ferramentas e métodos inovadores para investigar formas tradicionais e novas de dados e mídia, além de buscar oportunidades para interrogar e refletir sobre o conhecimento e a percepção que o 'digital' oferece. Em contextos informacionais, Pimenta (2020) destaca o interesse da Ciência da Informação nos estudos das humanidades digitais:

> Embora já presente no cenário acadêmico brasileiro desde o início dos anos 2000, as humanidades digitais – um campo de pesquisa transdisciplinar onde questões e objetos ligados às diversas disciplinas das ciências humanas,

> sociais e sociais aplicadas se encontram com recursos oriundos da computação, ocasionando a possibilidade de novos desdobramentos da produção do conhecimento nas humanidades no ambiente digital – têm chamado a atenção de um público crescente da ciência da informação nos últimos cinco anos. (Pimenta, 2020).

O que se percebe é que impulsionados pela estreita relação da informação com os dispositivos de Tecnologia da Informação e da Comunicação (TICs) e com a proximidade da produção de informações com a produção de dados, os estudos de aprendizagem de máquina no contexto da CI são observados desde a década de 1980, com crescimento notável no número de projetos nessa área de estudo na última década, também refletidos na ascensão do número de estudos no âmbito das humanidades digitais, mencionado acima.

Por fim, as novas perspectivas de estudos no âmbito da organização da informação possibilitadas pela digitalização de informações e documentos são reforçadas por Glushko (2016) que aponta que cada registro de uma escolha do usuário em acessar, navegar, comprar, destacar, vincular informações e outras interações torna-se então um "recurso de interação" que pode ser analisado e utilizado para reorganizar ou redesenhar um sistema de informação.

1.1 Problema de pesquisa

Em face do exposto, torna-se possível suscitar o questionamento norteador desta pesquisa: *De que maneira as anotações e outros registros feitos por usuários/leitores em documentos digitais podem ser usados para indexação automática de documentos digitais?*

1.1 Objetivo

A seguir encontram-se elencados os objetivos geral e específicos elaborados para a presente pesquisa, de forma a orientar o desenvolvimento do estudo:

1.1.1 Objetivo geral

Propor um modelo de fluxo de trabalho (*workflow*) para indexação automática de documentos a partir dos termos extraídos de marcações em documentos digitais.

1.1.2 Objetivos específicos

a) Compreender a relação entre as anotações em documentos digitais e a organização da informação por meio da análise de marcações de leitores;
b) Realizar a indexação automática do corpus;
c) Propor um workflow de indexação de termos extraídos de anotações em documentos digitais.

1.2 Justificativa

O estudo proposto se insere no campo da organização da informação, importante área de estudo no contexto da Ciência da Informação, e mais especificamente na dimensão temática da organização da informação, que segundo Café e Sales (2010), se preocupa com os conteúdos informacionais, dimensão esta notada nas práticas de catalogação de assuntos, classificação, indexação e análise documental. Quanto ao objetivo da organização da informação, temos posto, de acordo com Brascher e Café, que:

> O objetivo do processo de organização da informação é possibilitar o acesso ao conhecimento contido na informação. Para que esse objetivo seja alcançado faz-se necessário descrever fisicamente e tematicamente os objetos informacionais (textos, imagens, registros sonoros, páginas da web, entre outros). (Brascher; Café, 2008, p. 5).

Ancorado também na importância da automatização dos processos documentais, devido ao grande número de documentos existentes, o trabalho proposto busca investigar maneiras para que esses processos sejam executados de forma eficiente com base nos princípios informacionais já estabelecidos e conhecidos por meio da ampla literatura da Ciência da Informação, para que possa complementar o trabalho intelectual humano empreendido em tais processos, corroborando para a criação de ferramentas informacionais que contribuam com os objetivos gerais da organização da informação e também da CI, de modo mais amplo.

Na medida em que a tecnologia dispõe de ferramentas para a automatização de alguns processos realizados por humanos, entender como o processo de indexação pode ser realizado com o aparato de ferramentas computacionais pode nos aproximar de diferentes perspectivas daquelas enfrentadas por indexadores humanos, como aponta Mai (2001):

> Seria quase impossível, naturalmente, para qualquer pessoa ou, neste caso, qualquer indexador, precisar todas as ideias e significados que estivessem associados a qualquer documento, posto que sempre haverá ideias e significados potenciais que diferentes pessoas em diferentes momentos e lugares poderão descobrir nesse documento. Além do que, seria quase impossível prever com exatidão quais das inúmeras ideias e significados que estivessem associados ao documento seriam especificamente úteis para os usuários ou dariam ao documento alguma utilidade duradoura. É da máxima importância reconhecer e aceitar essa indefinição fundamental. O indexador deve compreender, desde o início, que jamais descobrirá todas as ideias e

> significados que estariam associados ao documento e que, portanto, não é possível descrever todas essas ideias e significados. (Mai, 2001).

Lancaster (2004) aponta a importância de levar em consideração o interesse do usuário ao realizar a indexação temática, visto que não existe um conjunto correto de termos pré-determinados para indexar um documento, bem como diz que um mesmo documento poderá ser indexado de diversas maneiras a depender do usuário e/ou grupo de usuários aos quais se destina o documento indexado. Também cabe mencionar Hjørland (2008) que afirma que a indexação automática é realizada por meio de procedimentos algorítmicos, o que corrobora com a temática proposta neste trabalho.

Sendo assim, aponta-se que a proposta de investigação sobre indexação automática a partir de trechos grifados pelos leitores se alinha a uma demanda crescente por soluções que possam lidar com a sobrecarga de dados e documentos no âmbito digital, visto que a abordagem estudada busca não apenas facilitar a recuperação da informação, mas também personalizar a experiência do usuário durante o processo. Por fim, destaca-se que a pesquisa também representa uma oportunidade valiosa para avançar o conhecimento na interseção entre Ciência da Informação e Tecnologia da Informação, em especial, nas aplicações de automação de tarefas e aprendizagem de máquina no âmbito da recuperação de informações digitais.

2 REFERENCIAL TEÓRICO

Esta seção aborda os aspectos e temas fundamentais para o embasamento teórico da pesquisa, a começar por documentos digitais, abordando as definições conceituais, especificidades e os impactos do formato digital nos processos de documentação da informação. Em seguida serão apresentados tópicos referentes à organização da informação, em especial, sobre a indexação em seus variados tipos. A última parte do referencial se volta para a aprendizagem de máquina e suas aplicações na organização da informação. Para a construção deste referencial teórico foi utilizada a revisão de literatura caracterizada, segundo Sampieri, Collado e Lucio (2013), como revisão narrativa, visto que foi realizada de forma indutiva, a partir da busca de termos relacionados às temáticas do trabalho em bases de indexação. Os textos utilizados foram recuperados principalmente por meio da busca na base de periódicos Capes, Brapci, Scielo e *Web of Science*, além da busca direta por trabalhos citados nas leituras realizadas.

2.1 Documentos digitais

Assim como a busca pelo entendimento sobre o que é um documento possibilitou rica discussão no campo da Ciência da Informação (CI), com as primeiras descrições sobre o assunto observadas nos apontamentos de Paul Otlet em seu *Traité de documentation* (1934), a busca pela compreensão sobre o que é um documento digital vem possibilitando novas reflexões. É necessário que se entenda os aspectos de um documento agora em um novo meio de circulação de informações que possui tecnologias e fluxos característicos dos tempos atuais.

Partir da definição do que é um documento em suporte físico é fundamental para a construção do entendimento sobre o documento digital. Para alguns autores, como Pédauque (2003), a noção de documento deve permanecer a mesma, independente do suporte onde a informação está registrada, visto que o autor considera o documento um produto das dimensões antropológicas, intelectuais e sociais do ser humano. Nesta perspectiva, mesmo que o meio seja diferente, o produto continua o mesmo.

A autora Blanca Rodríguez Bravo em seu livro *El documento: entre la tradición y la renovación* (2002) também defende que o documento digital continua sendo documento, com a especificidade de que a união da mensagem com o suporte não é inseparável, como acontece no documento analógico.

Contudo, é necessário que se estabeleça as características específicas do documento produzido e utilizado em meio digital, já que as características de um suporte impactam nos processos que garantem a circulação, recuperação, utilização e preservação de um documento. Se é preciso tratar algo, é necessário entender o que é algo. Os esforços empenhados por Paul Otlet e Henri La Fontaine para definir o que é um documento estão intimamente ligados ao desejo de documentar os objetos informativos com os quais lidavam, com objetivo de possibilitar a recuperação da informação por quem desejasse fazê-la. Nesse sentido, visto que os meios pelos quais as informações são recuperadas atualmente, é imprescindível que se entenda também como documentar em meio digital.

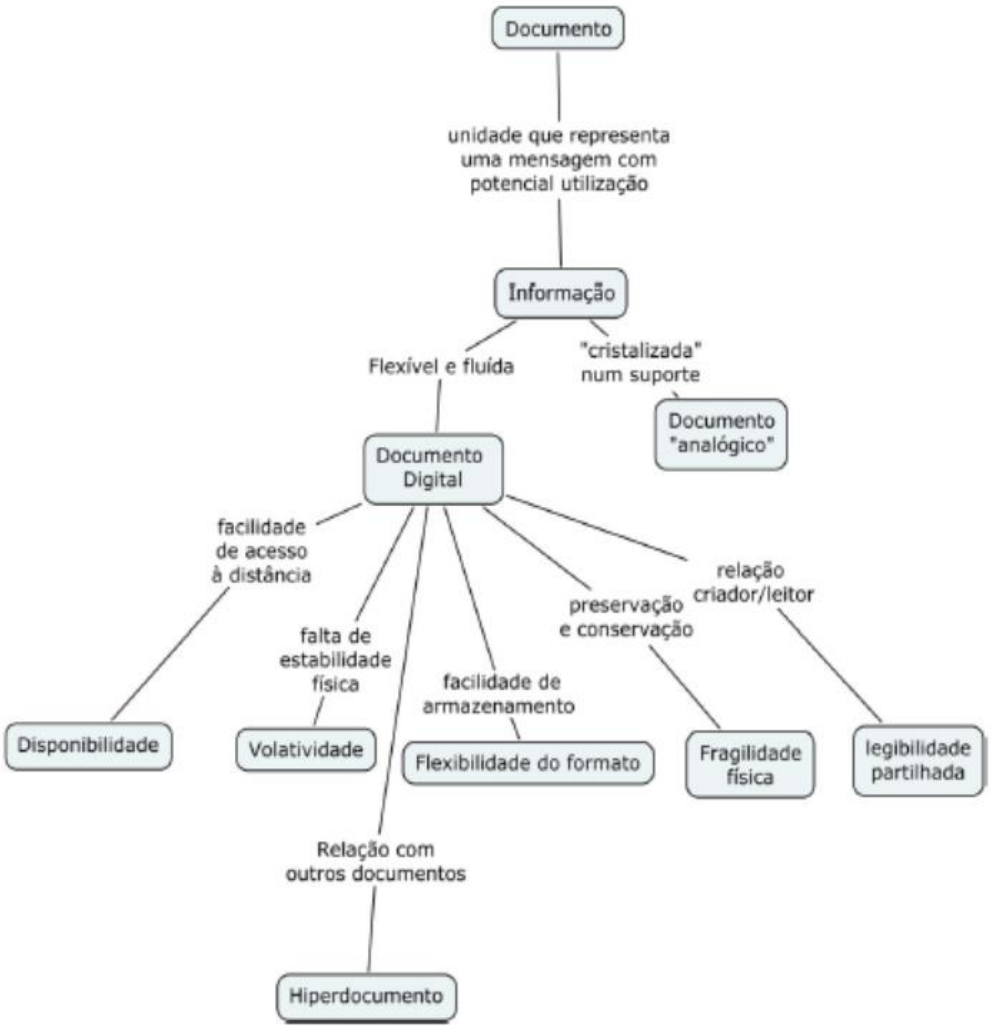
Para Buckland (1997), a mudança para o meio digital representa uma virada que possibilita a redefinição do próprio documento. Além de alargar o conceito de documento para incluir elementos não textuais, a exemplo de imagens, a ideia de redefinição proposta por Buckland seria depois endossada por López Yepes (1998), que aponta que o documento digital é um objeto informativo livre das limitações de suportes tradicionais, como o papel, já que a circulação em meio eletrônico é facilitada. Para o autor, o documento digital consolida de vez a ideia de documento.

Outro autor que escreveu sobre a temática e trouxe características do documento digital foi Michel (2000), o qual aponta características como: a facilidade de armazenamento em comparação com os documentos analógicos, o que possibilita facilidade também nos processos de localização e recuperação; a disponibilidade de acesso à distância de forma instantânea e a flexibilidade de formato. Outros pontos destacados pelo autor são a possibilidade de relacionar documentos digitais; a versatilidade para registrar informações e se adequar às necessidades da comunicação humana dos tempos atuais e, por fim, a necessidade dos profissionais que lidam com informação adequarem os seus trabalhos aos meios tecnológicos.

Santos e Flores (2012), em uma proposta de definição, dizem que " [...]um documento digital é a informação registrada em suportes acessíveis por meio de um equipamento computacional." Os autores apontam benefícios encontrados nos documentos digitais, tais como: redução de custos e a otimização da criação, tramitação e difusão dos documentos; novas possibilidades de interação entre pessoas, documentos e instituições de informação e a adaptação dos novos usuários frente ao mundo digital, com familiarização que facilita o acesso e uso dos documentos e suas informações. Com isso, os autores também citam a coexistência de documentos em suportes analógicos e digitais, excluindo a substituição do primeiro pelo segundo.

Siqueira (2012) realizou trabalho de revisão de literatura com o objetivo de levantar os autores e conceitos que abordam o documento digital, a fim de sistematizar as ideias que se têm sobre o assunto. A autora, ao fim do trabalho de revisão, apresenta um diagrama de conceitos elaborado em forma de mapa mental, interligando as definições que encontrou durante a revisão feita. O diagrama está replicado na imagem abaixo:

Figura 1 - Mapa conceitual com definições de documentos



Fonte: Siqueira, Jessica Campos (2012).

A proposta apresentada no diagrama é que se parta de uma definição única para documento, um conceito básico para que se entenda tanto o que é documento analógico quanto o que é documento digital. Corroborando o que diz López Yepes (1998), a diferença entre o analógico e o digital aqui é a fluidez, onde o digital se mostra livre da cristalização da informação em um suporte. A autora também destaca a importância de se considerar as características distintas entre as modalidades documentais, apontando os seguintes traços no

documento digital: facilidade de ser acessado, reproduzido, transmitido e armazenado, como pontos positivos e a fragilidade física e falta de estabilidade como pontos negativos.

A mudança de paradigma dos documentos passando do analógico para o digital também implica em mudanças na relação do leitor com tais documentos. A próxima subseção é dedicada a elucidar importantes aspectos da leitura em meio digital.

2.2 Leitura em meio digital

Na medida em que as tecnologias de informação e comunicação foram incorporadas no dia a dia das pessoas, e dado que documentos como textos, fotografias, audiovisuais e outros gêneros documentais migraram para formatos digitais, a forma de consumo dos usuários de informação também se adequou ao novo contexto. Um estudo cientométrico realizado por Dantas *et* al. (2017) sobre o cenário das pesquisas sobre leitura mostra o aumento de publicações sobre o tema em 2009, o que é explicado pela adoção de novos meios de leituras tais como computadores, *tablets* e *smartphones*. Os autores também notaram que o aumento de pesquisas sobre leitura coincide com o lançamento de dois *gadgets* de leitura digital populares: o *Kindle*, leitor eletrônico da empresa *Amazon*, e o *iPad*, *tablet* da empresa *Apple*.

O aumento da população leitora em meios digitais continua crescente. Um relatório produzido pela *BusinessWire*[3] aponta tendência de aumento em 28% das receitas geradas pelo mercado global de *e-books*, acrônimo para *eletronic book*, os livros digitais. No Brasil, uma pesquisa realizada pela Câmara Brasileira do Livro com base no ano de 2021, publicada em 2022[4], corrobora a tendência global, apresentando um crescimento geral de 23% no faturamento das vendas de livros digitais, o que representa 6% do mercado editorial no país.

A leitura digital requer do leitor diferentes competências daquelas que a leitura em documentos analógicos demanda. Agora, além do processo básico de interpretação e compreensão de signos, também se mostra necessária a adaptação ao dispositivo digital utilizado para a leitura, o que é chamado por Zayas (2010) de alfabetização digital. Outro fator diferencial da leitura em meio digital, mencionado na seção anterior, é o hipertexto. Além de alterar a estrutura do documento, a leitura digital altera também a forma que o leitor lida com o conteúdo, que agora pode fazer vários itinerários e rotas para a leitura, não mais seguindo o fluxo linear imposto em mídias analógicas.

---

[3] Disponível em: https://www.researchandmarkets.com/reports/4894553/global-digital-publishing-market-2022-2026

[4] Disponivel em: https://cbl.org.br/pesquisas_de_mercado_categoria/1-producao-e-vendas-do-setor-editorial-brasileiro/

É natural também observar mudanças na interação do leitor com o livro quando se muda o suporte da leitura. Os aparelhos utilizados para tais leituras desempenham um papel fundamental na determinação de como um leitor vai interagir com um texto, na medida em que cada dispositivo vai disponibilizar diferentes funções de interação: cliques em *links* que são acessados na internet, possibilidade de grifar, anotar ou destacar, salvar comentários, por exemplo. Estas funções, inclusive, são apontadas como benefícios da leitura digital em detrimento da leitura em meio físico:

> O que o livro sempre quis foi ser anotado, marcado, sublinhado, ter as pontas de suas páginas dobradas, ser resumido, ganhar referências cruzadas, hiperlinks, ser compartilhado, e dialogar. Ser digital lhes permite fazer tudo isso e muito mais. (Kelly, 2011).

A UNESCO (2013) também apresenta um conceito relacionado à leitura em meio digital denominado em inglês *mobile learning* para se referir ao conjunto de práticas de ensino possibilitadas pela utilização da tecnologia digital móvel, combinada ou não com outras tecnologias de informação e comunicação, e entre as práticas se encontra a leitura digital. Um estudo de Bernardo e Karwoski (2017), realizado com estudantes de graduação do curso de Letras, mostrou que 50% dos participantes escolheram, durante o período do experimento, ler textos acadêmicos em dispositivos digitais. Os autores perceberam o hábito de praticar a leitura de textos para aprendizado utilizando o meio digital entre os alunos:

> Quanto a textos menores, como artigos acadêmicos, não titubeamos em afirmar que não só a preferência como já hábito formado é a leitura em arquivo digital via telefone celular. Os professores mesmos, segundo os participantes, estimulam essa forma de leitura a partir da disponibilização de textos e materiais didáticos *on-line*. (Bernardo; Karwoski, 2017)

Para além do meio utilizado para realizar a leitura, existe também o conjunto de comportamentos executados por um leitor durante a leitura, conjunto este chamado de estratégias de leitura (Cook; Mayer, 1983). Além da influência do meio escolhido para leitura, seja digital ou analógico, o que define o comportamento adotado durante uma leitura é o objetivo final do leitor. Em uma leitura com objetivo de aprendizado, as práticas mais comuns são: sublinhar palavras-chave, grifar passagens do texto, anotar e sumarizar tópicos. A próxima subsubseção destrincha um destes comportamentos adotados por leitores: as anotações.

2.2.1 Anotações: conceitos e aplicações

Anotar é uma estratégia de leitura para auxiliar o leitor na compreensão do texto. DiVesta e Gray (1972) apontam duas funções distintas possíveis de serem executadas pelo ato de anotar durante uma leitura e/ou aula: armazenamento e codificação. A primeira função caracteriza as anotações como um armazenamento externo de informações, que podem servir como meio de revisão posterior à leitura inicial. Já a função de codificação sugere que o processo de anotar estimula o aumento da concentração, da capacidade de análise sobre a leitura, do desenvolvimento de ideias próprias sobre o assunto tratado e da organização de informações.

A anotação possui também um sentido de enriquecimento informacional de documentos, como proposto no conceito de anotação semântica, cujo objetivo é possibilitar que um documento digital criado para interpretação humana possa também ser interpretado por máquinas. Por meio da adição de palavras com significados relacionados à temática do texto, as anotações incluídas no documento permitem a recuperação da informação em sistemas de buscas precisos. (Fontes *et al.*, 2010).

2.2.2 Destaques em texto

Neste trabalho, utilizaremos os termos destaque ou destacar como sinônimos para grifo ou grifar, respectivamente. O termo destacar, uma tradução do termo em inglês *highlight*, tem significado definido pelo dicionário Merriam Webster como *to center attention on: emphasize, stress*; em português significa salientar algo que merece atenção. Já o termo grifar, definido em português pelo dicionário Michaelis como escrever com grifo; sublinhar, nos aproxima da ideia de destacar termos em um texto, justamente o contexto definido no estudo proposto.

Assim como as anotações, destacar informações em um texto utilizando grifos também está intimamente ligado à compreensão da leitura, segundo Leutner et al. (2007). Os autores apresentam a categorização de estratégias de aprendizado por leitura em dois níveis: 1) superficial; e, 2) profundo. No nível superficial, o leitor se concentra em memorizar as informações apenas lendo o texto, enquanto no nível profundo o leitor seleciona e estrutura as informações mais importantes, adicionando anotações para relacionar o conteúdo com outros já conhecidos, criando, assim, um esquema mental para a absorção do novo conteúdo lido.

O processo de selecionar e destacar as informações que o leitor considera importantes está ligado ao segundo nível de aprendizagem, o mais profundo. Ainda segundo Leutner et al.

(2007), o ato de grifar auxilia o processo de armazenar e recuperar informações, num contexto mental do leitor. Weinstein e Mayer (1986) também escrevem sobre a técnica de destacar informações e apontam duas funções para a prática: 1) identificar e focar a atenção do leitor nas informações importantes de um texto; e 2) armazenar as informações importantes identificadas na memória de curto prazo, para que seja processada posteriormente.

Outro ponto a ser considerado é que o ato de grifar funciona como um guia visual durante a revisão ou estudo posterior. Quando o leitor retorna ao texto, as marcações coloridas ou sublinhadas servem como indicadores visuais, apontando áreas importantes que merecem atenção. Isso economiza tempo e esforço, já que o leitor pode focar nas partes mais relevantes do material.

No âmbito da leitura digital e destaques em textos, já existem iniciativas que tratam e utilizam os destaques gerados por leitores, uma delas vem do Kindle. O sistema de marcação de trechos pelos usuários do Kindle é uma funcionalidade que permite aos leitores destacar e anotar passagens em livros digitais. No Kindle, os leitores têm a capacidade de selecionar trechos específicos de um livro, seja para enfatizar partes importantes, fazer anotações pessoais ou simplesmente marcar passagens interessantes. Essas marcações ficam armazenadas no sistema e podem ser facilmente acessadas posteriormente, proporcionando uma experiência de leitura interativa e personalizada.

Além disso, o Kindle oferece recursos adicionais, como a visualização das marcações feitas por outros leitores em determinados trechos. Isso cria uma comunidade virtual de leitores, permitindo que compartilhem insights e comentários sobre partes específicas dos livros. Essa abordagem social da marcação contribui para uma experiência de leitura mais envolvente e conectada, mesmo em um ambiente digital.

A capacidade de marcar trechos no Kindle não apenas replica a prática tradicional de marcação em livros físicos, mas também aprimora com recursos digitais, como a facilidade de busca, organização automática de marcações e a capacidade de compartilhar notas de leitura. Essa integração de tecnologia na marcação de trechos reflete a evolução da leitura para o mundo digital, mantendo e aprimorando aspectos valiosos da experiência de leitura física.

2.3 Organização da informação

A informação, enquanto objeto de estudo da CI, é um termo bastante discutido com múltiplas definições. Uma miríade de estudos aborda seus diversos aspectos e possíveis formas de análise. Para citar um exemplo das definições levantadas em tais estudos, Le Coadic (2004)

diz que a informação é um conhecimento registrado que comporta um elemento de sentido, registro este que é feito graças a um sistema de signos, a linguagem. Com base nas definições e abordagens existentes, emergem formas de exploração e estudo da informação, dentre as quais está a organização da informação (OI). Aguiar e Kobashi (2013), sobre a organização, afirmam que "no domínio da CI, ela pode ser compreendida como uma série de atividades processuais com a finalidade de descrever intelectualmente conteúdos documentais para serem representados nos sistemas de recuperação da informação".

Organizar informações já era uma preocupação muito antes da consolidação da CI como área de estudo, visto que Otlet (1934) já considerava a prática de organizar dentro dos processos da documentação. Para o autor, o objetivo principal era: "A organização da documentação em uma base cada vez mais abrangente, cada vez mais prática, de forma a alcançar para o trabalhador intelectual o ideal de um mecanismo para explorar o tempo e o espaço".

Conceituando o termo organização da informação, Café e Sales (2010) definem que é um processo de arranjo de acervos tradicionais ou eletrônicos realizado por meio da descrição física e de conteúdo (assunto) de seus objetos informacionais.

É possível mencionar diversas técnicas desenvolvidas neste âmbito organizacional para que se alcance os objetivos mencionados acima, como a classificação e a indexação, por exemplo. A classificação permite representar a informação em forma de símbolos de classificação e números, descrevendo o conteúdo de forma abrangente. Já a indexação busca descrever o conteúdo de forma específica, gerando termos e palavras-chave que possam atender possíveis perguntas de usuários em suas buscas por informação. As duas práticas atuam de forma complementar.

A partir de práticas como estas citadas, torna-se possível o desenvolvimento de sistemas e ferramentas organizadas para o uso no processo de recuperação da informação. A organização da informação na forma em que se conhece tem suas origens na criação do sistema de Classificação Decimal de Dewey, em 1876, que propõe a classificação por assuntos de maneira hierárquica. Outro sistema de classificação de grande importância nesse campo é a Classificação Decimal Universal, proposta por Paul Otlet e Henri La Fontaine em 1905, que possibilita a organização temática por meio de códigos numéricos.

Com as contribuições dos estudos do bibliotecário Ranganathan, um novo método surgiu no campo da organização informacional. O estudioso foi responsável pela formulação, entre 1933 e 1960, do *Colon Classification*, um método que extrapola as relações hierárquicas apresentadas em outros sistemas de classificação existentes até o momento e permite a categorização de assuntos que se agrupam por características similares, possibilitando a criação

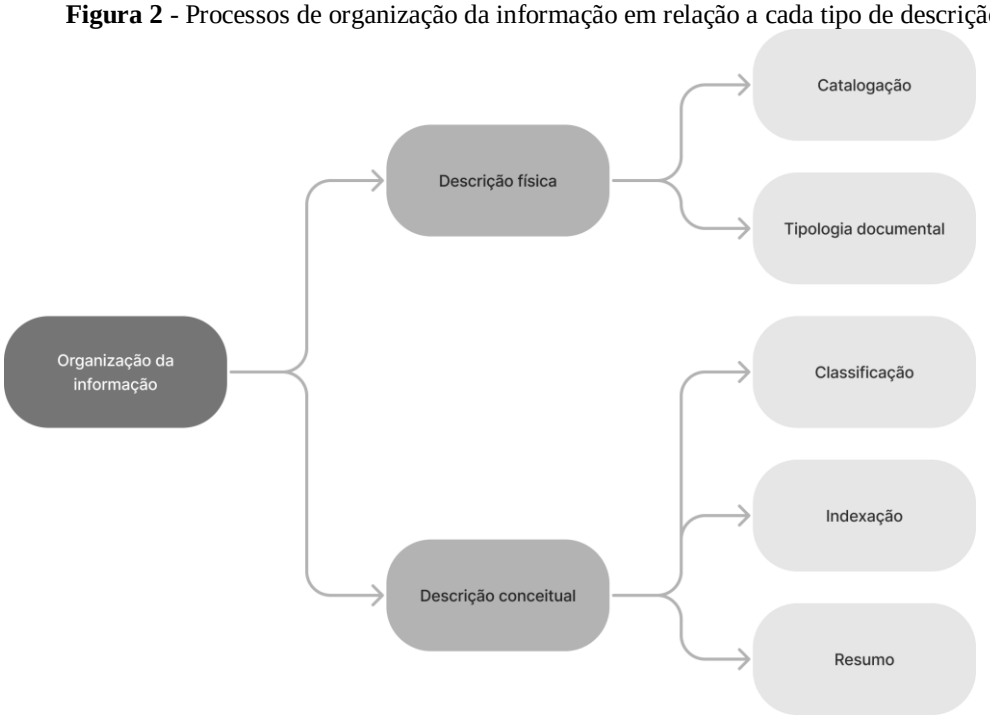
de relações entre estes. Além do citado método, foi Ranganathan quem proporcionou arcabouço teórico para posteriores construções de códigos de classificação diversos ao descrever seu modelo de princípios de categorias fundamentais, o PMEST.

A influência dos estudos dos autores mencionados nesta seção, considerados pioneiros no contexto da organização da informação, perduram nas práticas e estudos atuais. A exemplo, podemos citar os termos trazidos à luz por Ranganathan em seu modelo teórico que ainda hoje são utilizados como formas de classificação em diversos sistemas de recuperação da informação, tais como facetas, níveis e focos. Os resultados possibilitados e alcançados por tais estudos vão ao encontro da construção do conhecimento, que é construído a partir da circulação e uso da informação, sua matéria-prima.

A organização da informação está intimamente ligada ao processo de descrever objetos informacionais. Shera e Egan (1953) apontam que a descrição é o processo de "individualização de determinado item entre o vasto número dos que formam o conjunto de literatura". Essa descrição se dá em dois níveis: o conceitual e o físico. No nível conceitual, é necessário descrever o conteúdo do objeto informacional, o conhecimento registrado. Já no nível físico, a descrição se encarrega de especificar as características do suporte no qual o conhecimento está registrado. Destes dois tipos de descrições, derivam-se alguns processos e práticas de organização da informação.

Baseado na sistematização dos elementos gerais da organização da informação realizada por Brascher e Monteiro (2010), o esquema abaixo, seguido de breve definição dos termos elencados, foi elaborado com o intuito de ilustrar os procedimentos realizados no âmbito da OI e suas ligações com os níveis de descrição supramencionados:

Figura 2 - Processos de organização da informação em relação a cada tipo de descrição



Fonte: adaptado de Brascher e Monteiro (2010).

- **Catalogação:** processo técnico para registro e descrição de itens tendo em vista a organização de catálogos;
- **Tipologia documental:** designação dos tipos de documentos segundo o aspecto de sua representação nos diferentes suportes: textuais, audiovisuais, iconográficos e cartográficos;
- **Classificação:** Conjunto de operações que levam à colocação de um documento em uma determinada ordem, mediante a utilização de um esquema de classificação;
- **Indexação:** Representação do conteúdo temático de um documento por meio dos elementos de uma linguagem documentária ou de termos extraídos do próprio documento (palavras-chave, frases-chave);
- **Resumo:** representação concisa e acurada do conteúdo de um documento; síntese de um documento; notação de conteúdo.

O foco da pesquisa é a investigação de aspectos de um dos processos acima representados: a indexação, definido por Lapa e Corrêa (2014) do seguinte modo:

> A indexação é um processo de tratamento temático essencial, pois consiste no ato de identificar e descrever um documento de acordo com o seu assunto, e cujo principal objetivo é orientar o usuário sobre esse conteúdo intelectual,

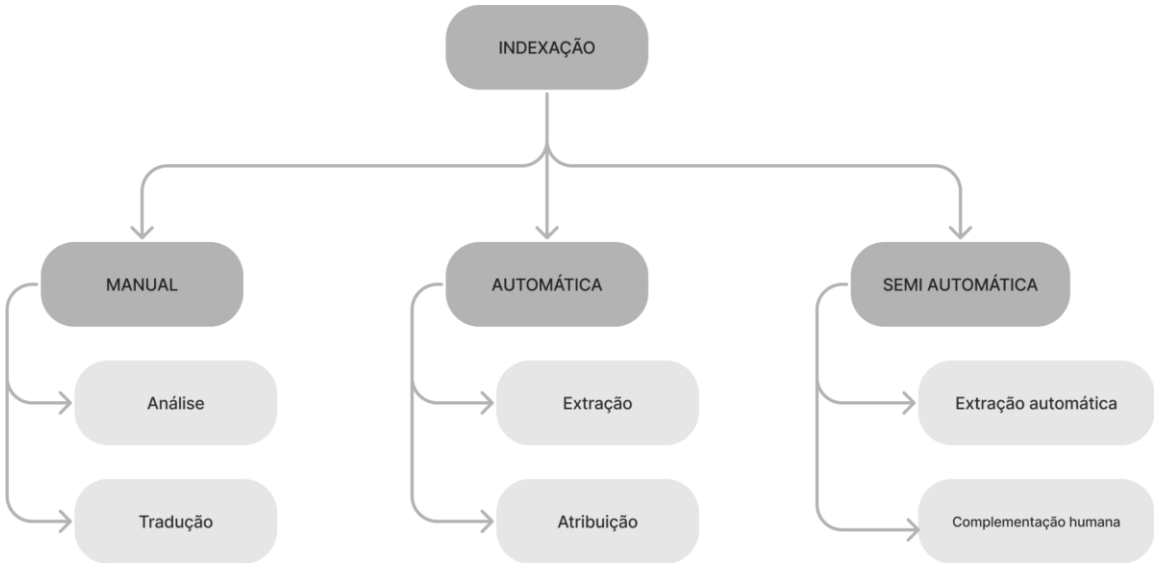
> permitindo, dessa forma, a sua recuperação de forma ágil e eficiente. (Lapa; Corrêa, 2014)

A próxima seção se concentra em apresentar conceitos fundamentais deste processo, bem como listar e caracterizar os tipos de indexação estudados e praticados no âmbito da organização da informação.

2.4 Indexação

A indexação, segundo Guinchat e Menou (1994), é a operação pela qual se escolhe os termos mais apropriados para descrever o conteúdo de um documento. A prática de atribuir termos que identifiquem um documento a partir das informações as quais ele carrega é de suma importância para a posterior recuperação da informação por meio de sistemas. A escolha de termos deve ser realizada de forma que atenda aos interesses do usuário que utilizará os sistemas de recuperação da informação. Assim, um mesmo documento pode ser indexado a partir de termos diferentes, a depender de quem realizará consultas (Lancaster, 2004). A seguir, na Figura 3, apresenta-se uma classificação dos tipos de indexação, considerando o tipo de indexação em relação à tecnologia utilizada e as etapas envolvidas.

Figura 3 - Classificação dos tipos de indexação em relação ao nível de automação



Fonte: adaptado de Lancaster (2004) e Pinto (2001);

No processo de indexação manual, Lancaster (2004) aponta duas etapas de trabalho: a análise conceitual e a tradução. Na etapa de análise, o profissional indexador realiza a leitura do documento a fim de compreender seu conteúdo e definir a temática tratada, definir o assunto do documento. É nessa etapa onde o indexador deve tomar decisões conceituais sobre como sintetizar o conteúdo da melhor forma para atender as necessidades informacionais específicas do grupo de usuários que será atendido. Aqui, o autor menciona que é importante a formulação de perguntas por parte do indexador tais como: de que se trata o documento? Por que foi incorporado a este acervo? Quais de seus aspectos são de interesse dos usuários?

A indexação orientada ao usuário, termo esse cunhado por Fidel (1994), abarca diversos pontos de questionamentos acerca de como atender satisfatoriamente às necessidades informacionais dos variados grupos de usuários que possam existir, e que são citados também na obra de Lancaster (2004). O primeiro ponto levantado é sobre a competência de conhecimento do público que fará uso do conjunto de documentos indexados, como um importante complemento do conhecimento das técnicas de indexação, indispensáveis ao trabalho. Outro ponto elucidado por Lancaster é a impermanência da indexação realizada, visto que os interesses de um usuário ou um grupo de usuários pode e provavelmente mudará ao longo do tempo. Como exemplo, aponta-se um grupo de pesquisadores que avançam em seus estudos e precisam organizar seus documentos com frequência para que continuem tendo suas necessidades informacionais atendidas.

Já na etapa seguinte, a de tradução, a tarefa do indexador consiste em sintetizar o assunto do documento, identificado na etapa de análise, em termos indexáveis que serão incluídos na base de dados. É nesta etapa que Lancaster (2004) apresenta os conceitos de indexação por atribuição e por extração. Na indexação por extração, o indexador retira do próprio texto a ser indexado alguns termos (palavras ou expressões) que considere mais representativos para realizar a indexação. Já na indexação por atribuição o indexador seleciona termos que não estão presentes no texto, mas que podem ser considerados relevantes dada a temática ou o interesse dos usuários. Nesse segundo tipo pode se utilizar ferramentas como o vocabulário controlado, que é uma lista de termos autorizados que se convencionou usar em um determinado contexto, como em uma instituição, por exemplo.

A indexação automática, realizada com apoio de ferramentas computacionais, é estudada desde a década de 1950, por meio da análise automatizada de textos. Os estudos, conforme Robredo (2005), mostram que o computador é uma ferramenta indispensável para garantir acesso rápido à bases de dados científicos, auxiliando nas tarefas de processamento de dados e informações. Lapa e Corrêa (2014) conceituam a indexação automática como um

conjunto de operações matemáticas, linguísticas e de programação que, quando aplicada na análise de documentos, faz o processamento dos conteúdos, selecionando automaticamente os termos representativos dos assuntos destes documentos para serem utilizados posteriormente em processos de recuperação de informação.

São identificados dois tipos de indexação automática: por atribuição e por extração, ambos referentes ao modo de extração dos termos de um documento textual. Lancaster (2004) define a indexação por atribuição como o processo de atribuir automaticamente termos oriundos de um vocabulário controlado de acordo com os conjuntos de palavras e expressões encontrados após a análise textual do documento. Neste tipo de indexação o vocabulário controlado possui influência no processo de análise dos documentos, ou seja, a escolha dos termos para atribuição está intimamente ligada e condicionada ao vocabulário que será utilizado. Segundo Lancaster, este é o tipo mais difícil de se realizar de forma computacional pois é necessário "[...] desenvolver, para cada termo a ser atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer frequentemente nos documentos [...]" (Lancaster, 2004).

Na abordagem de indexação por extração, os termos selecionados para indexar um documento são escolhidos dentro do próprio texto, usando a linguagem natural ao invés de termos estruturados em um vocabulário. Aqui, considerando os documentos digitais, é possível utilizar *softwares* computacionais para realizar um processo parecido com o desenvolvido por um indexador humano. Os termos extraídos podem atender aos critérios de frequência ou posição das palavras dentro do texto ou no resumo além da possibilidade de extração de termos considerando também o contexto do documento (Lancaster, 2004). Aqui, ao contrário da indexação por atribuição, o autor explicita o ótimo desempenho computacional na tarefa de extração, considerando-a bastante coerente quando validada.

Ainda no âmbito da indexação realizada com auxílio de computadores, além dos dois tipos de indexação já mencionados, existe também a indexação semiautomática. Segundo Pinto (2001), este é um tipo de indexação que une etapas da indexação manual e da indexação automática, de forma que a indexação por meio da extração de termos realizada por um *software* seja revisada e validada por um indexador humano. O *software* SISA mencionado na figura 4 é uma ferramenta para realização deste tipo de indexação com uma relacionada ao processo em que o sistema computacional realiza a atividade de análise do conteúdo do documento e, posteriormente, um indexador humano avalia os termos para indexação propostos pelo sistema (Narukawa, 2011).

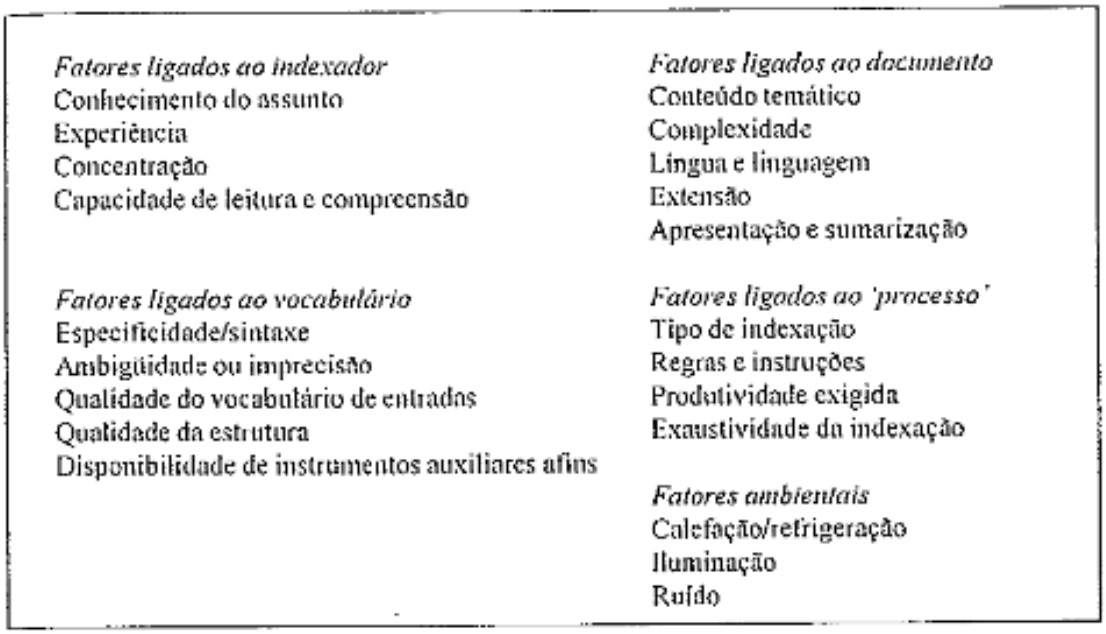
Mai (2000) discute a indexação no campo epistemológico e propõe cinco concepções, ou posicionamentos epistemológicos, para embasar a prática da indexação. São eles:

- **Concepção Simplista de Indexação:** foca exclusivamente na extração automática e manipulação estatística de palavras, considerando que a soma das palavras em um documento constitui seu assunto. Essa visão está intrinsecamente ligada ao empirismo;
- **Concepção Orientada ao Documento:** concentra-se na informação presente no documento, envolve a investigação de partes específicas do documento, onde a importância da informação é determinada pelo indexador. Além disso, está alinhada à posição racionalista, que busca objetivamente determinar o assunto dos documentos por meio do raciocínio puro;
- **Concepção Orientada ao Conteúdo:** procura descrever o conteúdo do documento da forma mais abrangente possível. Essa visão objetivista sustenta que há apenas uma análise correta para um determinado documento. Baseando-se na investigação cuidadosa das diferentes interpretações do documento, essa concepção busca objetivar o assunto do documento, ancorando-se nas circunstâncias históricas e culturais que influenciam sua produção;
- **Concepção Orientada ao Usuário:** coloca o foco nos usuários, considerando o nível geral de conhecimento ou o domínio de trabalho/pesquisa do usuário. O indexador leva em conta o ambiente do usuário, sendo sensível ao fato de que usuários em uma biblioteca pública podem necessitar de indexação diferente em comparação com usuários em uma biblioteca universitária. Essa abordagem é fundamentada na posição epistemológica pragmática, centrada no potencial uso futuro do documento pelos usuários;
- **Concepção Orientada a Requisitos:** os indexadores possuem conhecimento das necessidades individuais de informação e tarefas de trabalho dos usuários. Essa concepção é especialmente útil em organizações menores, onde os indexadores podem conhecer as necessidades de cada indivíduo. Semelhante à concepção orientada para o usuário, a orientação para requisitos é pragmática, concentrando-se em necessidades específicas de informação de pessoas específicas e adaptando-se ao longo do tempo com base no potencial uso futuro dos documentos.

De acordo com estas colocações, percebe-se na investigação realizada neste trabalho aspectos das concepções orientadas ao documento e ao usuário, no sentido em que trabalha com partes específicas do documento (aquelas grifadas pelos leitores) e também no sentido em que foca no usuário, considerando não só a área de pesquisa/conhecimento como levando em consideração a participação ativa do usuário (leitor) no processo de indexação.

Quanto à qualidade da indexação, Lancaster (2004) aponta que de forma pragmática o que se entende como boa indexação é aquela que recupera um alto número de documentos relevantes em relação ao termo de busca utilizado, entretanto, existe a carência de qualificadores e indicadores para determinar melhor a boa indexação. Para contribuir com a discussão, o autor traz fatores que impactam na qualidade da indexação, mostrados na Figura 5.

Figura 5 - Fatores que interferem na qualidade da indexação



Fonte: Lancaster (2004, p. 89)

2.5 Aprendizagem de máquina

O uso de tecnologias computacionais, presente em diversas áreas do conhecimento, também é observado na Ciência da Informação. A digitalização de acervos, a modelagem de dados e metadados para a alimentação de sistemas informatizados de busca e recuperação de informações são exemplos da utilização de recursos computacionais em práticas originárias da CI. O avanço do uso tecnológico supramencionado, não só na ciência, mas também no cotidiano das pessoas, resultou em uma elevação exponencial na produção de dados. A chamada era *big data* em que se vive, exige o desenvolvimento de estratégias para que se lide de forma eficiente com os dados, além de explorar o potencial existente neles para produção de novos conhecimentos.

Neste contexto, está o *machine learning (ML)*. A aprendizagem de máquina, em português, é a prática de programar computadores com o objetivo de otimizar a execução de tarefas específicas usando conjuntos de dados como base para o aprendizado (Ayodele, 2010).

Com a utilização de algoritmos computacionais com variadas funções, o aprendizado de máquina apresenta benefícios importantes tais como a melhoria constante no funcionamento das máquinas, a execução de tarefas em volume e tempo impossíveis para seres humanos e também a descoberta de relações e padrões em grandes quantidades de dados. Da mesma forma que pode ser aplicada para fins práticos e de pesquisa em muitas áreas do conhecimento, o aprendizado de máquina também é influenciado por diversas fontes. Algumas áreas que influenciam e criam variações nas técnicas e métodos de estudos, citadas por Ayodele (2010), são a estatística, a multidisciplinar inteligência artificial e também as áreas que estudam o comportamento humano a partir de modelos cerebrais e psicológicos.

As definições de aprendizagem de máquina variam conforme o autor consultado. Segundo Murphy (2012), é um conjunto de métodos que pode detectar padrões em dados de forma automática e utilizar os padrões descobertos para realizar previsões de novos dados. Também pode realizar tarefas de tomadas de decisão, como o planejamento da coleta de novos dados. Ainda para o autor, a aprendizagem de máquina surge no contexto da chamada era *big data* e na demanda por novas tecnologias capazes de lidar com as dimensões de dados desta era. Os autores Goodfellow, Bengio e Corville (2016) também conceituam a aprendizagem de máquina. Para estes, a aprendizagem é a capacidade de um sistema adquirir conhecimento próprio a partir de padrões descobertos em conjuntos de dados. Com o conhecimento adquirido, a máquina se torna capaz de resolver problemas do mundo real e também de tomar decisões que, a princípio, parecem subjetivas.

Além do conceito, Murphy (2012) também apresenta os tipos de aprendizagem de máquina. O autor cita os dois tipos gerais de aprendizado: o preditivo, também conhecido como aprendizado supervisionado, e o descritivo, conhecido como aprendizado não supervisionado, com menção também ao tipo de aprendizado semi-supervisionado. No tipo preditivo, o aprendizado acontece por meio do mapeamento das relações entre entradas e saídas de dados, a partir de conjuntos de dados para treinamento que são fornecidos à máquina. Neste tipo de aprendizado os dados são rotulados, ou seja, a máquina recebe o tipo de dado que deve gerar como resultado de seu aprendizado.

Já no tipo descritivo ou não supervisionado só são fornecidos os dados de entrada e o objetivo é que a máquina seja capaz de encontrar padrões nos conjuntos de dados disponíveis, realizando a descoberta de conhecimento. Além dos três tipos citados, o autor ainda menciona mais um outro tipo, o aprendizado por reforço. Neste tipo o aprendizado é impulsionado por recompensas ou punições, a depender do comportamento da máquina. Essas classificações têm como referência a quantidade de supervisão humana envolvida nos processos de aprendizagem.

Visto isso, uma breve descrição para os tipos de aprendizado, segundo Russel (2018), é a que segue:

- **Supervisionado:** as tarefas executadas pelo algoritmo de aprendizado supervisionado têm por objetivo ensinar a máquina a resolver problemas que já possuem solução conhecida. Conforme essa finalidade básica, existem dois tipos principais de aprendizado supervisionado: a classificação e a regressão. Na classificação, o objetivo é identificar se um item está ou não associado a uma categoria, como por exemplo, a detecção se um e-mail é ou não um spam; se determinada imagem ou vídeo possui ou não um determinado objeto ou conteúdo ou também se um texto ou mensagem possuem sentimento positivo ou negativo.
- **Não supervisionado:** nesta abordagem o objetivo é fazer com que a máquina encontre padrões, estruturas, semelhanças ou anomalias em um ou vários conjuntos de dados ainda não rotulados. É com este tipo de aprendizado que se encontram *insights* em dados que trazem informações não exploradas por humanos. As finalidades básicas dos algoritmos de aprendizado não supervisionado são a clusterização e a redução de dimensionalidade. A clusterização, ou agrupamento, cria conjuntos de objetos derivados de um grande conjunto de dados oferecido com base em similaridades de características, útil para entendimento de relações dos objetos dentro de um universo dado. Já a redução de dimensionalidade busca generalizar um conjunto de dados reduzindo o número de variáveis mas mantendo as características mais importantes, com objetivo de simplificar análises e promover mais eficiência ao lidar com os dados.
- **Aprendizado por reforço:** Mais utilizada em cenários de mundo real como jogos e robótica, essa abordagem utiliza um agente computacional para atuar em um dado ambiente interagindo em situações e tomando decisões com o objetivo de alcançar o máximo de recompensas possíveis, estas pré-definidas pelo modelo utilizado. Conforme o agente executa ações do tipo tentativa e erro, o aprendizado vai acontecendo por meio da associação de punição ou recompensa para cada ação e decisão tomada.

Uma máquina se torna capaz de aprender aquilo que lhe é designado por meio do uso de algoritmos, mais especificamente, um algoritmo de aprendizagem. Um algoritmo, segundo Dijsktra (1971), é uma descrição de padrões de comportamentos expressos em termos de um conjunto finito de ações, que quando combinados com dados, chegam a uma solução desejada. O tipo de algoritmo utilizado em um projeto está intimamente ligado ao tipo de aprendizado que ocorrerá. Russel (2018) menciona alguns dos principais algoritmos utilizados nos

aprendizados supervisionados, não supervisionados e por reforço, esquematizados no Quadro 1, a seguir:

Quadro 1 - Sistematização de algoritmos de aprendizado de máquina

Tipo de Aprendizado	Algoritmo	Descrição	Função
Aprendizado Supervisionado	Regressão Linear	Encontra a relação linear entre variáveis de entrada e saída em um conjunto de dados.	Regressão
Aprendizado Supervisionado	Árvores de Decisão	Cria uma estrutura de árvore para classificar dados em diferentes categorias com base em perguntas e respostas.	Regressão/ Classificação
Aprendizado Supervisionado	Redes Neurais Artificiais	Usa algoritmos de aprendizado de rede neural para reconhecer padrões complexos em dados.	Regressão
Aprendizado Supervisionado	Regressão Logística	Modela a probabilidade de um evento ocorrer, dado um conjunto de variáveis de entrada. É útil para prever resultados binários.	Regressão
Aprendizado Supervisionado	Floresta Aleatória	Combina várias árvores de decisão para obter uma previsão mais precisa. É útil para lidar com grandes conjuntos de dados e minimizar o overfitting.	Regressão/ Classificação
Aprendizado Supervisionado	*k-Nearest Neighbors* (KNN)	Classifica os dados com base nas classes dos pontos pré definidos (k) mais próximos. É útil para identificar agrupamentos de dados e para prever valores.	Classificação
Aprendizado Supervisionado	Naive Bayes	Modela a probabilidade de uma classe com base na probabilidade condicional das variáveis	Classificação

		de entrada. É útil para classificar dados com base na frequência de ocorrência de certas características.	
Aprendizado Supervisionado	Máquinas de Vetores de Suporte (SVMs)	Classifica os dados dividindo-os em hiperplanos multidimensionais. É útil para separar classes de dados não lineares.	Regressão/ Classificação
Aprendizado Não-Supervisionado	*K-Means*	Agrupa dados em diferentes grupos com base na distância euclidiana entre eles.	Clusterização
Aprendizado Não-Supervisionado	Análise de Componentes Principais (PCA)	Reduz a dimensionalidade dos dados, transformando-os em um espaço de menor dimensão sem perder muita informação.	Redução de dimensionalidade
Aprendizado Não-Supervisionado	Redes Neurais *Autoencoder*	Reduz a dimensionalidade dos dados, comprimindo-os em uma representação latente, e depois os reconstrói novamente. É útil para detecção de anomalias e para identificar características importantes nos dados.	Redução de dimensionalidade
Aprendizado Não-Supervisionado	Agrupamento Hierárquico	Agrupa dados em diferentes grupos com base em uma estrutura hierárquica.	Clusterização
Aprendizado por Reforço	*Q-Learning*	Usa um processo de tentativa e erro para aprender a tomar decisões em um ambiente dinâmico, com base em recompensas ou penalidades.	Não se aplica
Aprendizado por Reforço	*Deep Q-Learning*	Usa uma rede neural profunda para aprender a	Não se aplica

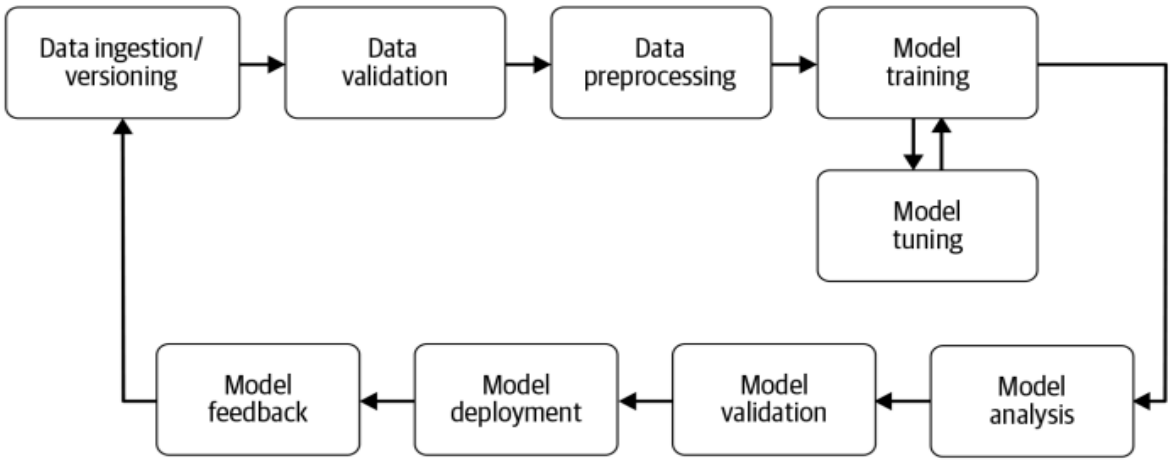
		tomar decisões em um ambiente dinâmico, com base em recompensas ou penalidades.	
Aprendizado por Reforço	Política de Gradiente	Usa um algoritmo de otimização para encontrar os melhores protocolos de ação em um ambiente dinâmico.	Não se aplica

Fonte: elaborado pela autora

2.5.1 Processos de aprendizagem de máquina

Para que uma aplicação de aprendizagem de máquina seja construída e funcione de forma efetiva para os fins que foi programada, é necessário seguir fluxos de processos predeterminados. Esses processos, quando elencados, formam o que se conhece por pipeline, um fluxo de trabalho padronizado e executável cuja implementação corrobora para acelerar, reutilizar, gerenciar e implantar modelos de aprendizado de máquina, conforme apontam Hapke e Nelson (2020). De forma geral, considerando que o problema ao qual o projeto deseja solucionar já esteja definido, algumas etapas são comuns a todos os projetos de aprendizagem de máquina, conforme a imagem abaixo bem ilustra:

Figura 6 - Etapas de um pipeline de aprendizagem de máquina



Fonte: Hapke e Nelson (2020)

O pipeline tem a característica de ser circular, como um ciclo de processos recorrentes, contribuindo para a automatização dos processos e a contínua melhoria dos modelos criados, que são constantemente avaliados e treinados. As etapas representadas na imagem acima seguem elencadas e elucidadas abaixo, com tradução nossa da obra de Hapke e Nelson (2020):

a) **Ingestão e versionamento de dados:** são as etapas onde os dados oriundos de fontes diversas são preparados para se conectarem ao formato desejado para se utilizar no modelo em questão;
b) **Validação de dados:** etapa em que se valida os dados no sentido de verificar se existem anomalias nos dados, se o modelo dos dados é aplicável ao modelo que será treinado, se as estatísticas do conjunto de dados são aplicáveis ao modelo que será treinado;
c) **Pré-processamento de dados:** é a etapa na qual se ajustam os rótulos dos dados, se convertem os vetores (textuais, numéricos, etc) de forma a garantir que os dados selecionados possam ser processados pelo modelo que será treinado;
d) **Treinamento e ajuste do modelo:** é o centro de um pipeline de machine learning, etapa onde se treina o modelo escolhido para consumir dados, interpretá-los e responder com uma predição com a menor chance de erro possível;
e) **Análise do modelo:** aqui a acurácia do modelo é analisada, os parâmetros utilizados são validados e modificados se necessário, a precisão é medida, bem como realizam-se testes com conjunto de dados maiores do que aqueles utilizados no treinamento anterior à análise. Também se analisam as predições feitas pelo modelo e a capacidade de aplicação do modelo em dados diferentes, em relação a sua capacidade de predição;
f) **Versionamento do modelo:** etapa dedicada a documentar o modelo em seu estado atual: registra-se seus parâmetros e características, indicando quais destes serão mantidos em uma próxima versão e quais serão eliminados ou modificados. É fundamental para que se conheça o histórico do modelo e entenda como ocorreu sua evolução;
g) **Implantação do modelo:** é a etapa na qual o modelo é disponibilizado para uso, envolvendo conhecimentos não só de programação como também de arquitetura do modelo e seus requisitos de hardware. Pode ser feita, geralmente, de três maneiras: em um modelo cliente-servidor, em um navegador ou em um edge device, um tipo de dispositivo que funciona como ponta de entrada para a rede de um servidor. Dentre estas, a mais comum é a implantação em um modelo cliente-servidor.
h) ***Feedback* sobre o modelo:** etapa na qual se coletam informações sobre o funcionamento do modelo treinado, pode ser realizado de forma automatizada e/ou com participação humana. A depender do formato em que se registra o feedback, este pode ser utilizado com o conjunto de dados para o próximo input do modelo.

As etapas descritas em um modelo de fluxo de trabalho como o apresentado são genéricas para que possam se adaptar ao contexto ou problema ao qual será aplicado. A próxima

seção apresenta meios de aplicação de aprendizagem de máquina no contexto da organização da informação, mostrando as utilidades e pontos de convergência entre as duas áreas.

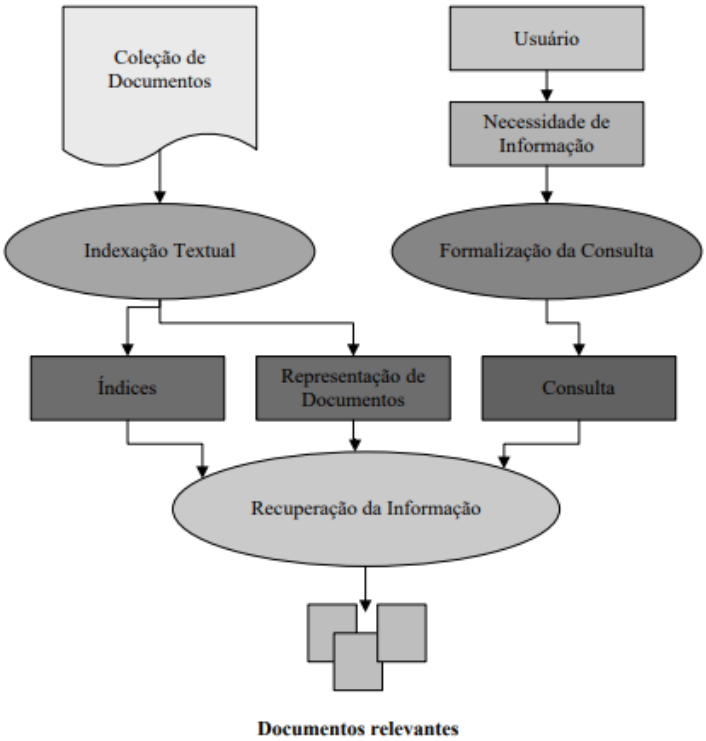
2.5.2 Aprendizado de máquina para organização da informação

Com aplicações diversas em muitas atividades e áreas de conhecimento, também é possível mencionar o uso do aprendizado de máquina nos processos de organização da informação. Sobre classificação, Khan et al. (2010) apresentam uma análise de alguns algoritmos utilizados para essa técnica de descrição conceitual da informacional, com foco naqueles de aprendizado supervisionado que é a abordagem mais utilizada para aplicações de classificação de dados textuais. Os autores realizaram testes utilizando *K-nearest neighborb (kNN)*, árvore de decisão, *Naive Bayes (NB)*, *Support Vector Machine (SVM)* e também redes neurais. Os resultados apontaram que o uso de técnicas híbridas, com utilização de mais de um algoritmo, mostram mais acurácia no processo de classificação automática de textos, em especial o uso de kNN, SVM e NB. Como ganhos e oportunidades da utilização de aprendizado supervisionado para a classificação de dados textuais não estruturados, os autores citam, dentre vários, o uso de semântica e ontologias para a classificação e recuperação da informação, aprimoramento de métodos que podem ser utilizados em outros processos de gestão da informação e descoberta de conhecimento por meio da extração de informações em grandes massas de documentos. Um estudo de Khadim (2019) sobre o desempenho de técnicas de aprendizagem de máquina para classificação corrobora os resultados de Khan et al. (2010), apontando que as abordagens híbridas apresentam resultados mais acertados e enfatiza também a importância do pré-processamento dos dados para que se obtenham resultados mais satisfatórios.

Além da classificação, o assunto aplicações automatizadas para organização da informação muito remete à recuperação da informação, área que, segundo Saracevic (1995), é o braço tecnológico da ciência da informação devido à sua grande intersecção com a ciência da computação. A recuperação da informação se preocupa com a representação, o armazenamento, a organização e o acesso a informações de forma a promover resultados relevantes sempre que um usuário realizar uma busca em algum sistema de informação (Baeza-Yates; Bertier, 2013). Para que a recuperação da informação seja realizada de forma eficiente e que os resultados sejam relevantes semanticamente, os processos de organização da informação que sustentam o sistema de informação em uso precisam ser realizados de forma consistente.

A figura abaixo ilustra a arquitetura de um sistema de recuperação de informação, partindo das duas partes principais que justificam sua existência que são os documentos e os usuários:

Figura 7 - Arquitetura de sistema de recuperação da informação



Fonte: PUC RIO (2017)

Uma das etapas ilustradas na figura é a chave para que os documentos sejam recuperados em um sistema: a indexação. Consoante a forte presença da tecnologia em sistemas de informação, mencionada por Saracevic (1995), encontra-se na literatura exemplos de aplicação de diversas técnicas computacionais para automatização da indexação, incluindo técnicas de aprendizagem de máquina. A próxima subsubseção apresenta exemplos de utilização das abordagens de ML na indexação.

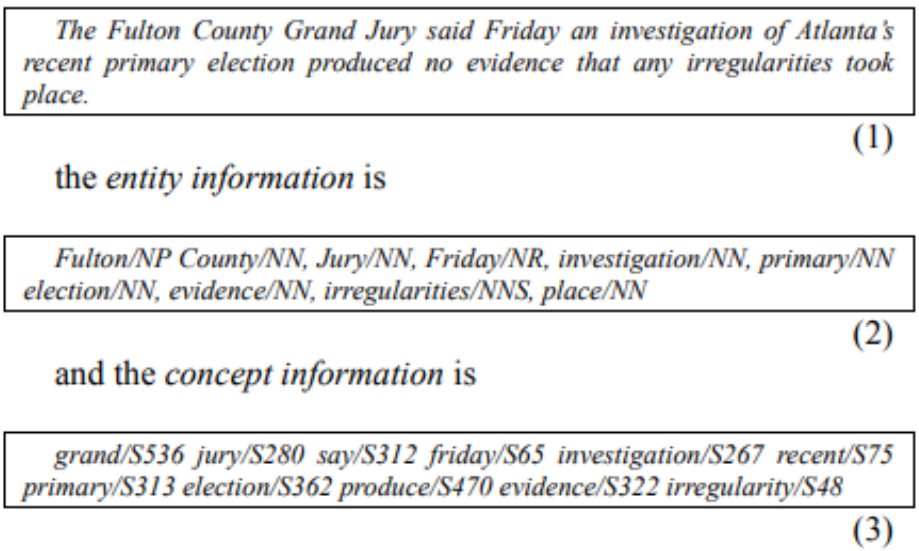
2.5.3 Estudos de caso de indexação

A união de algoritmos de ML aos princípios de organização e recuperação da informação estão documentados em trabalhos diversos na literatura. O avanço das tecnologias em nível semântico de processamento de dados permite novas abordagens aos sistemas de indexação automática de documentos. Setch e Tang (2007) descreveram sua experiência com aprendizagem de máquina supervisionado ao criarem uma ontologia e um algoritmo para

indexação de um *corpus* de documentos anotados. O algoritmo utilizado pelos autores adiciona informações de uma ontologia em termos e frases retiradas dos textos do *corpus*, o que permite que os documentos sejam pesquisados de forma integral a nível de palavras, além de analisados em diversos níveis de abstração. A abordagem utilizada para realizar a indexação considera classes gramaticais, possui sensibilidade para termos ambíguos, sendo capaz de minimizar a ambiguidade da palavra ao atribuir termos específicos vindos da ontologia em uso.

O processo de indexação adotado é o conceitual, onde se aplica um nível de abstração de ideias aos trechos dos textos, definido pelos autores como um índice de entidades e conceitos contidos em coleções de documentos que seja compreensível por máquina (Setch; Tang, 2007). Para a aplicação deste conceito, segue-se a seguinte ordem de preparação dos dados: (1) extração de entidades de conteúdo baseado em texto não estruturado usando *tags* lexicais; (2) identificação de conceitos e adição de *tags* de ontologia a eles usando regras semânticas e (3) fusão de informações de entidade e conceito criando um índice conceitual. Esta sequência é exemplificada no esquema a seguir:

Figura 8 - Etapas de preparação dos dados para indexação automática conceitual



Fonte: Setch e Tang (2007).

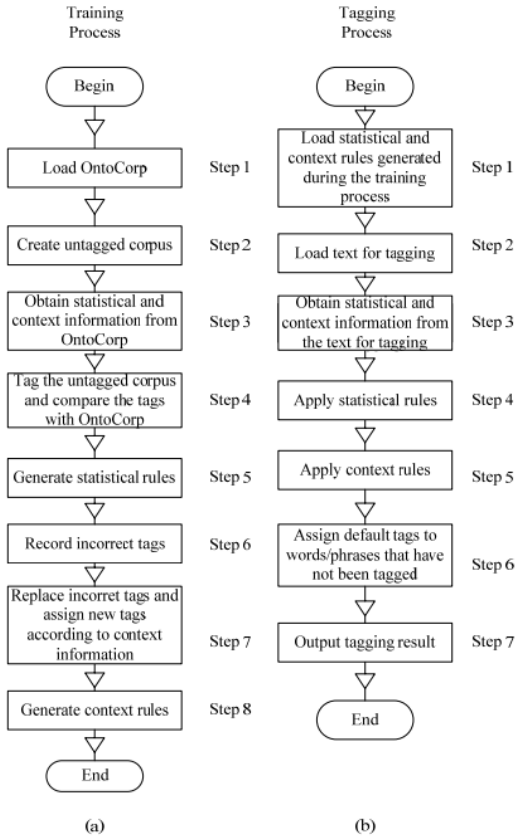
Em (1) está um trecho retirado de um dos textos do *corpus* utilizado, em sua forma não estruturada. Em (2) temos termos classificados como entidades, cada um rotulado com sua devida classe gramatical derivada de uma lista. Na imagem temos, por exemplo, NP que significa *proper noun* (nome próprio, em português), NN que significa *noun singular* (substantivo singular) e NNS, *noun plural* (substantivo plural). A classificação gramatical é

realizada automaticamente por meio do uso do Brill Tagger[5], uma metodologia para aplicar regras léxicas em corpus textuais. Na etapa (3) temos a atribuição de *tags* vindas da ontologia utilizada no estudo para os termos identificados na etapa anterior. Por exemplo, ao termo investigation foi atribuída a tag S267 que possui outros 672 termos relacionados à palavra investigação, tanto sinônimos gramaticais como indagação ou questionamento, como termos que em contextos semânticos se relacionam com o termo investigação, como detetive ou *CIA*. A ontologia foi construída a partir do Projeto Gutenberg[6] e conta com mais de 68 mil palavras.

Após passar pelo processo descrito acima, o corpus de termos foi utilizado para treinar o algoritmo de atribuição de *tags*. Ao optar pela utilização de um corpus textual anotado com outro tipo de ontologia e *tags*, os autores precisaram mapear a ontologia criada com referência à *WordNet*[7], uma base de dados léxicos que agrupa palavras com base em suas classes gramaticais e também com base em semântica. Passada a etapa de mapear e encontrar as *tags* equivalentes entre as duas bases ontológicas, seguiu-se para as etapas de treinamento e de atribuição de *tags*. O processo de trabalho do algoritmo está ilustrado na imagem abaixo:

[5] Metodologia desenvolvida pelo pesquisador Eric Brill como parte de sua tese de doutorado. (E. Brill, "A simple rule-based part of speech tagger", Proc. 3rd Conf. on Applied NLP, Trento, Italy, 1992)
[6] Disponível em: https://www.gutenberg.org/
[7] Disponível em: https://wordnet.princeton.edu/

Figura 9 - Fluxos de trabalho dos algoritmos de aprendizado de máquina utilizados no estudo



Fonte: Setch e Tang (2007)

O processo de treinamento começa na análise do corpus criado para o estudo de modo a se obter informações estatísticas e do contexto semântico dos textos. Depois remove-se as *tags* encontradas criando assim um corpus não rotulado, ou não anotado. Utilizando as informações obtidas na primeira etapa, são criadas novas *tags* que serão comparadas às *tags* originais do corpus inicial. Após isso, é possível estimar a precisão do algoritmo em treino. Segue-se para a fase de seleção das *tags* mais utilizadas para cada termo do corpus, depois uma análise para identificação das *tags* utilizadas incorretamente em referência a comparação do corpus não anotado ao corpus anotado. Mais adiante, é possível gerar *tags* referentes ao contexto semântico para marcar palavras ou frases que não haviam sido marcadas nas etapas anteriores. O processo de *tagging*, ou marcação, utiliza as regras de contexto geradas no processo de treinamento como ponto inicial. Em seguida os textos do corpus são analisados para que se obtenha informações estatísticas, assim como na etapa inicial do treinamento. Unindo todas as informações, o algoritmo atribui as *tags* vindas da ontologia utilizada. Segue-se para a verificação e atribuição de *tags* para termos que ocasionalmente não foram marcados, gerando assim o índice conceitual como resultado. Os autores apontam uma precisão de 78,9% do algoritmo, com espaço para futuras melhorias.

Em outro trabalho, temos a indexação semântica como apoio para a classificação de textos curtos (SMS) na categoria spam. Silva et al. (2017) utilizaram técnicas de indexação para atribuir contexto semântico aos textos de mensagens instantâneas, que têm como características principais a linguagem informal, uso de gírias e *emojis*. Ao indexar o texto atribuindo termos de uma ontologia, o algoritmo de aprendizado de máquina utilizado no estudo conseguiu uma melhor performance nos processos de classificação e predição, entregando mais acurácia ao dizer se um SMS é ou não um spam para os usuários. Além dos *softwares* de indexação automática citados na seção 2.2.1 deste documento, cabe trazer também o *software* MAUI[8], acrônimo para *Multi-purpose Automatic Topic Indexing*. O *software* desenvolvido por Alyona Medelyan em sua tese de doutorado, além de realizar indexação automática de documentos, possui a especificidade de utilizar aprendizagem de máquina na geração de índices, com funções de extração de termos e atribuição de termos provenientes de vocabulários controlados, marcação de termos e indexação de termos. O modelo de aprendizagem de máquina do MAUI o aproxima da indexação manual, visto que os termos utilizados para treinamento são coletados em uma base na qual os usuários podem avaliar se os termos (chamados pelo autor de termos candidatos) condizem ou não com as buscas realizadas. A partir desses termos já rotulados por humanos, o modelo aplica cálculos estáticos e linguísticos em uma abordagem de árvore de decisão para atribuir termos, levando em consideração critérios como: frequência de ocorrência do termo, posição de ocorrência do termo, o comprimento do termo em palavras, a probabilidade de um termo ser palavra-chave no domínio ou corpus e também as relações semânticas dos termos em um tesauro.

[8] Disponível em: https://github.com/zelandiya/maui

3 METODOLOGIA

Nesta seção apresenta-se os procedimentos metodológicos a serem realizados para alicerçar o desenvolvimento da pesquisa. Primeiramente, apresenta-se a caracterização da pesquisa, levando em consideração a concepção filosófica, bem como a classificação da abordagem e método do estudo. Em seguida, está elucidada a etapa de coleta de dados e a previsão inicial de tratamento. Por fim, são enumerados os procedimentos metodológicos relacionados a cada um dos objetivos propostos para a pesquisa.

3.1 Caracterização da pesquisa

Segundo Creswell e Creswell (2021), o planejamento metodológico de uma pesquisa deve considerar a definição de três eixos: a perspectiva filosófica, a abordagem de pesquisa e a definição do tipo de método. A perspectiva filosófica se refere às crenças básicas e ideias mais abrangentes que orientam as ações executadas durante a investigação científica, impactando também os métodos adotados. Além disso, Creswell e Creswell (2021) também apontam que a definição da perspectiva filosófica serve como apoio para identificação da abordagem de pesquisa, que pode ser qualitativa, quantitativa ou mista. Isso se deve ao fato de a perspectiva filosófica oferecer uma orientação ampla sobre a natureza e universo relacionados à pesquisa.

O presente estudo se caracteriza pela concepção pragmática, visto que tem foco em solucionar de forma prática o problema de pesquisa suscitado. Para Creswell e Creswell (2021), a pesquisa com esse tipo de concepção assume que o problema é mais importante e a pesquisa deve usar de todos os meios possíveis para entender o problema. Algumas alegações de conhecimento advindas do pragmatismo é a adoção de métodos mistos de pesquisa, se valendo de práticas quantitativas e também qualitativas para o levantamento de soluções, bem como a flexibilidade de adaptação de métodos diversos, de forma a alcançar o objetivo de um estudo. Além disso, corroborando o uso de métodos mistos, também a coleta e análise de dados utilizada nesse tipo de concepção pode se valer de formas mistas, assim como o uso de dados qualitativos e quantitativos (Creswell; Creswell, 2021).

Em relação à abordagem de pesquisa, o estudo adota a estratégia qualitativa, em específico, a estratégia de Estudo de caso, que é aquele em que o pesquisador explora em profundidade um programa, um fato, uma atividade, um processo ou uma ou mais pessoas (Creswell; Creswell, 2021). Sobre a estratégia de estudo de caso, Yin (2001) destaca que é a opção mais escolhida ao lidar com questões que envolvem o "como" e o "por quê", indo ao encontro da pergunta de pesquisa norteadora deste estudo. Cabe citar que é possível também

notar alguns aspectos quantitativos, na medida em que o estudo se propõe a experimentar uma técnica que poderá depois ser replicada em outros conjuntos de dados de diferentes escopos informacionais, por exemplo. Sendo assim, corroborando com a concepção filosófica adotada, a abordagem de pesquisa pode ser caracterizada como qualitativa com incorporação também de características quantitativas.

Quanto ao método de pesquisa, o presente estudo se vale predominantemente do método qualitativo. Entretanto, se conectando com a concepção filosófica e abordagem adotadas, também faz uso de aspectos quantitativos, fazendo válido caracterizar os métodos adotados como mistos. Sampieri, Collado e Lucio (2013) destacam a flexibilidade inerente ao método misto, permitindo que os pesquisadores se adaptem dinamicamente às necessidades específicas de suas investigações, o que se faz extremamente importante no contexto do estudo de caso aqui adotado, já que a investigação pretende explorar aspectos conceituais e também práticos ligados à indexação. Os procedimentos metodológicos adotados serão descritos nas próximas subseções.

3.2 Coleta e tratamento inicial dos dados

A coleta de dados é compreendida como a fase da pesquisa destinada ao conhecimento das características de uma determinada população ou amostra. Para tanto, a coleta de dados deve ser devidamente planejada de acordo com os objetivos da pesquisa (Barbetta, 2005).

3.2.1 Coleta

Para a realização da pesquisa faz-se necessária a utilização de dados textuais, que são registros em formato não estruturado (como os binários, por exemplo), a serem coletados aqui em formato de textos. A especificidade dos textos a serem utilizados é o requisito de possuírem marcações, os grifos, realizados por pesquisadores durante suas atividades de pesquisa e estudo. A escolha de utilizar um corpus criado a partir de textos utilizados em pesquisa é baseada justamente no uso prévio dos arquivos de texto, os quais já terão registros de interação dos usuários, como destaques e anotações, que serão utilizados como o recurso informacional primário para o processo de indexação proposto por esta pesquisa.

A coleta dos dados foi realizada por meio de formulário *online* construído na plataforma *Google Forms*, desenvolvido com os seguintes objetivos:

- Levantar dados de caracterização do perfil de leitura e pesquisa dos respondentes;

- Investigar a motivação dos leitores ao realizar as marcações em seus textos; e
- Coletar textos lidos e grifados pelos respondentes.

Para atender os objetivos da coleta de dados, a estrutura adotada no formulário foi aberta, contando com questões de respostas livres (exceto uma única questão fechada), o que reforça também o aspecto de estudo qualitativo que perpassa toda a metodologia aqui apresentada. O formulário foi elaborado com poucas questões, de forma a não apresentar um tempo elevado para respondê-lo, visto que o envio de documentos em si já é um processo demorado, pois envolve recuperação de documentos por parte do respondente em suas plataformas pessoais de armazenamento de informações. Sendo assim, para maximizar a possibilidade de receber uma quantidade significativa de respostas, o formulário contou com a estrutura detalhada no Quadro 2:

Quadro 2 - Sistematização das questões do formulário

Pergunta	Tipo de pergunta	Obrigatoriedade
Idade	Aberta	Obrigatória
Campo de pesquisa (descreva em palavras chave)	Aberta	Obrigatória
Dispositivo mais utilizado para leituras	Fechada	Obrigatória
Você costuma grifar textos durante a leitura? Se sim, quais são os critérios para escolher os trechos a serem grifados?	Aberta	Obrigatória
Compartilhe os documentos lidos e marcados por você	Aberta (upload de arquivos com limite de 10 por resposta[9])	Obrigatória

Fonte: elaborado pela autora

A aplicação do formulário foi direcionada a pesquisadores e alunos de pós graduação com foco em Ciência da Informação, tendo sido divulgado no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e no Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília (PPGCINF - UnB) contando com o prazo de 3 semanas para a coleta das respostas. No âmbito do Ibict, o formulário foi enviado para a lista total de colaboradores por meio da lista de *e-mail* institucional do órgão. Já no PPGCINF, o formulário foi enviado pela secretaria do programa também para a lista geral de contatos. Sendo assim, o

[9] Limitação da plataforma *Google Forms*.

universo da pesquisa se caracteriza como múltiplo, visto que considera dois grupos diferentes para a coleta de informações, dada a divulgação do instrumento de coleta em duas instituições distintas. A escolha por esse tipo de universo se deu com o intuito também de maximizar a quantidade de respostas recebidas, para que fosse possível coletar um número significativo de documentos para análise. Devido a essa escolha metodológica, uma limitação notada foi a de cálculo de amostra, já que não foi possível mensurar precisamente o universo trabalhado.

O formulário obteve 14 respostas de pessoas com idade entre 26 e 68 anos, com média de idade de 41 anos. Quanto às temáticas de pesquisa, os respondentes podiam incluir quantas quisessem, separando-as por ponto e vírgula no preenchimento do formulário. Foram identificadas 41 temáticas distintas, sendo 27 delas intimamente ligadas à Ciência da Informação e 14 outras relativamente distantes da CI, como geografia e neurociência, por exemplo. A Figura 9 traz uma nuvem de palavras, criada com a ferramenta Voyant Tools[10], com as respostas coletadas na questão sobre a área de pesquisa do formulário aplicado.

Figura 10 - Nuvem de palavras com os termos coletados na questão sobre áreas de pesquisa

Fonte: elaborado pela autora

Já na questão sobre os dispositivos utilizados para leitura de textos, os respondentes podiam escolher entre as opções: computador (*desktop*), *notebook*, *tablet*, celular e leitor digital (*Kindle* ou semelhantes). As respostas mostraram que a maioria, 8 entre 14 respondentes, utilizam o *desktop* para a leitura e marcação de seus textos no contexto das atividades de

[10] Disponível em: https://voyant-tools.org/

pesquisa. Outros 5 responderam que utilizam *notebook* e 1 pessoa respondeu que utiliza o celular para suas leituras. A questão para compartilhamento de textos coletou 55 arquivos em formato PDF, compondo assim o mini corpus a ser analisado na proposta de indexação desta pesquisa. As respostas abertas sobre os motivos para grifar os textos serão destrinchadas nas próximas subseções do trabalho.

3.2.2 Tratamento e análise preliminares

Visto que os dados utilizados para o processo de indexação serão do tipo textual, o tratamento previsto envolve técnicas de mineração de texto, um processo de análise textual envolvendo algoritmos computacionais para extração de informações, termos ou palavras em documentos de formato digital (Salton; Mcgill, 1983). A particularidade da extração de termos adotada para essa pesquisa será considerar apenas os destaques feitos por usuários nos documentos como termos aptos à extração.

Além da extração dos termos, a etapa de tratamento dos dados também inclui a normalização dos termos. O processo de normalização de termos extraídos por meio do uso de algoritmos computacionais foi descrito e ilustrado por Morais e Ambrósio (2007):

Figura 11 - Etapas de normalização para indexação automática

Fonte: Morais e Ambrósio (2007)

A seguir, cada etapa do processo de normalização será descrita em detalhes, também com base na obra de Morais e Ambrósio (2007):

a) **Identificação de termos:** compreende a identificação de termos simples e complexos. Para identificar termos simples, utiliza-se a aplicação de um algoritmo de análise léxica para identificar as palavras e eliminar símbolos e demais caracteres que não apresentem valor semântico para o documento. Também é nessa etapa onde se verifica erros ortográficos, define-se sinônimos para termos específicos ou se determina outras alterações na grafia dos termos, como definir todos os termos para letras maiúsculas ou minúsculas, por exemplo. Já a etapa de identificação de termos complexos serve para identificar palavras que quando usadas juntas possuam significado semântico importante no contexto do documento ou do conjunto de documentos analisados. Pode

ser realizada de duas maneiras: A primeira envolve a identificação de termos que ocorrem com muita frequência em uma coleção de documentos, podendo passar por validação do usuário para que decida se o que foi encontrado pelo algoritmo faz sentido para ele. A segunda consiste na utilização de um dicionário de expressões. É importante citar que termos complexos impactam a busca numa posterior recuperação da informação, já que os termos devem ser consultados sempre em sua forma complexa registrada.

b) Remoção de *stopwords*: esta etapa tem como foco a eliminação de algumas palavras que não devem ser consideradas no documento, conhecidas como *stopwords*, que são palavras consideradas não relevantes na análise de textos por não agregarem conteúdo semântico e nem serem termos que se relacionam à temática do documento analisado. Normalmente fazem parte desta lista as preposições, pronomes, artigos, advérbios, e outras classes de palavras auxiliares. Além dessas, existem também palavras cuja frequência na coleção de documentos é muito alta e por isso perdem sua capacidade de descrever tematicamente um documento dentro de uma coleção. O conjunto de *stopwords* identificadas formam uma lista de palavras conhecida como *stoplist*, que dificilmente são utilizadas em uma consulta, além de aumentarem o tamanho do índice em construção para além do necessário.

c) **Normalização morfológica:** nesta etapa se realiza a eliminação das variações morfológicas das palavras extraídas, eliminação que se dá por meio da identificação dos radicais das palavras. Essa técnica de identificação de radicais é denominada lematização ou *stemming*, que em inglês significa reduzir uma palavra ao seu radical (ou raiz). Feita por meio de algoritmos específicos, a lematização inclui a eliminação dos prefixos e sufixos, características de gênero, número e grau das palavras, o que na prática leva à redução de diversas palavras a uma única, em muitos casos. Por acarretar em perda de precisão nas buscas no momento de recuperação da informação, esta é uma etapa que pode não ser realizada neste estudo, caso se julgue adequado ao longo da normalização dos dados.

A etapa de extração dos termos e remoção de *stopwords* foi realizada por meio de código computacional criado com a linguagem Python. A linguagem foi escolhida por ser amplamente utilizada em projetos de pesquisa e também por apresentar vantagens de uso em relação a outras linguagens similares, como JAVA, por exemplo. Alguns dos benefícios proporcionados pelo Python, em especial no contexto de projetos de pesquisa, são:

- **Sintaxe clara e legível:** A sintaxe clara e legível do Python facilita a compreensão do código, tornando-o acessível para estudantes e pesquisadores que podem não ter uma experiência extensiva em programação;
- **Ampla variedade de bibliotecas:** Python possui uma vasta coleção de bibliotecas e *frameworks* que cobrem uma ampla gama de disciplinas acadêmicas, incluindo ciência de dados, aprendizado de máquina, processamento de linguagem natural, matemática, física, entre outras. Isso permite aos acadêmicos realizar análises e implementar algoritmos de forma eficiente, aproveitando as contribuições da comunidade;
- **Comunidade ativa:** Python tem uma comunidade ativa e engajada, o que significa que é fácil encontrar suporte *online*, tutoriais, e soluções para desafios específicos. Isso é especialmente útil para estudantes que estão aprendendo a programar ou que estão explorando novas áreas de pesquisa;
- **Documentação abundante:** A documentação extensiva e recursos educacionais disponíveis para Python facilitam a aprendizagem e a referência durante o desenvolvimento de projetos acadêmicos;
- **Open Source e gratuito:** Python é uma linguagem de programação de código aberto e gratuita, o que elimina barreiras financeiras e permite que estudantes e pesquisadores acessem facilmente a linguagem e suas bibliotecas, o que é importante no presente estudo visto que se pretende disponibilizar uma proposta de fluxo de trabalho que poderá ser implementada em outros cenários e conjuntos documentais.

O relato de uma extração com o mesmo critério de selecionar os trechos grifados em textos foi disponibilizado por Sarkari (2023), sendo aproveitado como base para o presente estudo. O código utilizado no relato utiliza a biblioteca PyMuPDF[11], uma biblioteca Python de alto desempenho para extração, análise, conversão e manipulação de documentos em PDF, com boa documentação e tutoriais de uso. O processo de extração se dá em duas etapas: primeiro é realizada uma varredura no documento com o objetivo de identificar cores, em padrão RGB, de trechos grifados no texto. Após a varredura, o código retorna os padrões de cores encontrados e segue-se para a próxima etapa, que é, de fato, a extração dos termos que foram grifados utilizando os padrões de cores identificados. Os trechos grifados são coletados e disponibilizados em um arquivo individual, um por cada cor identificada na varredura, o que traz a possibilidade de que um mesmo texto apresente mais de um conjunto de termos grifados.

[11] Disponível em: https://pymupdf.readthedocs.io/en/latest/

A Figura 12 mostra o código de checagem de cores presentes nos documentos, nomeado como *checkcolor.py*. Na linha 4, no parâmetro *filepath* foram adicionados os caminhos dos textos coletados no formulário, para que pudessem ser checados. Ao final, nas linhas 24 e 25, estão dois exemplos de código RGB retornados como resposta da checagem feita pelo código apresentado.

Figura 12 - Código de checagem de cor

```
1  import fitz  # PyMuPDF
2
3  # Open the PDF
4  doc = fitz.open('/filepath')
5
6  # Set to store unique colors
7  unique_colors = set()
8
9  # Loop through every page
10 for i in range(len(doc)):
11     page = doc[i]
12     # Get the annotations (highlights are a type of annotation)
13     annotations = page.annots()
14     for annotation in annotations:
15         if annotation.type[1] == 'Highlight':
16             # Get the color of the highlight
17             color = annotation.colors['stroke']  # Returns a RGB tuple
18             unique_colors.add(color)
19
20 # Print all unique colors
21 for color in unique_colors:
22     print(color)
23
24 # Amarelo (1.0, 1.0, 0.0)
25 # vermelho (0.9960780143737793, 0.450980007648468, 0.1254899948835373)
```

Fonte: captura de tela elaborada pela autora

Alguns arquivos, por serem textos estruturados em *templates* com muitas cores, ou por apresentarem imagens e figuras também com muitas cores, retornaram erro no código de checagem de cor, sendo desconsiderados para as próximas etapas de trabalho. No total, 24 textos apresentaram erro de checagem, restando assim 31 textos aptos a compor o mini corpus a ser utilizado para as etapas de extração e indexação. Esta limitação do código utilizado pode ser aprimorada em utilizações futuras, substituindo a detecção dos termos grifados por meio de cor para mapeamento da localização de grifos no documento, por exemplo. O Quadro 3 sistematiza os textos utilizados nas próximas etapas da metodologia, apresentando um número de identificação (ID) e seu respectivo título.

Quadro 3 - Textos do corpus analisado

ID	Título do texto
1	Os Fenômenos de Interesse para a Ciência da Informação
2	A Ciência da Informação como Ciência Social
3	Information as Thing
4	Sobre a Ideia e a Ensino do Paisagem
5	Marcos Históricos da Ciência da Informação: Breve Cronologia dos Pioneiros, das Obras Clássicas e dos Eventos Fundamentais
6	The Invisible Substrate of Information Science
7	Más Condutas Científicas: Uma Análise em Políticas de Repositórios de Dados
8	Praxeological Sociology of Knowledge and Documentary Method: Karl Mannheim's Framing of Empirical Research
9	O Papel da Informação Tecnológica: As Redes de Informação
10	Contextualização Histórica do Conceito de Paisagem, suas Implicações Filosóficas e Científicas
11	Sinais: Raízes de um Paradigma Indiciário
12	Dez Mandamentos para Bons Repositórios de Dados de Pesquisa
13	Metodologia de Pesquisa no Campo da Ciência da Informação
14	A Contribuição da História dos Conceitos à Ciência da Informação: Dimensões Categórico-Abstratas e Analítico-Causais
15	Museologia: Entre Abandono e Destino
16	A Comunicação Científica
17	Da Comunicação Extensiva ao Hibridismo da Animaverbivocovisualidade (AV3)
18	Autoria Coletiva, Autoria Ontológica e Intertextualidade na Ciência: Aspectos Interdisciplinares e Tecnológicos
19	The History and Historiography of Information Science: Some Reflections
20	Documento e Poder: Uma Arqueologia da Escrita
21	The Foundations of Information Science, Part I. Philosophical Aspects
22	Documents, Memory Institutions and Information Science

23	Ciência da Informação: Origem, Evolução e Relações
24	Documento e Significação na Trajetória Epistemológica da Ciência da Informação
25	Palestra Proferida pelo Prof. Dr. Bernd Frohmann na Abertura do Evento ENANCIB, 7, em Marília, SP em Outubro de 2006
26	Dados Bibliométricos e Altimétricos de Artigos Científicos sobre Inteligência Artificial: Análise do Impacto Acadêmico e Social
27	A Contribuição de Karl Mannheim para a Pesquisa Qualitativa: Aspectos Teóricos e Metodológicos
28	Correntes Teóricas da Ciência da Informação
29	Valores Sociais e Atividade Científica: um Retorno à Agenda de Robert Merton
30	O Processo de Atenção e o Letramento Informacional
31	Algumas Considerações sobre os Repositórios Digitais de Dados de Pesquisa

Fonte: elaborado pela autora

Já na Figura 13 é possível visualizar o código de extração dos termos a partir das cores identificadas, com os códigos RGB agora aparecendo como parâmetros dentro da sintaxe do código e com um exemplo de caminho do arquivo de texto na linha 5. Nos casos em que se identificou mais de duas cores de marcação por documento, os códigos RGB foram incorporados com os nomes de suas respectivas cores, de forma que se extraísse todos os termos grifados dentro do documento.

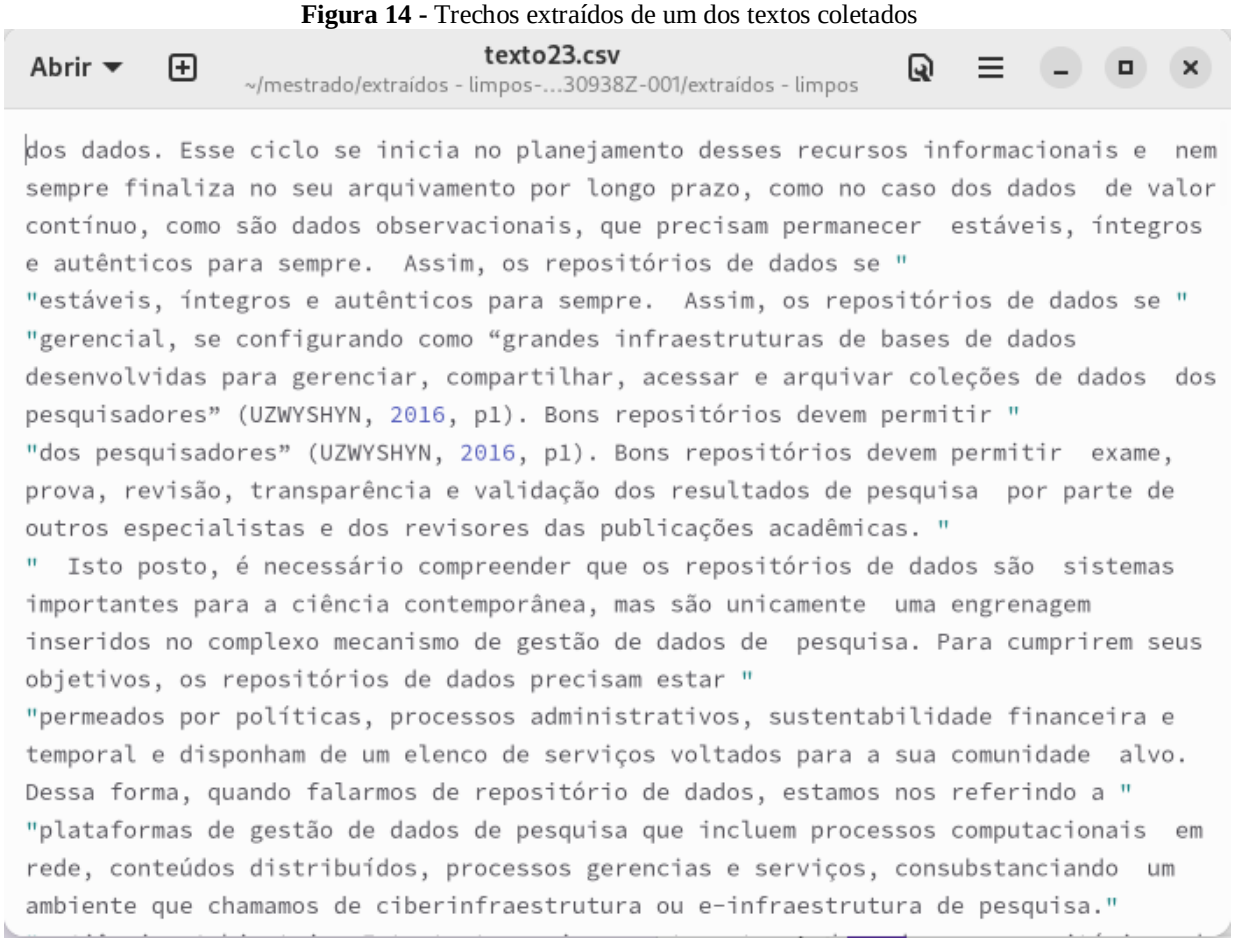
Figura 13 - Código de extração de termos grifados

```python
import fitz  # PyMuPDF
import pandas as pd

# Open the PDF
doc = fitz.open('/home/nathalyleite/mestrado/drive-download-20231030T141419Z-001/texto88.pdf')

# Define the RGB values for your colors
VERMELHO = (0.9960780143737793, 0.450980007648468, 0.1254899948835373)
AMARELO = (1.0, 1.0, 0.0)

color_definitions = {"Vermelho": VERMELHO, "Amarelo": AMARELO}

# Create separate lists for each color
data_by_color = {"Vermelho":[], "Amarelo":[]}

# Loop through every page
for i in range(len(doc)):
    page = doc[i]
    annotations = page.annots()
    for annotation in annotations:
        if annotation.type[1] == 'Highlight':
            color = annotation.colors['stroke']  # Returns a RGB tuple
            if color in color_definitions.values():
                # Get the detailed structure of the page
                structure = page.get_text("dict")

                # Extract highlighted text line by line
                content = []
                for block in structure["blocks"]:
                    for line in block["lines"]:
                        for span in line["spans"]:
                            r = fitz.Rect(span["bbox"])
                            if r.intersects(annotation.rect):
                                content.append(span["text"])

                content = " ".join(content)

                # Append the content to the appropriate color list
                for color_name, color_rgb in color_definitions.items():
                    if color == color_rgb:
                        data_by_color[color_name].append(content)

# Convert each list to a DataFrame and write to a separate .csv file
for color_name, data in data_by_color.items():
    if data:
        df = pd.DataFrame(data, columns=["Text"])
        df.to_csv(f'highlighted_text_{color_name.lower()}.csv', index=False)
```

Fonte: captura de tela elaborada pela autora

O resultado da aplicação do código de extração é um arquivo de texto reunindo todos os trechos grifados, separados por cor, que posteriormente foram reunidos em um arquivo geral por documento, visto que a divisão por cores de grifos não será considerada para o processo de indexação. O arquivo textual de um dos documentos está apresentado na Figura 14, onde nota-se a estrutura em que os trechos foram extraídos, respeitando a forma como estavam estruturados no documento de origem. Na figura o texto ainda conta com sinais gráficos e todas as palavras extraídas.

Figura 14 - Trechos extraídos de um dos textos coletados



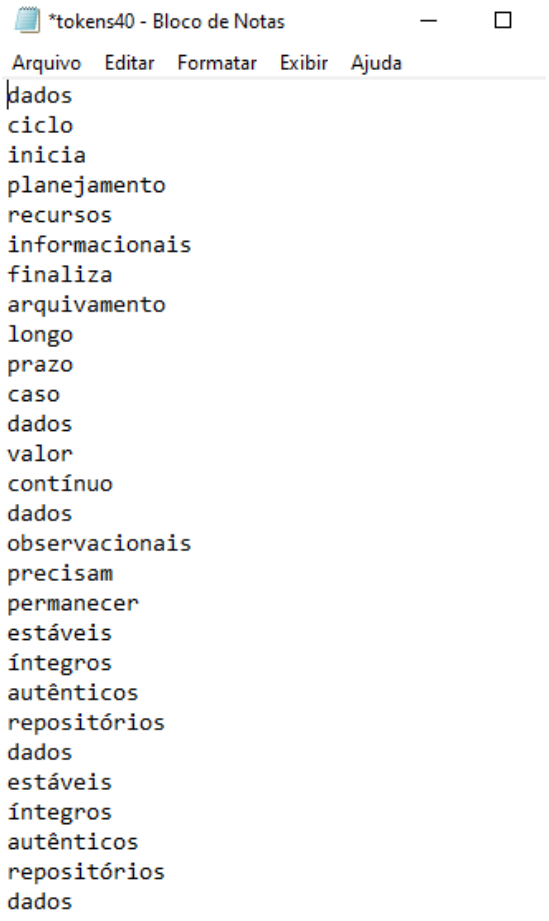
Fonte: captura de tela de dados da pesquisa

Após a extração de termos em todos os documentos do mini corpus, seguiu-se para a etapa de remoção das *stopwords*. Aqui também foi realizada a remoção de sinais gráficos como parênteses, aspas, pontos e vírgulas. Para esta etapa foi utilizada como suporte a biblioteca *Natural Language Toolkit* (NLTK)[12], uma plataforma robusta que serve como ferramenta para criar códigos em *Python* para processamento de dados em linguagem natural. Também foi realizada pesquisa no repositório *GitHub* com o objetivo de selecionar conjuntos completos de palavras nas línguas portuguesa, inglesa e espanhola a fim de utilizá-las no código de remoção. Os conjuntos selecionados estão descritos nos Anexos A, B e C.

Juntamente ao processo de remoção das *stopwords* e sinais gráficos foi realizado também o processo de tokenização das palavras, transformando os trechos de textos extraídos em termos únicos, em um arquivo onde cada palavra ocupa uma linha de forma individual. A Figura 15 traz um exemplo de arquivo com os termos em formato de *tokens*.

[12] Disponível em: https://www.nltk.org/

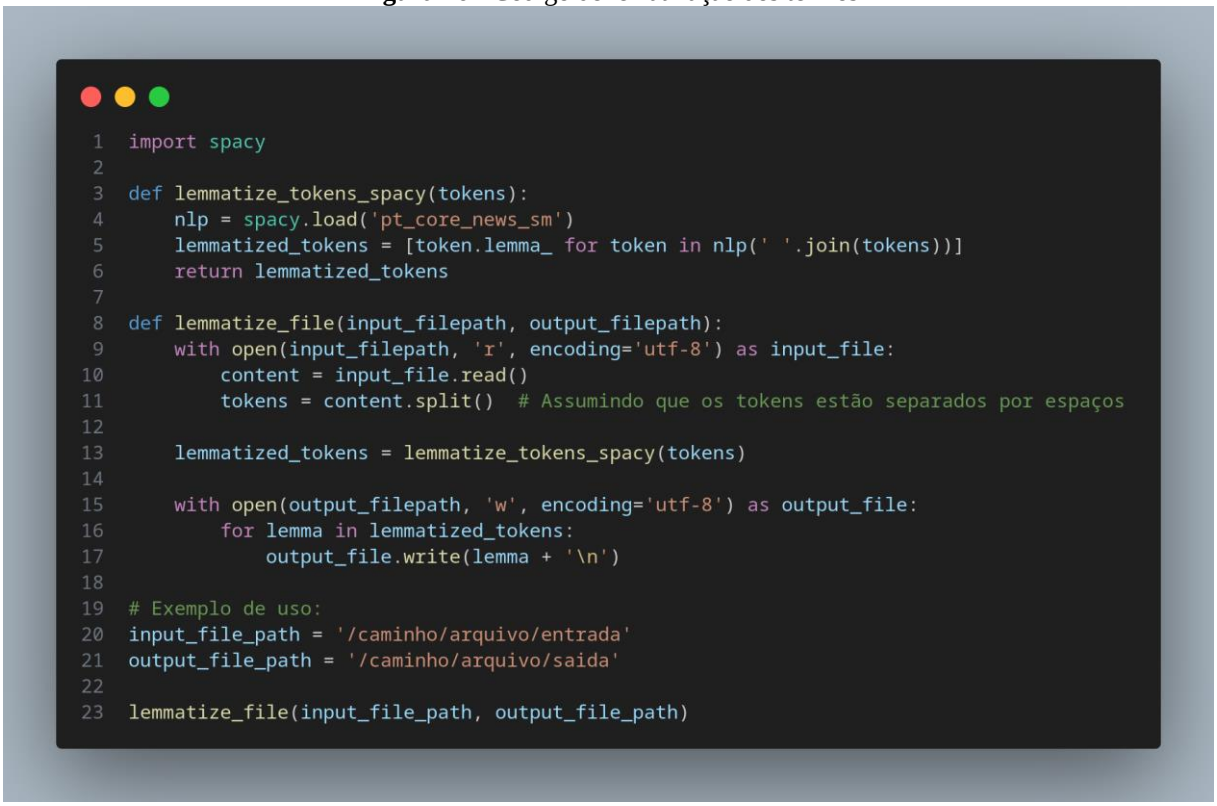
Figura 15 - Arquivo textual com termos em formato de token



Fonte: captura de tela de dados de pesquisa elaborada pela autora

Por fim, com os arquivos de textos com termos em formato de *tokens*, seguiu-se para o processo de lematização dos termos, com o objetivo de reduzir as palavras à sua forma radical, retirando plurais e outros tipos de variações que possam interferir no resultado da indexação. Assim, encerra-se o processo de tratamento preliminar dos dados, sendo possível avançar para a execução das etapas de trabalho relacionadas aos objetivos da pesquisa, descritas na próxima subseção. O código utilizado para a lematização está na Figura 16.

Figura 16 - Código de lematização dos termos



```python
import spacy

def lemmatize_tokens_spacy(tokens):
    nlp = spacy.load('pt_core_news_sm')
    lemmatized_tokens = [token.lemma_ for token in nlp(' '.join(tokens))]
    return lemmatized_tokens

def lemmatize_file(input_filepath, output_filepath):
    with open(input_filepath, 'r', encoding='utf-8') as input_file:
        content = input_file.read()
        tokens = content.split()  # Assumindo que os tokens estão separados por espaços

    lemmatized_tokens = lemmatize_tokens_spacy(tokens)

    with open(output_filepath, 'w', encoding='utf-8') as output_file:
        for lemma in lemmatized_tokens:
            output_file.write(lemma + '\n')

# Exemplo de uso:
input_file_path = '/caminho/arquivo/entrada'
output_file_path = '/caminho/arquivo/saida'

lemmatize_file(input_file_path, output_file_path)
```

Fonte: elaborado pela autora

3.3 Procedimentos metodológicos relacionados aos objetivos da pesquisa

As atividades previstas e propostas a seguir são consideradas a partir da análise do conjunto de dados resultantes das etapas descritas na seção 3.2 Coleta e tratamento inicial dos dados.

A) Compreender a relação entre as anotações em documentos digitais e a organização da informação dos participantes da pesquisa

Com caráter qualitativo, foi realizada análise das respostas coletadas no formulário da questão aberta com o seguinte enunciado: “Você costuma grifar textos durante a leitura? Se sim, quais são os critérios para escolher os trechos a serem grifados?”. O objetivo da análise foi identificar os processos utilizados pelos leitores no momento da leitura em relação aos grifos para aproximá-los de processos de organização da informação praticados e estudados no âmbito da CI, de modo a aprofundar a compreensão já elucidada no referencial teórico sobre a relação entre os dois aspectos. A análise focou na motivação para realizar os grifos, bem como no modo prático relatado pelos respondentes para a realização dos grifos. Para isso, as respostas foram estruturadas em uma lista com os tópicos identificados em cada uma, de forma que se pudesse encontrar as similaridades e pontuar também as diferenças entre os processos de cada

respondente. Cada tópico será discutido, buscando a aproximação com os conceitos da CI e da OI. Ao final, buscou-se discutir de forma geral as respostas como um todo.

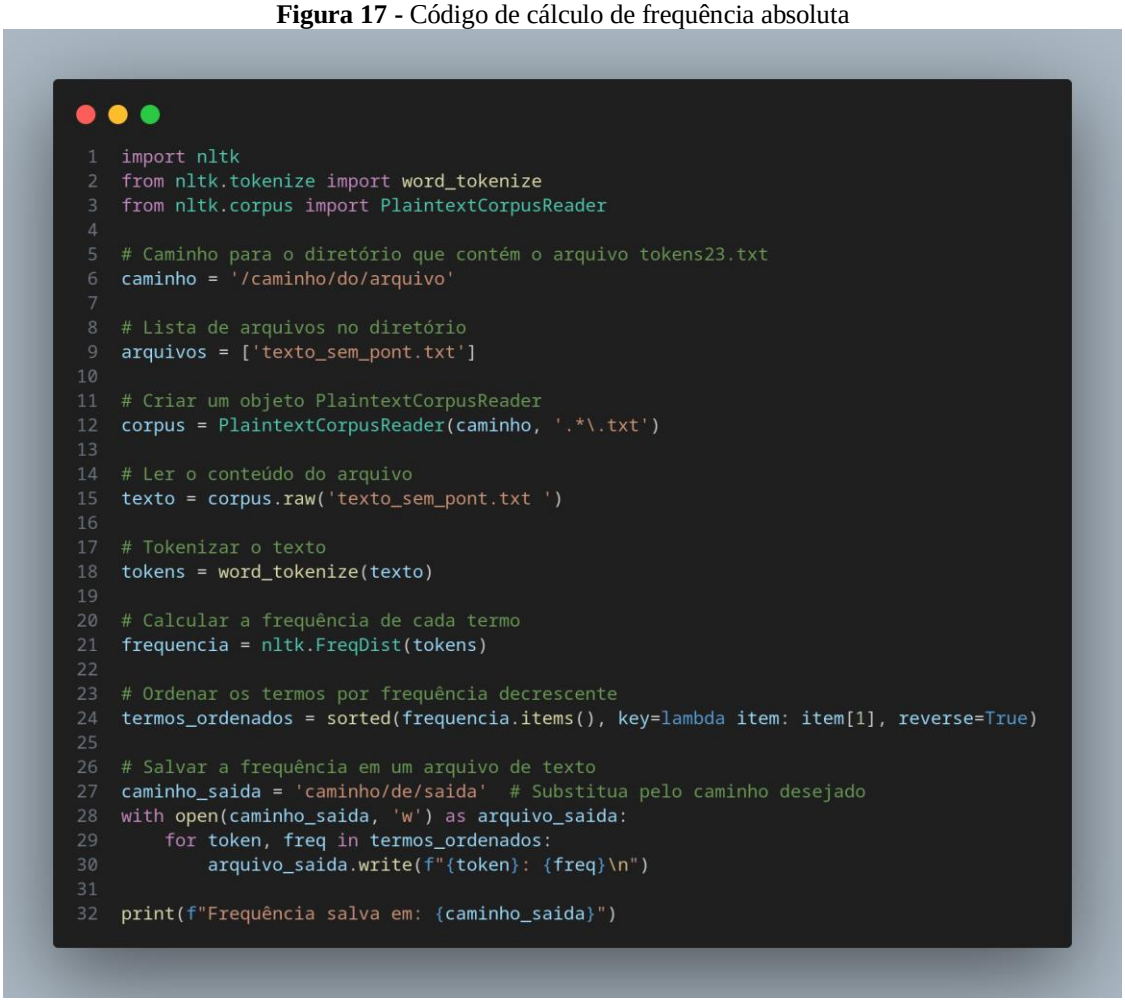
B) Realizar a indexação automática do corpus coletado

Nesta etapa se objetivou testar a indexação automática de extração a partir dos termos advindos dos grifos em documentos. Aqui, assim como na etapa de processamento de dados, utilizou-se o Python como linguagem principal de suporte para o código. Já para as funções específicas realizadas pelo código, utilizou-se a biblioteca SciKit Learn[13], uma biblioteca em código aberto para aprendizado de máquina em Python sendo amplamente utilizada para tarefas de mineração de dados e análise de dados, o que atende as necessidades do presente estudo.

A primeira tarefa realizada nesta etapa foi identificar a frequência de cada termo em cada conjunto extraído dos documentos, de forma individual. Para isso, utilizou-se um código que lê o documento de *tokens* e cria como resultado um arquivo contendo os termos organizados em forma decrescente de acordo com a quantidade de vezes que aquele termo aparece no documento, onde cada termo vem acompanhado também dessa frequência absoluta. A Figura 17 mostra o código utilizado.

[13] Disponível em: https://scikit-learn.org/stable/

Figura 17 - Código de cálculo de frequência absoluta



```python
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import PlaintextCorpusReader

# Caminho para o diretório que contém o arquivo tokens23.txt
caminho = '/caminho/do/arquivo'

# Lista de arquivos no diretório
arquivos = ['texto_sem_pont.txt']

# Criar um objeto PlaintextCorpusReader
corpus = PlaintextCorpusReader(caminho, '.*\.txt')

# Ler o conteúdo do arquivo
texto = corpus.raw('texto_sem_pont.txt ')

# Tokenizar o texto
tokens = word_tokenize(texto)

# Calcular a frequência de cada termo
frequencia = nltk.FreqDist(tokens)

# Ordenar os termos por frequência decrescente
termos_ordenados = sorted(frequencia.items(), key=lambda item: item[1], reverse=True)

# Salvar a frequência em um arquivo de texto
caminho_saida = 'caminho/de/saida'  # Substitua pelo caminho desejado
with open(caminho_saida, 'w') as arquivo_saida:
    for token, freq in termos_ordenados:
        arquivo_saida.write(f"{token}: {freq}\n")

print(f"Frequência salva em: {caminho_saida}")
```

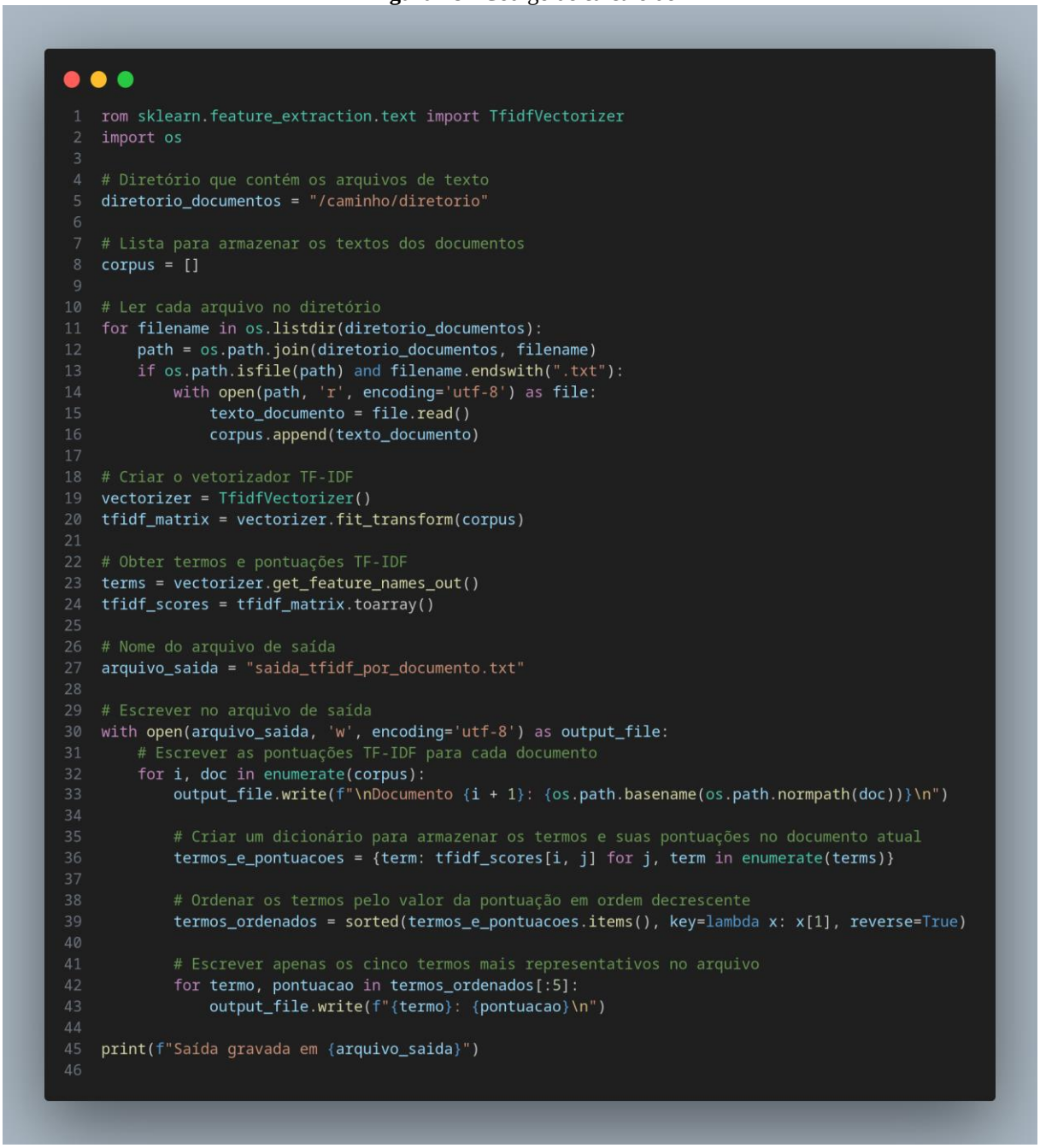
Fonte: elaborado pela autora

Além da frequência absoluta, também foi aplicado um código para realizar o cálculo TF-IDF (*Term Frequency - Inverse Document Frequency*), um cálculo amplamente utilizado em práticas de indexação automática que determina a frequência relativa das palavras em um documento específico em comparação com a proporção inversa dessa palavra em todo o corpus de documentos, ajudando assim a determinar quão relevante uma palavra específica é em um documento particular (Ramos, 2003). A fórmula de cálculo é expressa da seguinte maneira:

$$\square\square\square,\square = \frac{\text{frequência do termo } \square \text{ no documento } \square}{\text{total de termos no documento}}$$

O código foi programado para aplicar o referido cálculo e retornar como resultado os três termos mais representativos de cada documento, acompanhados de sua frequência relativa em comparação ao corpus como um todo. A Figura 18 mostra o código utilizado.

Figura 18 - Código de cálculo do TF-IDF



```
rom sklearn.feature_extraction.text import TfidfVectorizer
import os

# Diretório que contém os arquivos de texto
diretorio_documentos = "/caminho/diretorio"

# Lista para armazenar os textos dos documentos
corpus = []

# Ler cada arquivo no diretório
for filename in os.listdir(diretorio_documentos):
    path = os.path.join(diretorio_documentos, filename)
    if os.path.isfile(path) and filename.endswith(".txt"):
        with open(path, 'r', encoding='utf-8') as file:
            texto_documento = file.read()
            corpus.append(texto_documento)

# Criar o vetorizador TF-IDF
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(corpus)

# Obter termos e pontuações TF-IDF
terms = vectorizer.get_feature_names_out()
tfidf_scores = tfidf_matrix.toarray()

# Nome do arquivo de saída
arquivo_saida = "saida_tfidf_por_documento.txt"

# Escrever no arquivo de saída
with open(arquivo_saida, 'w', encoding='utf-8') as output_file:
    # Escrever as pontuações TF-IDF para cada documento
    for i, doc in enumerate(corpus):
        output_file.write(f"\nDocumento {i + 1}: {os.path.basename(os.path.normpath(doc))}\n")

        # Criar um dicionário para armazenar os termos e suas pontuações no documento atual
        termos_e_pontuacoes = {term: tfidf_scores[i, j] for j, term in enumerate(terms)}

        # Ordenar os termos pelo valor da pontuação em ordem decrescente
        termos_ordenados = sorted(termos_e_pontuacoes.items(), key=lambda x: x[1], reverse=True)

        # Escrever apenas os cinco termos mais representativos no arquivo
        for termo, pontuacao in termos_ordenados[:5]:
            output_file.write(f"{termo}: {pontuacao}\n")

print(f"Saída gravada em {arquivo_saida}")

```

Fonte: elaborado pela autora

C) Propor um *workflow* de indexação de termos extraídos de anotações em documentos digitais

Esta etapa consiste em sistematizar de forma gráfica, em formato de fluxo de trabalho aplicável a outras corpora, os procedimentos adotados desde o tratamento preliminar dos dados até a conclusão do objetivo específico **Realizar a indexação automática do corpus coletado.** Utilizando a ferramenta de diagramação Visio[14], da Microsoft, foi criado um fluxograma com

[14] Disponível em: https://www.microsoft.com/pt-br/microsoft-365/visio/flowchart-software

todas as etapas descritas para que se consiga realizar a indexação a partir de termos dos grifos encontrados em documentos digitais, incluindo as especificações tecnológicas necessárias, tais como as bibliotecas de código, a linguagem computacional e os tipos de códigos utilizados no processo.

4 RESULTADOS E DISCUSSÃO

Os resultados obtidos nos estudos serão apresentados de acordo com os objetivos específicos determinados.

4.1 Compreender a relação entre as anotações em documentos digitais e a organização da informação dos participantes da pesquisa

Como mencionado no referencial teórico, as anotações e marcações em textos desempenham importante papel nos momentos de leitura e estudo, atividades intrínsecas da pesquisa. Ao analisar as respostas coletadas foi possível identificar que alguns critérios para a realização das marcações podem ter relação com conceitos de recuperação e organização da informação. Os critérios e/ou motivos citados pelos respondentes para grifar trechos dos textos foram destacar:

- Conceitos;
- Exemplos;
- Termos técnicos;
- Pensamentos e posicionamentos do autor do texto;
- Trechos que possuem ligação direta com a pesquisa do leitor;
- Trechos para citação;
- Trechos para fichamento;
- Objetivo do texto;
- Resultados de pesquisa.

Em um primeiro momento, a análise inicial é que grifar o texto promove a rápida recuperação visual das partes importantes para o leitor. De certa forma, cria-se um sistema pessoal de organização de informações, de modo a encontrar de maneira mais eficiente os trechos de interesse para algum fim específico. Também nas respostas foi identificado o uso de cores diferentes para funções diferentes, o que pode ser associado com estratégias de indexação

e recuperação mais especializadas, gerando termos variados para um mesmo documento a partir de trechos grifados com cores diferentes.

Além da recuperação da informação de forma visual pelo leitor, as marcações realizadas com objetivo de serem transcritas em citações ou fichamentos funcionam como um sistema de recuperação para atividades mais práticas da pesquisa. Já na marcação de trechos que possuem ligação com a pesquisa do leitor é possível identificar a prática de associação de textos, o que seria útil para criar conexões para recuperar, por exemplo, documentos com assuntos semelhantes dentro de um sistema de recuperação de informação.

Também percebe-se certa sinergia com o que trazem as autoras Brascher e Café (2008) ao formularem que a OI se preocupa em entender de forma individual a estrutura dos objetos informacionais para organizá-los sistematicamente. Ao destacar as partes principais de um texto, ou quando destaca estruturas importantes como objetivos e resultados, o leitor está destrinchando o objeto informacional e proporcionando um possível caminho para a organização de uma coleção, ou do próprio objeto dentro de uma coleção já estruturada.

De modo geral, a etapa de leitura e marcação de um texto pode ser enxergada como uma fase preliminar das atividades de organização da informação, em específico, daquelas voltadas à descrição conceitual dos documentos (Brascher; Monteiro, 2010). Podemos destacar a ligação mais direta entre alguns:

- **Indexação:** a marcação de conceitos, termos técnicos e pensamentos do autor são de grande valia no processo de indexação, visto que funcionam como uma seleção preliminar de trechos que caracterizam o documento lido e que podem funcionar como um refinamento para destacá-lo de forma individual dentro do corpus em que se encontra;
- **Resumo:** para a elaboração de resumos, as marcações em partes estruturais do texto como os objetivos e resultados é de extrema utilidade para promover agilidade, visto que a identificação de tais partes já estarão destacadas;
- **Classificação:** as marcações de trechos que indicam ligações de um documento com outros de mesma temática podem ser apoio para a classificação do documento dentro de um sistema com documentos diversos.

Algumas marcações, apesar de poderem ser aproveitadas para fins de organização da informação, possuem relação mais direta com funções de aprendizagem e compreensão mais aprofundada do texto, como o destaque de exemplos que ilustram conceitos abordados pelo autor do texto. Por fim, dois respondentes sinalizaram que não marcam seus textos, preferindo

realizar anotações tanto no documento que está sendo lido quanto em um novo documento criado especialmente para reunir as anotações de estudo. Embora fujam do escopo do estudo de caso deste trabalho, as anotações também possuem potencial para servir de insumo para outros estudos da CI, já que também são um registro das interações de usuários com os objetos informacionais.

Por fim, cabe salientar que, devido ao número reduzido de respostas coletadas no formulário de pesquisa, a análise realizada tem caráter preliminar, podendo ser considerada como um indicativo das possibilidades de utilização de grifos e outros registros dos leitores como um critério ou insumo para as práticas da Organização da Informação.

4.2 Realizar a indexação automática do corpus coletado

A aplicação do código de indexação por frequência relativa gerou um conjunto de 73 termos diferentes, com alguns termos se repetindo em mais de um texto analisado. O termo que mais se repetiu em todo o conjunto de termos foi **informação**, representando 5 textos, seguido por **ciência** e **ciência da informação**, representando 4 textos cada um. No Quadro 4 estão listados todos os termos com suas frequências relativas, bem como os respectivos textos aos quais representam, identificados com o ID que se encontra no Quadro 3 da subseção de procedimentos metodológicos.

Quadro 4 - Termos indexadores com suas respectivas frequências e documentos representados

Termo	Frequência TF-IDF	Texto (IDs)
afirmações	0.27	1
aprendizagem	0.30	30
arquivo	0.26	20
artigos	0.28	26
atenção	0.62	30
atribuições	0.29	11
autoria	0.44	18
AV3	0.14	17
biblioteconomia	0.15	5
caracteres	0.29	4
ciência	0.33	29

	0.33	2
	0.13	5
	0.20	14
ciência da informação	0.37	23
	0.35	5
	0.24	13
	0.55	24
científica	0.20	16
citações	0.24	26
coesão	0.20	15
coleções	0.12	12
comunicação	0.11	16
conceitos	0.31	14
conjunctive	0.27	8
cópias	0.29	11
culturales	0.29	4
dados	0.43	7
dados de pesquisa	0.66	12
	0.60	31
digitais	0.09	31
direito	0.17	18
documentário	0.22	27
documentation	0.19	21
	0.19	22
documento	0.25	20
documents	0.26	21
enunciado	0.28	25
espacio	0.33	10
estudo	0.28	28
	0.20	15
evidência	0.27	3
evolução	0.16	23
fórmula	0.27	1

história	0.21	14
history	0.24	19
impacto	0.29	26
informação	0.44	28
	0.21	29
	0.26	2
	0.39	3
	0.21	9
Informação como coisa	0.21	3
interpretação	0.32	27
interpretation	0.19	8
knowledge	0.38	8
letramento	0.18	30
library	0.18	19
linguagem	0.15	17
más condutas	0.27	7
massa	0.23	25
matemática	0.29	1
materialidade	0.37	25
medio	0.22	10
memória	0.26	15
	0.19	20
método	0.19	27
originais	0.29	11
paisaje	0.59	10
paradigm	0.19	6
paradigma	0.16	13
pesquisa	0.19	13
plágio	0.16	18
problema	0.25	9
	0.25	24
projeto	0.15	29
recorded	0.21	6

recurso	0.16	17
referências	0.14	16
repositório	0.14	12
	0.11	31
	0.35	7
retrieval	0.23	21
	0.23	22
science	0.29	6
	0.50	19
	0.30	22
social	0.24	2
sociedade	0.21	23
tecnologia	0.15	24
teoria	0.18	28
territoriales	0.29	4
tomate	0.29	9

Fonte: elaborado pela autora

No Quadro 5 estão listados todos os textos considerados no processo de indexação, seguidos pelos termos definidos pelo código de frequência absoluta e pelo código de frequência relativa. Todos os termos estão em ordem decrescente, ou seja, do mais representativo até o menos representativo, em ambas as frequências.

Quadro 5 - Termos resultantes dos cálculos de frequência aplicados

ID	Termos (frequência absoluta)	Termos (TF-IDF)
1	pouquíssimas afirmações grandemente	afirmações fórmula matemática
2	ciência informação social	ciência informação social
3	informação evidência ciência	evidência informação informação como coisa

4	distintos paisaje caracteres	caracteres culturales territoriales
5	informação ciência da informação ciência	biblioteconomia ciência da informação ciência
6	information science content	paradigm recorded science
7	dados pesquisa repositórios	más condutas dados repositórios
8	knowledge conjunctive interpretation	conjunctive interpretation knowledge
9	informação problema atividade	informação problema tomate
10	paisaje espacio medio	espacio medio paisaje
11	dezenas traços presentes	atribuições cópias originais
12	dados repositórios pesquisa	coleções dados de pesquisa repositório
13	informação pesquisa ciência da informação	ciência da informação paradigma pesquisa
14	conceito ciência informação	ciência conceitos história
15	fatos sociais solidificados	coesão estudos memória
16	científica comunicação medida	científica comunicação referências

17	AV3 linguagem recurso	AV3 linguagem recurso
18	autoria plágio textos	autoria direito plágio
19	information science history	history library science
20	documento escrita informação	arquivo documento memória
21	knowledge physical human	knowledge book document
22	information science retrieval	documentation retrieval science
23	informação ciência da informação sociedade	ciência da informação evolução sociedade
24	informação ciência da informação problemas	ciência da informação problemas tecnologia
25	informação enunciados materialidade	enunciado massa materialidade
26	artigos impactos dados	artigos citações impacto
27	sentido interpretação social	documentário interpretação método
28	informação estudos teoria	estudos informação teoria
29	ciência social informação	ciência informação projeto

30	atenção processo informação	aprendizagem atenção letramento
31	dados repositórios publicações	dados de pesquisa digitais repositório

Fonte: elaborado pela autora

Ao analisar os resultados, num primeiro momento, é possível verificar algumas coincidências entre os termos sugeridos na frequência absoluta e os termos sugeridos na frequência relativa. Na maioria dos textos existe semelhança de pelo menos um termo entre as duas frequências calculadas. Já em outros textos, todos os termos coincidem porém não estão na mesma ordem de representatividade. A diferença se dá pela relação do termo específico com o conjunto de termos do corpus, relação essa calculada pelo TF-IDF. Este tipo de acontecimento pode ser observado nos textos de IDs 8 e 10. Outra coincidência, sendo essa idêntica, foi percebida nos textos de IDs 2 e 17, com resultados iguais tanto em termos sugeridos quanto em posição de representatividade pelas frequências.

Quanto ao processo de indexação aplicado, é possível perceber aspectos tanto da indexação manual quanto da indexação automática. Conforme delimitado por Lancaster (2004), a indexação manual se inicia na leitura do documento realizada por um profissional indexador. No caso aqui estudado, o ponto de início do processo também foi a leitura dos textos, entretanto, diferente do objetivo do profissional que lê o texto visando de fato a representação da informação, aqui os textos foram lidos com objetivos específicos por cada respondente, de acordo com seus campos de estudo e suas temáticas de pesquisa, além das motivações mencionadas na subseção 4.1.

Utilizando os fatores que influem na indexação elencados por Lancaster (2004) como um ponto inicial de análise dos resultados obtidos, observou-se a influência de fatores ligados ao vocabulário (especificidade/sintaxe; qualidade dos dados de entrada), fatores ligados ao documento (conteúdo temático) e, considerando os leitores como uma parcela ativa dos indexadores envolvidos no processo, também percebe-se influência de fatores ligados ao indexador (conhecimento do assunto/capacidade de leitura do texto).

A discussão sobre o que é uma boa indexação é ampla e apresenta diversos caminhos para qualificar o que é um bom termo indexador, visto que a seleção de tais termos depende de contexto dos usuários, funções do documento e da percepção do profissional indexador. No

entanto, ao analisar os termos resultantes do código de indexação utilizado no estudo, percebe-se que a especificidade dos trechos utilizados para a seleção dos termos indexadores teve um impacto no resultado final, pois considera-se que os termos não possuem uma característica importante na indexação: a particularidade do termo em relação aos documentos que compõem o corpus, apesar dos termos serem advindos de marcações consideradas muito específicas pelo contexto de cada leitor, o que apresenta ligação com os fatores relacionados ao indexador.

Nos resultados obtidos, percebe-se a aproximação aos princípios de revocação e o distanciamento dos princípios de precisão, já que a indexação resultou em termos mais abrangentes que podem servir para recuperar textos em um corpus com grande variedade de assuntos, a exemplo do corpus utilizado no estudo. É possível analisar a indexação obtida com um olhar para sua utilidade, que aqui seria recuperar textos a partir destes termos mais abrangentes, um tipo de organização da informação voltada a grandes áreas de conhecimento, o que pode ser utilizado como estratégia de busca quando se pensa em um sistema de recuperação da informação.

Outro fator, ligado ao vocabulário, e que impactou o resultado da indexação, foi identificado no processo de pré-tratamento dos dados, em específico, na tokenização. Ao separar cada palavra em um item individual, seria de grande valia contar com um conjunto de termos específicos da área da CI pronto para ser incorporado no código. A utilização de condições de expressões regulares no código computacional aplicado no estudo, que previne a separação de termos compostos, não possui aporte suficiente para identificar termos compostos mais específicos da área, fazendo com que a semântica e, por consequência, a indexação seja influenciada. Com isso, considera-se que a percepção do profissional que realiza a indexação ainda seja um fator-chave para o processo, de forma a avaliar os termos e identificar o que seria composto e o que seria um termo simples. A organização de um conjunto de termos em formato editável vindos de glossários, termos indexadores de sistemas de recuperação já em funcionamento e outras fontes de informação em CI seria de grande valia para uso em projetos de automação de tarefas.

Quanto ao uso dos termos grifados como principal insumo para a indexação, notou-se que é viável por meio das aplicações computacionais disponíveis. O pré-processamento dos dados pode apresentar alguma complexidade no que diz respeito à extração dos dados, entretanto, a utilização dos trechos grifados permite o processamento mais direcionado do cálculo de frequência, que agora se ocupa de partes específicas do texto e não mais do documento por completo. A indexação resultante do processo tem características que devem

ser associadas em estratégias de busca mais especializadas no sentido do tema da pesquisa, mas oferece possibilidades diferenciadas de tratamento e organização de um mesmo texto.

Também cabe destacar a possibilidade de indexação temática por meio de códigos de cores aplicados no texto do momento da marcação, criando novos resultados de busca para um mesmo texto a depender do intuito do usuário, como por exemplo, marcações em amarelo para trechos que serão citados em sua pesquisa geram um conjunto de termos indexadores para um documento e marcações em vermelho para exemplos de explicação geram um outro conjunto de termos indexadores para o mesmo documento, que pode ser recuperado por meio de estratégias de busca diferentes. Além disso, salienta-se também a possibilidade de indexar os próprios trechos grifados no âmbito de um sistema de recuperação para uso pessoal de pesquisadores, o que pode promover mais assertividade nas atividades executadas ao longo da pesquisa.

4.3. Propor um *workflow* de indexação de termos extraídos de anotações em documentos digitais

O fluxo de trabalho construído no presente estudo teve o intuito de dar suporte a possíveis aplicações de aprendizagem de máquina no contexto da indexação e da recuperação da informação, utilizando linguagem e bibliotecas de código compatíveis com as tecnologias empregadas em aplicações deste tipo. Servindo como uma proposta preliminar, cabe ressaltar a importância da continuidade de construção do fluxo de forma a abranger demais etapas como por exemplo, a de avaliação de desempenho dos termos indexadores em sistemas de recuperação.

A representação gráfica do fluxo mostra a proposta de indexação desde a sua etapa conceitual, que é o momento de marcação e criação dos destaques no texto por parte dos leitores, passando pelas etapas práticas aplicadas no estudo acompanhadas cada uma de seus insumos informacionais e deixando como sugestão as próximas etapas que se considera necessárias para o desenvolvimento da aplicação de aprendizagem de máquina. Cada subprocesso possui um código para execução, estando todos disponibilizados nos apêndices deste trabalho, assim como o código da etapa de cálculo de frequência para definição de termos indexadores. As etapas ilustradas são as seguintes:

- **Pré-processo:** é a etapa onde os documentos ainda estão em uso pelos leitores, sendo utilizados para fins de estudo e pesquisa. Nesta fase os textos recebem os destaques, anotações e outras marcações resultantes das interações dos leitores e são armazenadas

em um diretório. Essa atividade é continuada, o diretório deve armazenar os textos sempre que o leitor executar suas leituras e marcações, salvando os arquivos para as próximas etapas do fluxo;

- **Extração de trechos grifados:** esta é a etapa em que, por meio de códigos escritos em Python, se realiza os subprocessos de checagem de cor e extração de trechos. Primeiro, aplica-se o código para detectar padrões de cores utilizados para grifar partes dos textos analisados. Em seguida, utilizando os padrões identificados, é realizada a extração dos trechos grifados. Com isto, ao final desta etapa, cria-se um novo diretório, agora contendo os arquivos com os trechos destacados pelos leitores em cada texto. Aqui o código conta com o suporte da biblioteca *PyMuPDF* que deve ser instalada antes da execução dos códigos disponibilizados nos apêndices A e B;
- **Pré-processamento de dados:** esta etapa utiliza os insumos informacionais obtidos na etapa de extração para a continuidade no processo de indexação. Nesta fase são realizados três subprocessos: limpeza e eliminação de *stopwords, tokenização* e lematização, todos utilizando códigos computacionais com o suporte da biblioteca NLTK, que deve ser instalada antes da execução dos códigos. O primeiro subprocesso é o de limpeza de sinais gráficos e numéricos junto à remoção das *stopwords*, palavras que quando retiradas da construção de sentenças não possuem sentido semântico associado ao contexto dos documentos. Depois, segue-se para a *tokenização* dos trechos já limpos, aplicando o código para que todas as palavras dos trechos sejam transformadas em um único item, apto a ser utilizado como termo indexador do documento de origem. Com o arquivo contendo os *tokens* de cada trecho extraído dos textos, segue-se para a lematização, processo para reduzir palavras à sua forma mais simples no contexto morfológico, de modo a eliminar plurais e variações de conjugação de um mesmo verbo, por exemplo. O resultado desta etapa é um diretório com os arquivos com tokens de cada texto para ser utilizado na etapa subsequente;
- **Calcular frequência para gerar termos indexadores:** etapa na qual os termos gerados na etapa anterior são processados pelo código aplicando os cálculos de frequência para obtenção dos termos candidatos à indexação. O resultado é um arquivo de texto com os termos de cada texto acompanhados de sua frequência calculada.

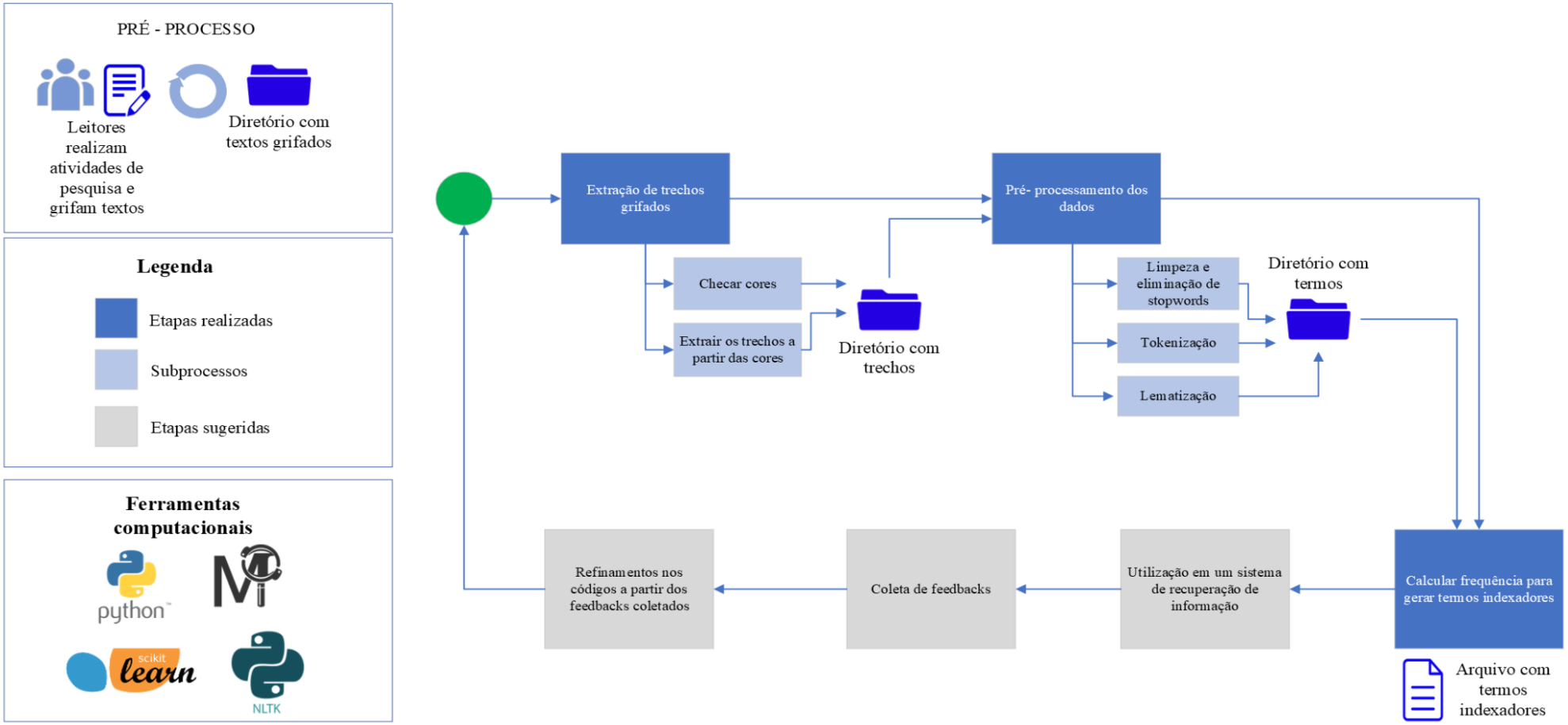
A aplicação prática do estudo foi desenvolvida até a etapa de cálculo de frequência. Após isto, as etapas ilustradas no fluxograma (Figura 19) são sugeridas para a possível aplicação em um modelo de aprendizagem de máquina. São elas:

- **Utilização em um sistema de recuperação de informação:** os termos obtidos no processo de cálculo de frequência podem ser utilizados em um sistema de recuperação de informação para que sejam testados quanto a sua eficiência na busca por documentos. Uma possibilidade para essa utilização pode ser a construção de um protótipo de sistema com os textos utilizados nas etapas anteriores com o auxílio de ferramentas abertas e de código livre como, por exemplo, o *VuFind*[15],. Este tipo de implementação possibilitaria a utilização prática dos termos indexadores bem como a coleta de métricas do desempenho dos termos, que configura a próxima etapa sugerida no fluxograma;
- **Coleta de feedbacks:** ao utilizar os termos em um sistema de recuperação de informação é importante avaliar e registrar o desempenho tanto do sistema quanto da eficiência dos termos empregados. Pode ser realizada de maneiras diversas, seja com formulários para coletar a opinião dos usuários e também por meio dos registros técnicos do próprio sistema, como o histórico de busca de cada usuário e os resultados retornados. A partir da análise dos resultados e feedbacks é possível promover melhorias e avanços tanto na indexação quanto no sistema utilizado;
- **Refinamentos nos códigos a partir dos feedbacks coletados:** os refinamentos são parte essencial para garantir o bom funcionamento do fluxo como um todo e, pensando em uma aplicação de aprendizagem máquina, são fundamentais para o próprio processo de aprendizagem. Finalizando o ciclo, esta também é a etapa que garante que se mantenha o ciclo proposto em bom funcionamento.

Quanto à implementação das etapas sugeridas para que se incorpore tecnologias de aprendizagem de máquina, é importante destacar os benefícios alcançados e possibilidades adicionadas nesse cenário. O primeiro e mais notável é um aspecto inerente de todos os sistemas de aprendizagem de máquina, que é a melhoria contínua da execução das tarefas desenvolvidas pelo sistema, ou seja, a adaptação dos processos à medida em que novos documentos e à medida em que as necessidades e exigências evoluem, se caracterizando como uma abordagem valiosa para garantir que o sistema permaneça relevante e eficiente ao lidar com mudanças nas condições, na natureza dos documentos ou nas preferências dos usuários, sendo aqui identificadas nos trechos grifados e também na coleta de *feedbacks* após o uso dos termos indexadores gerados.

[15] Disponível em: https://vufind.org/vufind/

Figura 19 - Fluxo de trabalho para indexação automática a partir de grifos em documentos digitais



Fonte: elaborado pela autora

Outros benefícios adicionais a se destacarem são:

- **Adaptação do sistema a mudanças:** um sistema de aprendizagem de máquina pode ter a capacidade de se ajustar a mudanças nas condições ou no contexto em que está operando, o que pode incluir ajustes automáticos nas regras de indexação, na ponderação de características ou em outras partes do processo de aprendizagem;
- **Feedback do usuário:** etapa prevista no fluxograma da Figura 19, significando que se os usuários indicarem imprecisões ou fornecerem informações adicionais sobre a relevância dos resultados de indexação, o sistema pode ajustar seus modelos com base nesse feedback para melhorar a precisão futura;
- **Atualização de modelos:** os modelos de aprendizagem de máquina podem ser periodicamente treinados com novos dados, permitindo que eles se adaptem a mudanças nas distribuições de dados e mantenham sua eficácia ao longo do tempo;
- **Incorporação de novas características:** se novas características ou tipos de documentos ou novas marcações nos documentos já existentes surgirem, o sistema pode ser projetado para incorporar essas informações de forma aprimorada. Isso é crucial para garantir que o sistema seja capaz de lidar com uma variedade crescente de conteúdo dentro de um sistema de recuperação de informação;
- **Monitoramento de desempenho:** o sistema pode contar com mecanismos de monitoramento contínuo de desempenho, permitindo a detecção rápida de quaisquer degradações na qualidade da indexação. Isso pode desencadear ações corretivas ou atualizações automáticas bem como oferecer métricas para futuros estudos na área de indexação automática ou de aprendizagem de máquina;
- **Escalabilidade:** a estrutura do sistema pode ser projetada para escalar eficientemente à medida que o volume de dados e a complexidade do ambiente aumentam, fazendo com que o sistema possa lidar com grandes quantidades de documentos e se adapte a novos cenários e contextos informacionais.

Quanto às opções de algoritmos a serem empregados num sistema como o proposto neste trabalho, existem diversos aspectos existentes em variados algoritmos capazes de contribuírem na automação das tarefas em cada uma das etapas do fluxograma, tendendo assim a uma abordagem híbrida, como a apresentada por Khan *et al.* (2010). Por exemplo, as Redes Neurais Convolucionais (CNNs) são ideais para extrair características espaciais em trechos grifados, permitindo a classificação eficaz em categorias específicas para a indexação automática de documentos.

Já os do tipo Máquinas de Vetores de Suporte (SVMs), como modelos clássicos de aprendizado supervisionado, destacam-se na classificação com base em características relevantes para a indexação. Modelos de Linguagem Pré-treinados, como o *Bidirectional Encoder Representations from Transformers* (BERT) e o *Generative Pre-trained Transformer* (GPT), capturam representações semânticas avançadas, sendo ajustados para classificar trechos grifados em categorias específicas de indexação.

Algoritmos de Agrupamento, como o k-means, podem agrupar trechos grifados sem rótulos em clusters semelhantes, enquanto métodos de Análise de Tópicos, como *Latent Dirichlet Allocation* (LDA) e *Non-Negative Matrix Factorization* (NMF), identificam temas predominantes nos trechos grifados, fornecendo uma base sólida para a categorização e indexação.

Abordagens semi-supervisionadas, como métodos de propagação de rótulos, permitem começar com um conjunto pequeno de trechos grifados rotulados manualmente e propagar esses rótulos para trechos sem rótulos. Algoritmos de Aprendizado por Reforço, como *Q-learning* ou *Deep Q-networks* (DQN), são úteis quando é possível definir um sistema de recompensas para otimizar decisões de indexação ao longo do tempo. O *Fine-tuning* de Modelos Pré-treinados, como BERT ou GPT, adapta esses modelos à tarefa específica de indexação de documentos a partir de trechos grifados. Redes Neurais Recorrentes (RNNs) e *Long Short-Term Memory* (LSTM) são aplicadas quando há dependência temporal ou estrutura sequencial nos trechos grifados.

Em cenários multimodais, a integração de texto e imagem por meio de modelos como *Contrastive Language-Image Pre-training* (CLIP) ou *Multimodal Transformer* (MMT) é útil para processar dados heterogêneos na indexação automática. A escolha do modelo depende das características dos documentos, da disponibilidade de dados e da natureza da tarefa de indexação. O pré-processamento adequado dos dados é essencial para garantir a eficácia desses modelos na automatização do processo de indexação de documentos.

5 CONCLUSÕES

Lidar com a representação, organização e recuperação da informação em contexto digital representa um desafio ao mesmo tempo que possibilita a exploração de diferentes modos de atender as necessidades informacionais intrínsecas a todos os indivíduos em suas atividades cotidianas. A convergência entre a interação humana na identificação de trechos relevantes e a automação do processo de indexação explorada neste estudo se mostra promissora para otimizar a organização e recuperação de informações em ambientes digitais.

Quanto aos objetivos específicos delimitados, o primeiro deles que buscou identificar de maneira conceitual a aplicação da organização da informação foi cumprido ao explorar as relações entre os processos mentais e motivações para grifar textos por parte dos respondentes da pesquisa com os processos de organização da informação já conhecidos. Foi possível perceber aproximações no sentido em que ambos visam criar maneiras de recuperar a informação de forma mais eficiente ao categorizar, destacar e organizar de forma estruturada um documento.

Sobre a aplicação de ferramentas computacionais para a manipulação dos trechos grifados e para a realização da indexação, considera-se satisfatórios os resultados obtidos com a utilização das bibliotecas de código bem como os próprios códigos. A adoção de bibliotecas de código aberto e com suporte para futuras melhorias e aplicação em sistemas de aprendizagem de máquina abrem a possibilidade de futuros trabalhos a partir dos resultados aqui apresentados. Além disso, a utilização de códigos com sintaxe clara e acessível faz com que os códigos disponibilizados possam ser adaptados de forma simplificada em outros contextos informacionais.

O fluxo construído no trabalho serve como ponto de partida para a criação de um sistema de automação de indexação de documentos digitais, podendo ser considerado um protótipo de etapas iniciais do desenvolvimento de um sistema de recuperação de informações. Enfatiza-se a necessidade de ajustes nos códigos utilizados, aprofundando e aprimorando a eficiência dos parâmetros de indexação, tanto para calcular os melhores termos indexadores bem como a construção de mecanismos de recuperação para os termos identificados.

Como possíveis desdobramentos do estudo, destaca-se a continuidade do desenvolvimento prático das etapas sugeridas no fluxograma para que a proposta preliminar de sistema avance para o nível de aprendizagem de máquina. Também cabe enfatizar a possibilidade da aplicação do fluxograma em outros contextos informacionais, como a indexação para os próprios trechos grifados para uso pessoal de pesquisadores ou grupos de

pesquisa ou a indexação realizada a partir de códigos de cores estabelecidos no momento de grifar os documentos, estabelecendo funções específicas para cada cor ou tipo de marcação. Além disso, a utilização do fluxo em outros contextos possibilita a manutenção do código e o surgimento de novas adaptações para o uso de trechos ou outras marcações de leitores.

REFERÊNCIAS

AGUIAR, F.; KOBASHI, N. Organização e representação do conhecimento: perspectivas de interlocução interdisciplinar entre Ciência da Informação e Arquivologia. In ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16º, Santa Catarina, 2013

AYODELE, Taiwo Oladipupo. Types of Machine Learning Algorithms. [S.l.]:IntechOpen, 2010. DOI10.5772/9385. Disponível em: https://www.intechopen.com/chapters/10694.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Recuperação de informação: conceitos e tecnologias das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

BERNARDO, J. C. O.; KARWOSKI, A. M. A leitura em dispositivos digitais móveis. ETD - Educação Temática Digital, [S. l.], v. 19, n. 4, p. 795–807, 2017. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/etd/article/view/8646355.

BOURDIEU, Pierre; CHARTIER, Roger. A leitura: uma prática cultural. Debate entre Pierre Bourdieu e Roger Chartier. In: CHARTIER, Roger(Org.). Práticas da leitura. Tradução de Cristiane Nascimento. São Paulo: Estação Liberdade, 1996.

BRASCHER, Marisa; CAFÉ, Ligia. Organização da informação ou organização do conhecimento? In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. Anais... São Paulo: ENANCIB, 2008. p. 1-14

BRÄSCHER, M.; MONTEIRO, F. S. Organização da informação em repositórios digitais. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 15, n. 29, 2010.

BRAVO, B. R. El documento: entre la tradición y la renovación. Ediciones Trea, 2002

BRITO SILVA, S.R. Sistemas de indexação automática por atribuição: uma análise comparativa. Dissertação (Mestrado em Ciência da Informação – Departamento de Ciência da Informação, Universidade Federal de Pernambuco. Recife. 2020.

BRYMAN, A. Social Research Methods. 4. ed. Oxford: Oxford University Press, 2012.

BUCKLAND, M. K. What is a document? Journal of the American Society for Information Science, Washington, v.48, n.9, p. 804-809, Sept., 1997.

CAFÉ, Lígia; SALES, R. Organização da informação: Conceitos básicos e breve fundamentação teórica. In: Jaime Robredo; Marisa Bräscher (Orgs.). Passeios no Bosque da Informação: Estudos sobre Representação e Organização da Informação e do Conhecimento. Brasília DF: IBICT, 2010. 335 p. ISBN: 978-85-7013-072-3. Capítulo 6, p. 115-129. Edição eletrônica.

COOK, L.K.; MAYER, R.E. Reading Strategies Training for Meaningful Learning from Prose. In: Pressley, M. and Levin, J.R. Editoras Cognitive Strategy Research, Springer, New York, p. 87-131, 1983. http://dx.doi.org/10.1007/978-1-4612-5519-2_4

CRESWELL, J. W.; CRESWELL, J. D. Projeto de pesquisa: métodos qualitativo, quantitativo e misto. 5. ed. Porto Alegre: Penso, 2021. 241 p.

CUKIER, Kenneth. *Data, data, everywhere: a special report on managing information. The Economist*, v. 394, Issue 867, Feb. 2010.

DAMACENO, S. S.; VASCONCELOS, R. O. Inteligência artificial: uma breve abordagem sobre seu conceito real e o conhecimento popular. Caderno de Graduação - Ciências Exatas e Tecnológicas - UNIT - SERGIPE, [S. l.], v. 5, n. 1, p. 11, 2018. Disponível em: https://periodicos.set.edu.br/cadernoexatas/article/view/5729

DANTAS, T., MANGAS-VEGA, A., GOMÉZ-DÍAZ, R.,; CORDÓN-GARCÍA, J. A. Pesquisa em leitura e pesquisa em leitura digital: Panorama do atual cenário científico. Informação & Sociedade:Estudos., João Pessoa, v.27, n.2, p. 117-131, maio/ago. 2017

DE MAURO, Andrea; GRECO, Marco; GRIMALDI, Michele. A formal definition of Big Data based on its essential features. Library Review, v. 65, n. 3, p. 122–135, 1 jan. 2016. https://doi.org/10.1108/LR-06-2015-0061.

DI VESTA, F. J.; GRAY, G. S. Listening and note taking. Journal of Educational Psychology, Vol. 63, n. 1, p. 8–14, 1972.

DIJKSTRA, E. W. The Humble Programmer. Communications of the ACM, v. 15, n. 10, p. 859–866. 1971.

FAVERIO, Michelle; PERRIN, Andrew. *Three-in-ten Americans now read e-books.* Pew Research Center, 6 jan. 2022. Disponível em: https://www.pewresearch.org/fact-tank/2022/01/06/three-in-ten-americans-now-read-e-books/

FIDEL, R. *User-oriented indexing.* Journal of the American Society for Information Science, v. 45, p. 572-576. 1994.

GIL-LEIVA, I. *Manual de indización: teoría y práctica*. Gijón: Trea, 2009.

GLUSHKO, Robert J. *The Discipline of Organizing: Professional Edition*. 4° edição. O'Reilly Media, 2016.

GONZALÉZ, José Antonio Moreiro. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural.*

GOODFELLOW, I., BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, Cambridge. 2016. Disponível: http://www.deeplearningbook.org

GOUVEIA, L. M. B; GAIO, S. Sociedade da informação: balanço e oportunidades. Rio de Janeiro: Universidade Fernando Pessoa, 2004.

GUINCHAT, C.; MENOU, M. Introdução geral às ciências e técnicas da informação e documentação. 2.ed. rev. aum. Brasília: Ibict;CNPq, 1994. 540 p.

HAPKE, H.; NELSON, C. Building Machine Learning Pipelines: Automating Model Life Cycles with TensorFlow. O'Reilly Media, Incorporated, 2020.

HJØRLAND, B. *Automatic Indexing. In: Lifeboat for Knowledge Organization*, 2008.

KELLY, Kevin. O que os Livros se tornarão. In: SILVEIRA, Julio (Org.). Livrolivre, Novas Possibilidades do Digital para a Escrita, a Leitura e a Publicação. Rio de Janeiro: Imã Editorial, 2011. P.17-37.

KHAN, A.; BAHARUDIN, B.; LEE, L.H.; KHAN, K. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of advances in information technology, v. 1, n. 1. Fevereiro, 2010.

KRAUT, R. Policy guidelines for mobile learning. Paris: United Nations Educational, Scientific And Cultural Organization, 2013.

LANCASTER, F.W. Indexação e Resumos: teoria e prática. 2ªed. Brasília, DF: Briquet de Lemos, 2004

LAPA, R. C.; CORRÊA, R. F. Indexação automática no âmbito da ciência da informação no brasil. Informação & Tecnologia, v. 1, n. 2, p. 59-76, 2014. Disponível em: http://hdl.handle.net/20.500.11959/brapci/41624.

LE COADIC, Y.F. A ciência da informação. 2. ed. Brasília, DF: Briquet de Lemos, 2004

LEUTNER, D., LEOPOLD, C., & DEN ELZEN-RUMP, V. Self-regulated learning with a text-highlighting strategy: A training experiment. Zeitschrift für Psychologie/Journal of Psychology. Vol. 215, n. 3, p. 174–182, 2007.

LÓPEZ YEPES, J. Hombre y documento: del homo sapiens al homo documentador. Scire, Zaragoza, v.4, n.2., jul-dec., 1998.

MAI, J. *Semiotics and indexing: an analysis of the subject indexing process.* Journal of Documentation, Vol. 57, n. 5, p. 591-622. 2001. Disponível em: https://doi.org/10.1108/EUM0000000007095

MEDELYAN, O. *Human-competitive automatic topic indexing. PhD Thesis. University of Waikato, New Zealand,* 2009. Disponível em: https://hdl.handle.net/10289/3513

MICHEL, J. L'Information et documentation un domaine d'activité professionnelle en mutation : LCN Les Métiers du Numérique. Hermès, v. 1, n.3, p. 47-64, 2000.

MORAIS, E.A.M. AMBRÓSIO, A.P. Mineração de Textos. Goiás, Brasil, p. 1-30. 2007. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf

MURPHY, K. P.; MASSACHUSETTS INSTITUTE OF TECHNOLOGY. Machine learning : a probabilistic perspective. Cambridge (Ma): Mit Press, 2012.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação Automatizada de Artigos de Periódicos Científicos: análise da aplicação do *software* SISA com uso da terminologia DeCS na área de Odontologia. Informação e Sociedade: Estudos, João Pessoa, v.19, n.2, p. 99-118, 2009.

OTLET, P. Traité de documentation: le livre sur le livre: théorie et pratique. Bruxelles: Mundaneum, 1934.

PÉDAUQUE. R. Document: forme, signe et medium, le reformulations de numériques. STIC-CNRS, 8 jun., 2003.

PIMENTA, R. M. Por que humanidades digitais na ciência da informação? Perspectivas pregressas e futuras de uma prática transdisciplinar comum. Informação & Sociedade: Estudos, João Pessoa, v. 30, n. 2, 2020. DOI: 10.22478/ufpb.1809-4783.2020v30n2.52122. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/52122.

PINTO, V. B. Indexação documentária: uma forma de representação do conhecimento registrado. Perspectivas em Ciência da Informação, Belo Horizonte, v.6, n.2, p.223-234, jul./dez., 2001.

ROBREDO, J. Documentação de hoje e de amanhã. 4. ed. rev. ampl. Brasília, DF: Ed. Do Autor, 2005.

RUSSEL, R. Machine Learning: Step-by-Step Guide To Implement Machine Learning Algorithms with Python. CreateSpace Independent Publishing Platform, 2018.

SALTON, G; MCGILL, M. J. Introduction to Modern Information Retrieval. John

Wiley and Sons, New York, 1983.

SAMPIERI, R. H.; COLLADO, C. F..; LUCIO, M. P. B. Metodologia de pesquisa. 5 ed. Dados eletrônicos - Porto Alegre: Penso, 2013.

SAMUEL, A. *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal, Vol. 3, p. 210–229. Mar. 1959. Disponível em: https://ieeexplore.ieee.org/document/5392560

SANTOS, H. M. D.; FLORES, D. O documento digital no contexto das funções arquivísticas. Páginas A&B, Arquivos e Bibliotecas (Portugal), n. 5, p. 165-177, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/65458.

SARACEVIC, T. A natureza interdisciplinar da ciência da informação. Ciência da Informação, v. 24, n. 1, 1995. DOI: 10.18225/ci.inf..v24i1.608

SETCHI, R.; TANG, Q. Concept indexing using ontology and supervised machine learning. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, v. 1, n. 1, p. 89–94, 26 jan. 2007.

SHERA, J. H.; EGAN, M. E. Examen del estado actual de la biblioteconomía y de la documentación. Santa Fe, Argentina: Centro de Documentación e Información de Asuntos Municipales Doctor Alcides Greca, 1953.

SILVA, I. C. O.; GOUVEIA, F. C. A busca e o acesso às informações sobre saúde no contexto tecnológico. Revista Conhecimento em Ação, v. 4, n. 2, p. 23-45, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/127448.

SILVA, R. M. *et al*. Towards filtering undesired short text messages using an online learning approach with semantic indexing. Expert Systems With Applications, v. 83, p. 314–325, 15 out. 2017.

SIQUEIRA, J. C. A noção de documento digital: uma abordagem terminológica. Em Questão, Porto Alegre, v. 18, n. 1, p. 125–140, 2012. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/24172

WEINSTEIN, C; MAYER, R. The Teaching of Learning Strategies. In: Wittrock, M., Editora Handbook of Research on Teaching, Macmillan, New York. p. 315-327, 1986.

ZAYAS, Emilio Lópes-Barajas et al. O Paradigma da Educação Continuada. Porto Alegre: Penso, 2012.

APÊNDICE A - Código de checagem de cor

```
import fitz  # PyMuPDF

# Open the PDF
doc = fitz.open('//caminho/do/arquivo')

# Set to store unique colors
unique_colors = set()

# Loop through every page
for i in range(len(doc)):
   page = doc[i]
   # Get the annotations (highlights are a type of annotation)
   annotations = page.annots()
   for annotation in annotations:
       if annotation.type[1] == 'Highlight':
           # Get the color of the highlight
           color = annotation.colors['stroke']  # Returns a RGB tuple
           unique_colors.add(color)

# Print all unique colors
for color in unique_colors:
   print(color)
```

APÊNDICE B - Código de extração de termos grifados

```
import fitz  # PyMuPDF
import pandas as pd

# Open the PDF
doc = fitz.open('/caminho/do/arquivo')

# Define the RGB values for your colors
VERMELHO = (0.00392200006172061, 1.0, 1.0)
AMARELO = (1.0, 1.0, 0.0)

color_definitions = {"Vermelho": VERMELHO, "Amarelo": AMARELO}

# Create separate lists for each color
data_by_color = {"Vermelho": [], "Amarelo": []}
```

```
# Loop through every page
for i in range(len(doc)):
   page = doc[i]
   annotations = page.annots()
   for annotation in annotations:
       if annotation.type[1] == 'Highlight':
           color = annotation.colors['stroke']  # Returns a RGB tuple
           if color in color_definitions.values():
               # Get the detailed structure of the page
               structure = page.get_text("dict")

               # Extract highlighted text line by line
               content = []
               for block in structure["blocks"]:
                   for line in block["lines"]:
                       for span in line["spans"]:
                           r = fitz.Rect(span["bbox"])
                           if r.intersects(annotation.rect):
                               content.append(span["text"])

               content = " ".join(content)

               # Append the content to the appropriate color list
               for color_name, color_rgb in color_definitions.items():
                   if color == color_rgb:
                       data_by_color[color_name].append(content)

# Convert each list to a DataFrame and write to a separate .csv file
for color_name, data in data_by_color.items():
   if data:
       df = pd.DataFrame(data, columns=["Text"])
       df.to_csv(f'highlighted_text_{color_name.lower()}.csv',
index=False)
```

APÊNDICE C - Código de tokens

```
from nltk.tokenize import word_tokenize

def tokenizar_arquivo(input_path, output_path):
   with open(input_path, 'r', encoding='utf-8') as arquivo_entrada:
       texto = arquivo_entrada.read()
```

```
        # Tokenize o texto em palavras
        palavras = word_tokenize(texto)

        # Salve os tokens em um novo arquivo
        with open(output_path, 'w', encoding='utf-8') as arquivo_saida:
            arquivo_saida.write("\n".join(palavras))

# Substitua 'caminho/do/seu/arquivo/input.txt' pelo caminho real do seu arquivo de entrada
caminho_arquivo_entrada = '/caminho/arquivo/entrada'

#  Substitua  'caminho/do/seu/arquivo/output_tokens.txt'  pelo  caminho desejado para o arquivo de saída
caminho_arquivo_saida = 'caminho/arquivo/saída

tokenizar_arquivo(caminho_arquivo_entrada, caminho_arquivo_saida)
```

APÊNDICE D - Código de remoção de stopwords

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# Lista de stop words fornecida por você
minhas_stop_words = [ sua lista de palavras]


# Use set para garantir uma busca eficiente
stop_words = set(minhas_stop_words)

caminho_arquivo_csv = "caminho/arquivo"

with open(caminho_arquivo_csv, 'r') as file1:
    # Use this to read file content as a stream:
    line = file1.read()

    # Tokenize o texto em palavras
    words = word_tokenize(line)

    # Filtre as stop words
    filtered_words = [word for word in words if word.lower() not in stop_words]
```

```
    # Salve o resultado em um novo arquivo
    with open('texto_limpo.txt', 'a') as appendFile:
        appendFile.write(" ".join(filtered_words))
```

APÊNDICE E - Código de lematização

```
import spacy

def lemmatize_tokens_spacy(tokens):
    nlp = spacy.load('pt_core_news_sm')
    lemmatized_tokens = [token.lemma_ for token in nlp(' '.join(tokens))]
    return lemmatized_tokens

def lemmatize_file(input_filepath, output_filepath):
    with open(input_filepath, 'r', encoding='utf-8') as input_file:
        content = input_file.read()
        tokens = content.split()  # Assumindo que os tokens estão separados
por espaços

    lemmatized_tokens = lemmatize_tokens_spacy(tokens)

    with open(output_filepath, 'w', encoding='utf-8') as output_file:
        for lemma in lemmatized_tokens:
            output_file.write(lemma + '\n')

# Exemplo de uso:
input_file_path = '/caminho/arquivo/entrada'
output_file_path = '/caminho/arquivo/saida'

lemmatize_file(input_file_path, output_file_path)
```

APÊNDICE F - Código de frequência absoluta

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import PlaintextCorpusReader

# Caminho para o diretório que contém o arquivo tokens23.txt
caminho = '/caminho/do/arquivo'

# Lista de arquivos no diretório
arquivos = ['texto_sem_pont.txt']
```

```
# Criar um objeto PlaintextCorpusReader
corpus = PlaintextCorpusReader(caminho, '.*\.txt')

# Ler o conteúdo do arquivo
texto = corpus.raw('texto_sem_pont.txt ')

# Tokenizar o texto
tokens = word_tokenize(texto)

# Calcular a frequência de cada termo
frequencia = nltk.FreqDist(tokens)

# Ordenar os termos por frequência decrescente
termos_ordenados = sorted(frequencia.items(), key=lambda item: item[1],
reverse=True)

# Salvar a frequência em um arquivo de texto
caminho_saida = 'caminho/de/saida'  # Substitua pelo caminho desejado
with open(caminho_saida, 'w') as arquivo_saida:
    for token, freq in termos_ordenados:
        arquivo_saida.write(f"{token}: {freq}\n")

print(f"Frequência salva em: {caminho_saida}")
```

APÊNDICE G - Código de frequência TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer
import os

# Diretório que contém os arquivos de texto
diretorio_documentos = "caminho/arquivo/entrada"

# Lista para armazenar os textos dos documentos
corpus = []

# Ler cada arquivo no diretório
for filename in os.listdir(diretorio_documentos):
    path = os.path.join(diretorio_documentos, filename)
    if os.path.isfile(path) and filename.endswith(".txt"):
        with open(path, 'r', encoding='utf-8') as file:
            texto_documento = file.read()
```

```
            corpus.append(texto_documento)

# Criar o vetorizador TF-IDF
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(corpus)

# Obter termos e pontuações TF-IDF
terms = vectorizer.get_feature_names_out()
tfidf_scores = tfidf_matrix.toarray()

# Nome do arquivo de saída
arquivo_saida = "caminho/arquivo/saida"

# Escrever no arquivo de saída
with open(arquivo_saida, 'w', encoding='utf-8') as output_file:
    # Escrever as pontuações TF-IDF para cada documento
    for i, doc in enumerate(corpus):
        output_file.write(f"\nDocumento {i + 1}: {os.path.basename(os.path.normpath(doc))}\n")

        # Criar um dicionário para armazenar os termos e suas pontuações no documento atual
        termos_e_pontuacoes = {term: tfidf_scores[i, j] for j, term in enumerate(terms)}

        # Ordenar os termos pelo valor da pontuação em ordem decrescente
        termos_ordenados = sorted(termos_e_pontuacoes.items(), key=lambda x: x[1], reverse=True)

        # Escrever apenas os cinco termos mais representativos no arquivo
        for termo, pontuacao in termos_ordenados[:3]:
            output_file.write(f"{termo}: {pontuacao}\n")

print(f"Saída gravada em {arquivo_saida}")
```

ANEXO A - Lista de stopwords em português

'a', 'à', 'adeus', 'agora', 'aí', 'ainda', 'além', 'algo', 'alguém', 'algum', 'alguma', 'algumas', 'alguns', 'ali', 'ampla', 'amplas', 'amplo', 'amplos', 'ano', 'anos', 'ante', 'antes', 'ao', 'aos', 'apenas', 'apoio', 'após', 'aquela', 'aquelas', 'aquele', 'aqueles', 'aqui', 'aquilo', 'área', 'as', 'às', 'assim', 'até', 'atrás', 'através', 'baixo', 'bastante', 'bem', 'boa', 'boas', 'bom', 'bons', 'breve', 'cá', 'cada', 'catorze', 'cedo', 'cento', 'certamente', 'certeza', 'cima', 'cinco', 'coisa', 'coisas', 'com', 'como', 'conselho', 'contra', 'contudo', 'custa', 'da', 'dá', 'dão', 'daquela', 'daquelas', 'daquele', 'daqueles', 'dar', 'das', 'de', 'debaixo', 'dela', 'delas', 'dele', 'deles', 'demais', 'dentro', 'depois', 'desde', 'dessa', 'dessas', 'desse', 'desses', 'desta', 'destas', 'deste', 'destes', 'deve', 'devem', 'devendo', 'dever', 'deverá', 'deverão', 'deveria', 'deveriam', 'devia', 'deviam', 'dez', 'dezanove', 'dezasseis', 'dezassete', 'dezoito', 'dia', 'diante', 'disse', 'disso', 'disto', 'dito', 'diz', 'dizem', 'dizer', 'do', 'dois', 'dos', 'doze', 'duas', 'dúvida', 'e', 'é', 'ela', 'elas', 'ele', 'eles', 'em', 'embora', 'enquanto', 'entre', 'era', 'eram', 'éramos', 'és', 'essa', 'essas', 'esse', 'esses', 'esta', 'está', 'estamos', 'estão', 'estar', 'estas', 'estás', 'estava', 'estavam', 'estávamos', 'este', 'esteja', 'estejam', 'estejamos', 'estes', 'esteve', 'estive', 'estivemos', 'estiver', 'estivera', 'estiveram', 'estivéramos', 'estiverem', 'estivermos', 'estivesse', 'estivessem', 'estivéssemos', 'estiveste', 'estivestes', 'estou', 'etc', 'eu', 'exemplo', 'faço', 'falta', 'favor', 'faz', 'fazeis', 'fazem', 'fazemos', 'fazendo', 'fazer', 'fazes', 'feita', 'feitas', 'feito', 'feitos', 'fez', 'fim', 'final', 'foi', 'fomos', 'for', 'fora', 'foram', 'fôramos', 'forem', 'forma', 'formos', 'fosse', 'fossem', 'fôssemos', 'foste', 'fostes', 'fui', 'geral', 'grande', 'grandes', 'grupo', 'há', 'haja', 'hajam', 'hajamos', 'hão', 'havemos', 'havia', 'hei', 'hoje', 'hora', 'horas', 'houve', 'houvemos', 'houver', 'houvera', 'houverá', 'houveram', 'houvéramos', 'houverão', 'houverei', 'houverem', 'houveremos', 'houveria', 'houveriam', 'houveríamos', 'houvermos', 'houvesse', 'houvessem', 'houvéssemos', 'isso', 'isto', 'já', 'la', 'lá', 'lado', 'lhe', 'lhes', 'lo', 'local', 'logo', 'longe', 'lugar', 'maior', 'maioria', 'mais', 'mal', 'mas', 'máximo', 'me', 'meio', 'menor', 'menos', 'mês', 'meses', 'mesma', 'mesmas', 'mesmo', 'mesmos', 'meu', 'meus', 'mil', 'minha', 'minhas', 'momento', 'muita', 'muitas', 'muito', 'muitos', 'na', 'nada', 'não', 'naquela', 'naquelas', 'naquele', 'naqueles', 'nas', 'nem', 'nenhum', 'nenhuma', 'nessa', 'nessas', 'nesse', 'nesses', 'nesta', 'nestas', 'neste', 'nestes', 'ninguém', 'nível', 'no', 'noite', 'nome', 'nos', 'nós', 'nossa', 'nossas', 'nosso', 'nossos', 'nova', 'novas', 'nove', 'novo', 'novos', 'num', 'numa', 'número', 'nunca', 'o', 'obra', 'obrigada', 'obrigado', 'oitava', 'oitavo', 'oito', 'onde', 'ontem', 'onze', 'os', 'ou', 'outra', 'outras', 'outro', 'outros', 'para', 'parece', 'parte', 'partir', 'paucas', 'pela', 'pelas', 'pelo', 'pelos', 'pequena', 'pequenas', 'pequeno', 'pequenos', 'per', 'perante', 'perto', 'pode', 'pude', 'pôde', 'podem', 'podendo', 'poder', 'poderia', 'poderiam', 'podia', 'podiam', 'põe', 'põem', 'pois', 'ponto', 'pontos', 'por', 'porém', 'porque', 'porquê', 'posição', 'possível', 'possivelmente', 'posso', 'pouca', 'poucas', 'pouco', 'poucos', 'primeira', 'primeiras', 'primeiro', 'primeiros', 'própria', 'próprias', 'próprio', 'próprios', 'próxima', 'próximas', 'próximo', 'próximos', 'pude', 'puderam', 'quais', 'quáis', 'qual', 'quando', 'quanto', 'quantos', 'quarta', 'quarto', 'quatro', 'que', 'quê', 'quem', 'quer', 'quereis', 'querem', 'queremas', 'queres', 'quero', 'questão', 'quinta', 'quinto', 'quinze', 'relação', 'sabe', 'sabem', 'são', 'se', 'segunda', 'segundo', 'sei', 'seis', 'seja', 'sejam', 'sejamos', 'sem', 'sempre', 'sendo', 'ser', 'será', 'serão', 'serei', 'seremos', 'seria', 'seriam', 'seríamos', 'sete', 'sétima', 'sétimo', 'seu', 'seus', 'sexta', 'sexto', 'si', 'sido', 'sim', 'sistema', 'só', 'sob', 'sobre', 'sois', 'somos', 'sou', 'sua', 'suas', 'tal', 'talvez', 'também', 'tampouco', 'tanta', 'tantas', 'tanto', 'tão', 'tarde', 'te', 'tem', 'tém', 'têm', 'temos', 'tendes', 'tendo', 'tenha', 'tenham', 'tenhamos', 'tenho', 'tens', 'ter', 'terá', 'terão', 'terceira', 'terceiro', 'terei', 'teremos', 'teria', 'teriam', 'teríamos', 'teu',

'teus', 'teve', 'ti', 'tido', 'tinha', 'tinham', 'tínhamos', 'tive', 'tivemos', 'tiver', 'tivera', 'tiveram', 'tivéramos', 'tiverem', 'tivermos', 'tivesse', 'tivessem', 'tivéssemos', 'tiveste', 'tivestes', 'toda', 'todas', 'todavia', 'todo', 'todos', 'trabalho', 'três', 'treze', 'tu', 'tua', 'tuas', 'tudo', 'última', 'últimas', 'último', 'últimos', 'um', 'uma', 'umas', 'uns', 'vai', 'vais', 'vão', 'vários', 'vem', 'vêm', 'vendo', 'vens', 'ver', 'vez', 'vezes', 'viagem', 'vindo', 'vinte', 'vir', 'você', 'vocês', 'vos', 'vós', 'vossa', 'vossas', 'vosso', 'vossos', 'zero', '1', '2', '3', '4', '5', '6', '7', '8', '9', '0', '_', '.', (', ')'

ANEXO B - Lista de stopwords em inglês

"0o", "0s", "3a", "3b", "3d", "6b", "6o", "a", "a1", "a2", "a3", "a4", "ab", "able", "about", "above", "abst", "ac", "accordance", "according", "accordingly", "across", "act", "actually", "ad", "added", "adj", "ae", "af", "affected", "affecting", "affects", "after", "afterwards", "ag", "again", "against", "ah", "ain", "ain't", "aj", "al", "all", "allow", "allows", "almost", "alone", "along", "already", "also", "although", "always", "am", "among", "amongst", "amoungst", "amount", "an", "and", "announce", "another", "any", "anybody", "anyhow", "anymore", "anyone", "anything", "anyway", "anyways", "anywhere", "ao", "ap", "apart", "apparently", "appear", "appreciate", "appropriate", "approximately", "ar", "are", "aren", "arent", "aren't", "arise", "around", "as", "a's", "aside", "ask", "asking", "associated", "at", "au", "auth", "av", "available", "aw", "away", "awfully", "ax", "ay", "az", "b", "b1", "b2", "b3", "ba", "back", "bc", "bd", "be", "became", "because", "become", "becomes", "becoming", "been", "before", "beforehand", "begin", "beginning", "beginnings", "begins", "behind", "being", "believe", "below", "beside", "besides", "best", "better", "between", "beyond", "bi", "bill", "biol", "bj", "bk", "bl", "bn", "both", "bottom", "bp", "br", "brief", "briefly", "bs", "bt", "bu", "but", "bx", "by", "c", "c1", "c2", "c3", "ca", "call", "came", "can", "cannot", "cant", "can't", "cause", "causes", "cc", "cd", "ce", "certain", "certainly", "cf", "cg", "ch", "changes", "ci", "cit", "cj", "cl", "clearly", "cm", "c'mon", "cn", "co", "com", "come", "comes", "con", "concerning", "consequently", "consider", "considering", "contain", "containing", "contains", "corresponding", "could", "couldn", "couldnt", "couldn't", "course", "cp", "cq", "cr", "cry", "cs", "c's", "ct", "cu", "currently", "cv", "cx", "cy", "cz", "d", "d2", "da", "date", "dc", "dd", "de", "definitely", "describe", "described", "despite", "detail", "df", "di", "did", "didn", "didn't", "different", "dj", "dk", "dl", "do", "does", "doesn", "doesn't", "doing", "don", "done", "don't", "down", "downwards", "dp", "dr", "ds", "dt", "du", "due", "during", "dx", "dy", "e", "e2", "e3", "ea", "each", "ec", "ed", "edu", "ee", "ef", "effect", "eg", "ei", "eight", "eighty", "either", "ej", "el", "eleven", "else", "elsewhere", "em", "empty", "en", "end", "ending", "enough", "entirely", "eo", "ep", "eq", "er", "es", "especially", "est", "et", "et-al", "etc", "eu", "ev", "even", "ever", "every", "everybody", "everyone", "everything", "everywhere", "ex", "exactly", "example", "except", "ey", "f", "f2", "fa", "far", "fc", "few", "ff", "fi", "fifteen", "fifth", "fify", "fill", "find", "fire", "first", "five", "fix", "fj", "fl", "fn", "fo", "followed", "following", "follows", "for", "former", "formerly", "forth", "forty", "found", "four", "fr", "from", "front", "fs", "ft", "fu", "full", "further", "furthermore", "fy", "g", "ga", "gave", "ge", "get", "gets", "getting", "gi", "give", "given", "gives", "giving", "gj", "gl", "go", "goes", "going", "gone", "got", "gotten", "gr", "greetings", "gs", "gy", "h", "h2", "h3", "had", "hadn", "hadn't", "happens", "hardly", "has", "hasn", "hasnt", "hasn't", "have", "haven", "haven't", "having", "he", "hed", "he'd", "he'll", "hello", "help", "hence", "her", "here", "hereafter",

"hereby", "herein", "heres", "here's", "hereupon", "hers", "herself", "hes", "he's", "hh", "hi", "hid", "him", "himself", "his", "hither", "hj", "ho", "home", "hopefully", "how", "howbeit", "however", "how's", "hr", "hs", "http", "hu", "hundred", "hy", "i", "i2", "i3", "i4", "i6", "i7", "i8", "ia", "ib", "ibid", "ic", "id", "i'd", "ie", "if", "ig", "ignored", "ih", "ii", "ij", "il", "i'll", "im", "i'm", "immediate", "immediately", "importance", "important", "in", "inasmuch", "inc", "indeed", "index", "indicate", "indicated", "indicates", "information", "inner", "insofar", "instead", "interest", "into", "invention", "inward", "io", "ip", "iq", "ir", "is", "isn", "isn't", "it", "itd", "it'd", "it'll", "its", "it's", "itself", "iv", "i've", "ix", "iy", "iz", "j", "jj", "jr", "js", "jt", "ju", "just", "k", "ke", "keep", "keeps", "kept", "kg", "kj", "km", "know", "known", "knows", "ko", "l", "l2", "la", "largely", "last", "lately", "later", "latter", "latterly", "lb", "lc", "le", "least", "les", "less", "lest", "let", "lets", "let's", "lf", "like", "liked", "likely", "line", "little", "lj", "ll", "ll", "ln", "lo", "look", "looking", "looks", "los", "lr", "ls", "lt", "ltd", "m", "m2", "ma", "made", "mainly", "make", "makes", "many", "may", "maybe", "me", "mean", "means", "meantime", "meanwhile", "merely", "mg", "might", "mightn", "mightn't", "mill", "million", "mine", "miss", "ml", "mn", "mo", "more", "moreover", "most", "mostly", "move", "mr", "mrs", "ms", "mt", "mu", "much", "mug", "must", "mustn", "mustn't", "my", "myself", "n", "n2", "na", "name", "namely", "nay", "nc", "nd", "ne", "near", "nearly", "necessarily", "necessary", "need", "needn", "needn't", "needs", "neither", "never", "nevertheless", "new", "next", "ng", "ni", "nine", "ninety", "nj", "nl", "nn", "no", "nobody", "non", "none", "nonetheless", "noone", "nor", "normally", "nos", "not", "noted", "nothing", "novel", "now", "nowhere", "nr", "ns", "nt", "ny", "o", "oa", "ob", "obtain", "obtained", "obviously", "oc", "od", "of", "off", "often", "og", "oh", "oi", "oj", "ok", "okay", "ol", "old", "om", "omitted", "on", "once", "one", "ones", "only", "onto", "oo", "op", "oq", "or", "ord", "os", "ot", "other", "others", "otherwise", "ou", "ought", "our", "ours", "ourselves", "out", "outside", "over", "overall", "ow", "owing", "own", "ox", "oz", "p", "p1", "p2", "p3", "page", "pagecount", "pages", "par", "part", "particular", "particularly", "pas", "past", "pc", "pd", "pe", "per", "perhaps", "pf", "ph", "pi", "pj", "pk", "pl", "placed", "please", "plus", "pm", "pn", "po", "poorly", "possible", "possibly", "potentially", "pp", "pq", "pr", "predominantly", "present", "presumably", "previously", "primarily", "probably", "promptly", "proud", "provides", "ps", "pt", "pu", "put", "py", "q", "qj", "qu", "que", "quickly", "quite", "qv", "r", "r2", "ra", "ran", "rather", "rc", "rd", "re", "readily", "really", "reasonably", "recent", "recently", "ref", "refs", "regarding", "regardless", "regards", "related", "relatively", "research", "research-articl", "respectively", "resulted", "resulting", "results", "rf", "rh", "ri", "right", "rj", "rl", "rm", "rn", "ro", "rq", "rr", "rs", "rt", "ru", "run", "rv", "ry", "s", "s2", "sa", "said", "same", "saw", "say", "saying", "says", "sc", "sd", "se", "sec", "second", "secondly", "section", "see", "seeing", "seem", "seemed", "seeming", "seems", "seen", "self", "selves", "sensible", "sent", "serious", "seriously", "seven", "several", "sf", "shall", "shan", "shan't", "she", "shed", "she'd", "she'll", "shes", "she's", "should", "shouldn", "shouldn't", "should've", "show", "showed", "shown", "showns", "shows", "si", "side", "significant", "significantly", "similar", "similarly", "since", "sincere", "six", "sixty", "sj", "sl", "slightly", "sm", "sn", "so", "some", "somebody", "somehow", "someone", "somethan", "something", "sometime", "sometimes", "somewhat", "somewhere", "soon", "sorry", "sp", "specifically", "specified", "specify", "specifying", "sq", "sr", "ss", "st", "still", "stop", "strongly", "sub", "substantially", "successfully", "such", "sufficiently", "suggest", "sup", "sure", "sy", "system", "sz", "t", "t1", "t2", "t3", "take",

"taken", "taking", "tb", "tc", "td", "te", "tell", "ten", "tends", "tf", "th", "than", "thank", "thanks", "thanx", "that", "that'll", "thats", "that's", "that've", "the", "their", "theirs", "them", "themselves", "then", "thence", "there", "thereafter", "thereby", "thered", "therefore", "therein", "there'll", "thereof", "therere", "theres", "there's", "thereto", "thereupon", "there've", "these", "they", "theyd", "they'd", "they'll", "theyre", "they're", "they've", "thickv", "thin", "think", "third", "this", "thorough", "thoroughly", "those", "thou", "though", "thoughh", "thousand", "three", "throug", "through", "throughout", "thru", "thus", "ti", "til", "tip", "tj", "tl", "tm", "tn", "to", "together", "too", "took", "top", "toward", "towards", "tp", "tq", "tr", "tried", "tries", "truly", "try", "trying", "ts", "t's", "tt", "tv", "twelve", "twenty", "twice", "two", "tx", "u", "u201d", "ue", "ui", "uj", "uk", "um", "un", "under", "unfortunately", "unless", "unlike", "unlikely", "until", "unto", "uo", "up", "upon", "ups", "ur", "us", "use", "used", "useful", "usefully", "usefulness", "uses", "using", "usually", "ut", "v", "va", "value", "various", "vd", "ve", "ve", "very", "via", "viz", "vj", "vo", "vol", "vols", "volumtype", "vq", "vs", "vt", "vu", "w", "wa", "want", "wants", "was", "wasn", "wasnt", "wasn't", "way", "we", "wed", "we'd", "welcome", "well", "we'll", "well-b", "went", "were", "we're", "weren", "werent", "weren't", "we've", "what", "whatever", "what'll", "whats", "what's", "when", "whence", "whenever", "when's", "where", "whereafter", "whereas", "whereby", "wherein", "wheres", "where's", "whereupon", "wherever", "whether", "which", "while", "whim", "whither", "who", "whod", "whoever", "whole", "who'll", "whom", "whomever", "whos", "who's", "whose", "why", "why's", "wi", "widely", "will", "willing", "wish", "with", "within", "without", "wo", "won", "wonder", "wont", "won't", "words", "world", "would", "wouldn", "wouldnt", "wouldn't", "www", "x", "x1", "x2", "x3", "xf", "xi", "xj", "xk", "xl", "xn", "xo", "xs", "xt", "xv", "xx", "y", "y2", "yes", "yet", "yj", "yl", "you", "youd", "you'd", "you'll", "your", "youre", "you're", "yours", "yourself", "yourselves", "you've", "yr", "ys", "yt", "z", "zero", "zi", "zz"

ANEXO C - Lista de stopwords em espanhol

"a", "actualmente", "adelante", "además", "afirmó", "agregó", "ahora", "ahí", "al", "algo", "alguna", "algunas", "alguno", "algunos", "algún", "alrededor", "ambos", "ampleamos", "ante", "anterior", "antes", "apenas", "aproximadamente", "aquel", "aquellas", "aquellos", "aqui", "aquí", "arriba", "aseguró", "así", "atras", "aunque", "ayer", "añadió", "aún", "bajo", "bastante", "bien", "buen", "buena", "buenas", "bueno", "buenos", "cada", "casi", "cerca", "cierta", "ciertas", "cierto", "ciertos", "cinco", "comentó", "como", "con", "conocer", "conseguimos","conseguir", "considera", "consideró", "consigo", "consigue", "consiguen", "consigues", "contra", "cosas", "creo", "cual", "cuales", "cualquier", "cuando", "cuanto", "cuatro", "cuenta", "cómo", "da", "dado", "dan", "dar", "de", "debe", "deben", "debido", "decir", "dejó", "del", "demás", "dentro", "desde", "después", "dice", "dicen", "dicho", "dieron", "diferente", "diferentes", "dijeron", "dijo", "dio", "donde", "dos", "durante", "e", "ejemplo", "el", "ella", "ellas", "ello", "ellos", "embargo", "empleais", "emplean", "emplear", "empleas", "empleo", "en", "encima", "encuentra", "entonces", "entre", "era", "erais", "eramos", "eran", "eras", "eres", "es", "esa", "esas", "ese", "eso", "esos", "esta", "estaba", "estabais", "estaban", "estabas", "estad", "estada", "estadas", "estado", "estados", "estais", "estamos", "estan", "estando", "estar", "estaremos", "estará", "estarán", "estarás", "estaré", "estaréis", "estaría",

"estaríais", "estaríamos", "estarían", "estarías", "estas", "este", "estemos", "esto", "estos", "estoy", "estuve", "estuviera", "estuvierais", "estuvieran", "estuvieras", "estuvieron", "estuviese", "estuvieseis", "estuviesen", "estuvieses", "estuvimos", "estuviste", "estuvisteis", "estuviéramos", "estuviésemos", "estuvo", "está", "estábamos", "estáis", "están", "estás", "esté", "estéis", "estén", "estés", "ex", "existe", "existen", "explicó", "expresó", "fin", "fue", "fuera", "fuerais", "fueran", "fueras", "fueron", "fuese", "fueseis", "fuesen", "fueses", "fui", "fuimos", "fuiste", "fuisteis", "fuéramos", "fuésemos", "gran", "grandes", "gueno", "ha", "haber", "habida", "habidas", "habido", "habidos", "habiendo", "habremos", "habrá", "habrán", "habrás", "habré", "habréis", "habría", "habríais", "habríamos", "habrían", "habrías", "habéis", "había", "habíais", "habíamos", "habían", "habías", "hace", "haceis", "hacemos", "hacen", "hacer", "hacerlo", "haces", "hacia", "haciendo", "hago", "han", "has", "hasta", "hay", "haya", "hayamos", "hayan", "hayas", "hayáis", "he", "hecho", "hemos", "hicieron", "hizo", "hoy", "hube", "hubiera", "hubierais", "hubieran", "hubieras", "hubieron", "hubiese", "hubieseis", "hubiesen", "hubieses", "hubimos", "hubiste", "hubisteis", "hubiéramos", "hubiésemos", "hubo", "igual", "incluso", "indicó", "informó", "intenta", "intentais", "intentamos", "intentan", "intentar", "intentas", "intento", "ir", "junto", "la", "lado", "largo", "las", "le", "les", "llegó", "lleva", "llevar", "lo", "los", "luego", "lugar", "manera", "manifestó", "mayor", "me", "mediante", "mejor", "mencionó", "menos", "mi", "mientras", "mio", "mis", "misma", "mismas", "mismo", "mismos", "modo", "momento", "mucha", "muchas", "mucho", "muchos", "muy", "más", "mí", "mía", "mías", "mío", "míos", "nada", "nadie", "ni", "ninguna", "ningunas", "ninguno", "ningunos", "ningún", "no", "nos", "nosotras", "nosotros", "nuestra", "nuestras", "nuestro", "nuestros", "nueva", "nuevas", "nuevo", "nuevos", "nunca", "o", "ocho", "os", "otra", "otras", "otro", "otros", "para", "parece", "parte", "partir", "pasada", "pasado", "pero", "pesar", "poca", "pocas", "poco", "pocos", "podeis", "podemos", "poder", "podria", "podriais", "podriamos", "podrian", "podrias", "podrá", "podrán", "podría", "podrían", "poner", "por", "por qué", "porque", "posible", "primer", "primera", "primero", "primeros", "principalmente", "propia", "propias", "propio", "propios", "próximo", "próximos", "pudo", "pueda", "puede", "pueden", "puedo", "pues", "que", "quedó", "queremos", "quien", "quienes", "quiere", "quién", "qué", "realizado", "realizar", "realizó", "respecto", "sabe", "sabeis", "sabemos", "saben", "saber", "sabes", "se", "sea", "seamos", "sean", "seas", "segunda", "segundo", "según", "seis", "ser", "seremos", "será", "serán", "serás", "seré", "seréis", "sería", "seríais", "seríamos", "serían", "serías", "seáis", "señaló", "si", "sido", "siempre", "siendo", "siete", "sigue", "siguiente", "sin", "sino", "sobre", "sois", "sola", "solamente", "solas", "solo", "solos", "somos", "son", "soy", "su", "sus", "suya", "suyas", "suyo", "suyos", "sí", "sólo", "tal", "también", "tampoco", "tan", "tanto", "te", "tendremos", "tendrá", "tendrán", "tendrás", "tendré", "tendréis", "tendría", "tendríais", "tendríamos", "tendrían", "tendrías", "tened", "teneis", "tenemos", "tener", "tenga", "tengamos", "tengan", "tengas", "tengo", "tengáis", "tenida", "tenidas", "tenido", "tenidos", "teniendo", "tenéis", "tenía", "teníais", "teníamos", "tenían", "tenías", "tercera", "ti", "tiempo", "tiene", "tienen", "tienes", "toda", "todas", "todavía", "todo", "todos", "total", "trabaja", "trabajais", "trabajamos", "trabajan", "trabajar", "trabajas", "trabajo", "tras", "trata", "través", "tres", "tu", "tus", "tuve", "tuviera", "tuvierais", "tuvieran", "tuvieras", "tuvieron", "tuviese", "tuvieseis", "tuviesen", "tuvieses", "tuvimos", "tuviste", "tuvisteis", "tuviéramos", "tuviésemos", "tuvo", "tuya", "tuyas", "tuyo", "tuyos", "tú", "ultimo", "un", "una", "unas", "uno", "unos", "usa", "usais", "usamos", "usan", "usar", "usas",

"uso", "usted", "va", "vais", "valor", "vamos", "van", "varias", "varios", "vaya", "veces", "ver", "verdad", "verdadera", "verdadero", "vez", "vosotras", "vosotros", "voy", "vuestra", "vuestras", "vuestro", "vuestros", "y", "ya", "yo", "él", "éramos", "ésta", "éstas", "éste", "éstos", "última", "últimas", "último", "últimos"