Universidade Federal da Bahia
Instituto de Matemática e Estatística

Programa de Pós-Graduação em Ciência da Computação

ANOTAÇÕES SEMÂNTICAS EM REPOSITÓRIOS ACADÊMICOS: UM ESTUDO DE CASO COM O RI UFBA

Aline Meira Rocha

DISSERTAÇÃO DE MESTRADO

Salvador
03 de março de 2020

ALINE MEIRA ROCHA

ANOTAÇÕES SEMÂNTICAS EM REPOSITÓRIOS ACADÊMICOS: UM ESTUDO DE CASO COM O RI UFBA

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Profª Dra. Lais do Nascimento Salvador
Co-orientador: Prof. Dr. Marlo Vieira dos Santos e Souza

Salvador
03 de março de 2020

Sistema de Bibliotecas - UFBA

Rocha, Aline M..

Anotações Semânticas em Repositórios Acadêmicos: um estudo de caso com o RI UFBA / Aline Meira Rocha – Salvador, 2019.

90p.: il.

Orientadora: Prof. Dr. Profª Dra. Lais do Nascimento Salvador.

Co-orientador: Prof. Dr. Prof. Dr. Marlo Vieira dos Santos e Souza.

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática e Estatística, 2019.

1. Web Semântica. 2. Anotações Semânticas. 3. Aprendizado de Máquina. 4. Classificação Textual. 5. Extração de Palavras-Chave.. I. Salvador, Lais N. . II. Souza, Marlo. III. Universidade Federal da Bahia. Instituto de Matemática e Estatística. IV. Título.

CDD – XXX.XX

CDU – XXX.XX.XXX

TERMO DE APROVAÇÃO

ALINE MEIRA ROCHA

ANOTAÇÕES SEMÂNTICAS EM REPOSITÓRIOS ACADÊMICOS: UM ESTUDO DE CASO COM O RI UFBA

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 03 de março de 2020

Profª. Dra. Lais do Nascimento Salvador
Departamento de Ciência da Computação - IME
Universidade Federal da Bahia

Profª. Dra. Daniela Barreiro Claro
Departamento de Ciência da Computação - IME
Universidade Federal da Bahia

Profª. Dra. Flávia Goulart Mota Garcia Rosa
EDUFBA - Editora UFBA
Universidade Federal da Bahia

Dedico esse trabalho a meus pais, a minha irmã e a todos que me incentivaram. Sem essa força não teria conseguido concluir mais essa jornada.

AGRADECIMENTOS

Gostaria de agradecer a Deus pela proteção e pela força em mais essa caminhada. A minha família e amigos por terem sabido enfrentar a minha ausência e pelo apoio recebido.

A minha orientadora Profª Lais Salvador e meu co-orientador Prof. Marlo Vieira pela ajuda, paciência, direcionamentos na pesquisa, disponibilidade e pelas orientações prestadas. O meu mais sincero obrigado.

A todos meus colegas de PGCOMP por compartilharem momentos de estudo, de pesquisa, amizade, força e aprendizado.

Aos meus colegas de STI por terem compreendido minhas ausências e por sempre me incentivarem durante todo o período do mestrado.

Aos Grupos de Pesquisa FORMAS (*Formalisms and Semantic Applications Research Group*) e Onda Digital por terem contribuído de maneira essencial em vários momentos dessa árdua jornada.

Ao Núcleo Tecnológico do SIBI pelo apoio durante o projeto e as bibliotecárias Eleonora, Livia e Solange na realização do estudo de caso.

Ao Comitê Gestor do RI UFBA pela liberação do acesso à base de dados do RI UFBA e a CRI/STI pelo auxilio na montagem do ambiente para realização dos experimentos.

A todas as pessoas que me apoiaram e me incentivaram a concluir o curso.

A tecnologia move o mundo.

—STEVE JOBS

RESUMO

Anotações Semânticas permitem enriquecer os metadados de um documento, o que facilita a recuperação do mesmo pelos mecanismos de busca. Por sua vez, Repositórios Institucionais (RI) são repositórios acadêmicos que possibilitam o armazenamento e a divulgação das produções científicas de universidades e centros de pesquisa. As informações sobre cada item depositado são armazenadas em seus metadados, mas como usualmente isso é feito de maneira manual pelo próprio pesquisador nem sempre os termos escolhidos ajudam nessa descrição, o que leva à intervenção dos bibliotecários no processo. A anotação semântica de metadados referentes à comunidade, subcomunidade e palavras-chave permite o enriquecimento das descrições de itens do RI, além de facilitar o processo de recuperação. O objetivo geral desse trabalho é desenvolver uma solução para realizar a anotação semântica de maneira semiautomática em um RI de forma a auxiliar o trabalho dos bibliotecários durante a validação dos metadados de cada publicação. Para isso, a sugestão de palavras-chave durante a validação dos metadados identificaria termos representativos de cada publicação e enriqueceria semanticamente esses metadados, favorecendo a recuperação dos itens em um RI. Já métodos de aprendizagem de máquina de classificação textual binária podem sugerir que uma publicação também seja associada a outra subcomunidade caso seja identificado que se trata de trabalho multidisciplinar. Através da implementação de um classificador multi-hierárquico também é possível identificar comunidades e subcomunidade de publicações ainda não depositados no RI. Para atingir o objetivo proposto, os seguintes passos foram executados: (i) montagem de um ambiente de teste contendo um conjunto de documentos do RI UFBA com seus respectivos metadados e implementação de classificadores multi-hierárquicos e binários; (ii) avaliação dos classificadores a fim de identificar quais apresentam os melhores resultados; (iii) implementação do extrator de palavras-chaves; (iv) realização de um estudo de caso no RI da UFBA, no qual as palavras-chave extraídas passaram pela validação de especialistas de domínio, no caso as bibliotecárias do Sistema de Bibliotecas da UFBA (SIBI) e (vi) anotação semântica dos resultados obtidos no experimento dos classificadores e no estudo de caso. Os resultados obtidos mostram que a classificação multi-hierárquica teve um bom desempenho, sendo que o algoritmo de Naive Bayes apresentou os melhores resultados, com os valores das métricas acima de 85% no 1º nível e acima de 80% nos comunidades, com destaque na comunidade IME, na qual atingiu valores acima de 96%. Na classificação binária foram observados resultados promissores, dada a complexidade da tarefa: foram retornados treze (13) trabalhos de vinte e oito (28) identificados como multidisciplinares, considerando o conjunto de documentos utilizado nos experimentos. Já no estudo de caso foi avaliado que as palavras-chave sugeridas foram adequadas. Por fim, foi realizada a anotação semântica utilizando o padrão RDF do Dublin Core a partir dos resultados obtidos na classificação textual e validação das

sugestões de palavras-chave. O uso do classificador binário mostra um caminho para identificação de trabalhos multidisciplinares, campo pouco explorado na literatura, por sua vez o classificador multi-hierárquico pode ser usado em situações de povoamento de RI. A sugestão de palavras-chave auxiliaria à complementação da palavras-chaves realizadas pelos bibliotecários durante a validação dos metadados de cada documento. Por sua vez, as sugestões de comunidade, subcomunidade e palavra-chave podem ser anotados semanticamente no documento do RI com a finalidade de enriquecimento de seus metadados.

Palavras-chave: anotações semânticas. repositórios acadêmicos. aprendizado de máquina. classificação textual. extração de palavras-chave.

ABSTRACT

Semantic annotations allow you to enrich a document's metadata, which facilitates its retrieval by search engines. In turn, Institutional Repositories (IR) are academic repositories that enable the storage and dissemination of scientific productions from universities and research centers. The information about each deposited item is stored in its metadata, but as this is usually done manually by the researcher himself, the terms chosen do not always help in this description, which leads to the intervention of librarians in the process. Semantic annotation of metadata for the community, subcommunity, and keywords allows enrichment of RI item descriptions, as well as facilitating the recovery process. The general objective of this work is to develop a solution to perform semantic annotation in a semi-automatic manner in an IR in order to assist the work of librarians during the validation of the metadata for each publication. For this, the suggestion of keywords during the validation of the metadata would identify representative terms of each publication and would semantically enrich these metadata, favoring the recovery of the items in an IR. Learning methods for a binary textual classification machine may suggest that a publication is also associated with another subcommunity if it is identified that it is multidisciplinary work. Through the implementation of a multi-hierarchical classifier, it is also possible to identify communities and subcommunities for publications not yet deposited in RI. To achieve the proposed objective, the following steps were performed: (i) setting up a test environment containing a set of RI UFBA documents with their respective metadata and implementing multi-hierarchical and binary classifiers; (ii) evaluation of the classifiers to identify which ones present the best results; (iii) implementation of the keyword extractor; (iv) conducting a case study at UFBA RI, in which the extracted keywords were validated by domain experts, in this case, the librarians of the UFBA Library System (SIBI) and (vi) semantic annotation of the results obtained in the classifier experiment and the case study. The results obtained show that the multi-hierarchical classification had a good performance, and the Naive Bayes algorithm showed the best results, with the values of the metrics above 85% in the 1st level and above 80% in the communities, with emphasis on IME community, in which it reached values above 96%. In the binary classification, promising results were observed, given the complexity of the task: thirteen (13) papers from twenty-eight (28) identified as multidisciplinary were returned, considering the set of documents used in the experiments. In the case study, it was assessed that the suggested keywords were adequate. Finally, the semantic annotation was performed using the Dublin Core RDF standard based on the results obtained in the textual classification and validation of keyword suggestions. The use of the binary classifier shows a way to identify multidisciplinary works, a field little explored in the literature, in turn, the multi-hierarchical classifier can be used in IR population situations. The suggestion of keywords would help to complement the

keywords made by librarians during the validation of the metadata of each document. In turn, suggestions for community, subcommunity, and keyword can be noted semantically in the RI document for the purpose of enriching their metadata.

Keywords: semantic annotation. academic repositories. machine learning. text classification. keyword extraction.

LISTA DE ABREVIATURAS

BOW - Bag of Words
CRI - Coordenação de Redes e Infraestrutura
DC - Dublin Core
DCMES - Dublin Core Metadata Element Set
EDUFBA - Editora da Universidade Federal da Bahia
FN - False Negative
FOAF - Friend of a Friend
FP - False Positive
HP - Hewlett-Packard
HTML - Hypertext Markup Language
IA - Inteligência Artificial
IBICT - Instituto Brasileiro de Informação em Ciência e Tecnologia
LOD - Linked Open Data
MIT - Massachusetts Institute of Technology
NLTK - Natural Language Toolkit
OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
PDF - Portable Document Format
PLN - Processamento de Linguagem Natural
POS - Part-of-Speech
QDC - Qualified Dublin Core
RDF - Resource Description Framework
RDFa - Resource Description Framework in Attributes
RI - Repositório Institucional
SGBD - Sistema Gerenciador de Banco de Dados
SPARQL - Simple Protocol and RDF Query Language
SQL - Structured Query Language
SIBI - Sistema Universitário de Bibliotecas
STI - Superintendência de Tecnologia da Informação
SVM - Support Vector Machine
TCC - Trabalho de Conclusão de Curso
TF - Term Frequency
TF-IDF - Term Frequency–Inverse Document Frequency
TN - True Negative
TP - True Positive
UFBA - Universidade Federal da Bahia
URI - Uniform Resource Identifier
W3C - World Wide Web Consortium

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Este capitulo objetiva apresentar a introdução desta dissertação, bem como o problema identificado, a hipótese levantada, os objetivos geral e específicos, a metodologia e a estrutura utilizada no desenvolvimento desta dissertação.

INTRODUÇÃO

1.1 CONTEXTO

Em termos acadêmicos, pode ser dizer que os Repositórios Institucionais (RI) são utilizados para armazenar as produções cientificas de instituições de ensino e de pesquisa. Um RI é capaz de aumentar a visibilidade destas produções (FARID; KHAN; JAVED, 2013), proporcionando uma maior transparência aos investimentos destinados à ciência (SAYÃO et al., 2009).

Parte da produção científica Universidade Federal da Bahia (UFBA) está armazenada em seu RI, que foi implementado através da ferramenta *DSpace*[1]. Atualmente, o RI UFBA[2] conta com cerca de 27.996 (vinte e sete mil novecentos e noventa e seis) itens depositados[3] em suas comunidades - o que reflete sua estrutura organizacional. Estas comunidades, que podem ser subdivididas em subcomunidades, possuem coleções compostas por itens (unidade informacional do *DSpace*) (SHINTAKU; MEIRELLES, 2010).

Atualmente, a gestão do RI UFBA é feita pelo Comitê Gestor[4], que é órgão responsável pela política do Repositório e sua administração - permissão dos usuários, gerência da validação dos metadados, entre outros - é realizada pelo Núcleo Tecnológico, localizado na Biblioteca Universitária de Ciências e Tecnologias Professor Omar Catunda pertencente ao SIBI. A Tabela 1.1 a seguir apresenta as informações gerais da estrutura do Núcleo Tecnológico do SIBI:

[1]https://duraspace.org/dspace/

[2]https://repositorio.ufba.br/ri/

[3]Segundo http://oasisbr.ibict.br/ em fev20

[4]Composto pela Editora da Universidade Federal da Bahia (EDUFBA), a Pró-Reitoria de Pesquisa, Criação e Inovação (PROPCI), a Pró-Reitoria de Ensino de Pós-Graduação (PROPG), a Superintendência de Tecnologia e Informação (STI), o Sistema de Bibliotecas (SIBI) e o Instituto de Ciência da 45Informação (ICI).

Tabela 1.1 Estrutura do Núcleo Tecnológico - SIBI

Núcleo Tecnológico - SIBI	
Setor responsável por consultoria em projetos na área tecnológica no SIBI.	
Criação:	2017
Gerencia:	RI UFBA, Portal SEER, *Pergamum*, site SIBI, redes sociais
Equipe:	02 Bibliotecárias, 02 Assistentes Administrativos, 01 Administrador

Fonte: SIBI (2019).

Com o *DSpace* é possível realizar o autoarquivamento, ou seja, o próprio pesquisador pode realizar o depósito de sua publicação enquanto um bibliotecário responsabiliza-se pela validação dos metadados informados (ROSA; MEIRELLES; PALACIOS, 2011). Após a validação, as publicações ficam disponíveis para os usuários finais.

A Figura 1.1 mostra o Fluxo de Submissão do *DSpace*, desde o cadastro do pesquisador que solicita permissão para depositar a publicação na subcomunidade desejada, passando pelo depósito da publicação (que é realizada pelo próprio pesquisador), pela validação dos metadados de cada publicação (realizada pelo bibliotecário gestor de cada comunidade) até culminar na sua disponibilização no RI.

Figura 1.1 Fluxo de Submissão no *DSpace*

Fonte: elaborada pela autora (2019)[5]

.

O RI UFBA está disponível na *Web* Sintática e sua busca nem sempre gera resultados relevantes. Os itens depositados neste repositório possuem metadados que ajudam a descrevê-los, no entanto, nem sempre estes são preenchidos com os termos mais adequados, favorecendo uma identificação eficiente das informações.

Os metadados servem para identificar e descrever um determinado documento, facilitando sua recuperação durante a busca. Dependendo do tipo de documento, torna-se possível utilizar um conjunto de metadados específicos, sendo que um dos padrões mais utilizados atualmente é o *Dublin Core*(HILLMANN, 2008), que é o mesmo utilizado pelo *DSpace*.

1.2 MOTIVAÇÃO

Os Repositórios acadêmicos possibilitam o armazenamento de produções cientificas das universidades e dos centros de pesquisa. Eles permitem o autoarquivamento, no qual o

[5]Ícones obtidos gratuitamente no site https://www.flaticon.com/

próprio pesquisador pode efetuar o depósito de sua publicação. Vale destacar que, por mais que os metadados de cada publicação sejam preenchidos pelo próprio pesquisador, nem sempre os termos escolhidos para descrevê-la são os mais relevantes e adequados. Além disso, a falta de conhecimento multidisciplinar por parte dos bibliotecários, responsáveis por validarem estes metadados, impacta nesta validação.

Extrair termos relevantes destes conteúdos, pode ajudar a descrevê-los mais adequadamente (EDMUNDSON, 1969) e o ideal seria que estes termos fossem sugeridos durante a validação dos metadados. Como os metadados são preenchidos manualmente, apesar da validação dos bibliotecários, muitas vezes, por falta de conhecimento multidisciplinar, esta validação torna-se ineficiente. Em contrapartida, ela se tornaria mais eficiente caso os termos mais relevantes de cada documento fossem sugeridos no decorrer da validação. Provavelmente, isso otimizaria a busca, considerando que os metadados seriam enriquecidos semanticamente.

Por sua vez, a classificação textual serve para atribuir textos (ou documentos) a uma ou mais categorias pré-definidas, de acordo com o seu conteúdo (LEWIS, 1992). Com isso, torna-se possível classificar publicações ainda não depositadas no RI (classificação multi-hierárquica) e que não estejam organizadas por subcomunidade bem como sugerir que uma publicação seja associada a outra subcomunidade, caso seja possível identificar que se trate de trabalho multidisciplinar(classificação binária).

A identificação de trabalhos multidisciplinares pode contribuir com a redução do isolamento da pesquisa, afinal, pesquisadores que estejam desenvolvendo trabalhos na mesma área podem trabalhar em conjunto.

As Anotações Semânticas possibilitam que os metadados de um documento sejam enriquecidos semanticamente, facilitando a sua recuperação (OREN et al., 2006). Com isso, propõe-se, a realização de experimentos através da classificação textual (multi-hierárquica e binária) e da extração de palavras-chave de um conjunto de documentos do RI UFBA, gerando a anotação semântica dos metadados referentes a subcomunidades e as palavras-chave desses documentos.

No decorrer desta dissertação, serão apresentados os resultados obtidos através de 2 experimentos e um estudo de caso, objetivando apresentar uma solução para as anotações de forma semiautomática dos itens de um RI (Repositório Institucional).

1.3 PROBLEMA DE PESQUISA

Levando em conta este projeto de pesquisa, foram identificados os seguintes problemas e hipóteses de pesquisa:

P1: Quais as formas de anotar um RI semanticamente?
H1: A anotação semântica é uma possível forma de anotar um RI.

P2: Como a anotação semântica auxiliaria os bibliotecários?
H2: As anotação semântica semiautomática auxiliaria o trabalho dos bibliotecários durante a validação dos metadados das produções cientificas depositadas em um RI.

P3: Como classificar trabalhos multidisciplinares?
H3: Os métodos de aprendizado de máquina de classificação textual binária permitiria identificar e sugerir trabalhos que possam pertencer a mais de uma área de conhecimento.

P4: Como classificar itens não depositados em um RI?
H4: Ao identificar os termos mais relevantes de cada comunidade e subcomunidade é possível treinar um classificador multi-hierárquico para classificar novos documentos.

1.4 OBJETIVOS

1.4.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver uma solução para realizar a anotação semântica de maneira semiautomática em um RI de forma a auxiliar o trabalho dos bibliotecários durante a validação dos metadados de cada publicação.

1.4.2 Objetivos Específicos

Para atingir o objetivo geral desta pesquisa, foram definidos os seguintes objetivos específicos:

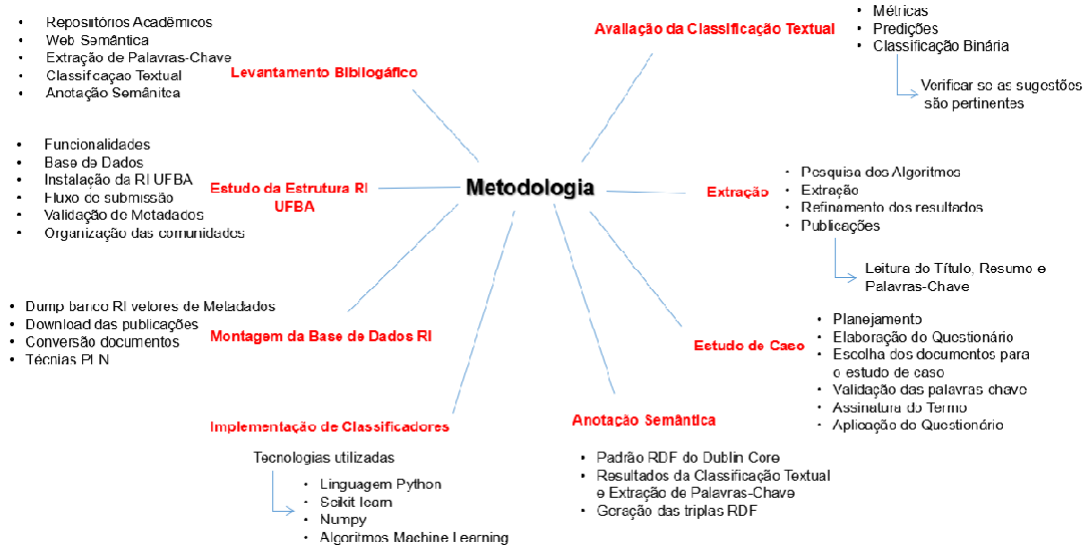
- Organizar os trabalhos em comunidades e subcomunidades a partir dos termos relevantes extraídos;
- Identificar trabalhos multidisciplinares, isto é, trabalhos que pertençam a mais de uma subcomunidade;
- Investigar os métodos de classificação textual para anotação semântica;
- Investigar métodos de Extração de palavras-chave para anotação semântica;
- Definir os itens de um RI utilizando o padrão RDF do *Dublin Core*;
- Avaliar parte da solução com profissionais do SIBI;
- Minimizar o esforço de classificar novos documentos por bibliotecários;
- Validar a área de conhecimento na qual o documento foi incluído;
- Investigar métodos de sugestões/extração de palavras-chave em documentos.

1.5 METODOLOGIA ADOTADA

Visando atingir os objetivos propostos por este trabalho, foram definidos os passos metodológicos listados nesta seção. Ressalta-se que o primeiro passo corresponde à atividade iterativa aplicada no decorrer deste trabalho.

A Figura 1.2 apresenta uma visão geral da metodologia adotada no decorrer desse trabalho:

Figura 1.2 Esquema da metodologia utilizada



Fonte: elaborada pela autora (2019).

A seguir são descritos com mais detalhes os passos metodológicos percorridos:

- **(1) Realização de levantamento bibliográfico** Realização de uma pesquisa em diversas bases de dados, considerando os trabalhos desenvolvidos em cada área de pesquisa. A principio, foi realizada uma pesquisa por trabalhos na área de Ontologia e Anotações Semânticas que tenham como objeto de estudo, os Repositórios Institucionais (ou Acadêmicos). Também foram pesquisados trabalhos cuja a extração de texto utiliza as técnicas de PLN (Processamento de Linguagem Natural), de classificação textual e de extração de palavras-chave.

  Estes trabalhos serviram de apoio para a fundamentação teórica deste projeto, favorecendo verificar como se encontrava o estado da arte nestas áreas, identificando possíveis lacunas de pesquisa.

- **(2) Estudo da estrutura de um repositório acadêmico** Estudo da estrutura de um Repositório Institucional Acadêmico, destacando como ele é organizado e mapeando suas principais características e funcionalidades. Verificou-se como foi feita a instalação do RI UFBA, analisando a existência de alguma alteração na instalação padrão, como customizações, adição de plugins, entre outros.

- **(3) Montagem da base de dados para realização dos experimentos** A base de dados para a realização dos experimentos é composta do *dump* do Sistema Gerenciador de Banco de Dados (SGBD) - *PostgreSQL* - no qual estão armazenadas as informações das publicações depositadas no RI UFBA e *downloads* de cerca 153 (cento e cinquenta e três) publicações escritas na Língua Portuguesa (teses e dissertações), em formato PDF *Portable Document Format*, coletadas em 3 (três) comunidades.

  - **Conversão dos textos** Conversão dos documentos baixados para os experimentos do formato pdf para o formato txt, com o auxílio da biblioteca para extração de texto, *pdfminer*, codificada em *Python*, em sua versão .six. O formato de texto mencionado facilitou o tratamento destes textos e a aplicação de técnicas de PLN.
  - **Processamento de Linguagem Natural** Para realizar o pré-processamento destes textos, foi utilizada a biblioteca *Natural Language Toolkit* (NLTK)[6] também codificada em *Python*, que permite a tokenização dos textos (quebra do texto em *tokens* (palavras)), extração de prefixos ou sufixos, remoção de *stopwords* (palavras pouco relevantes no texto), dentre outras tarefas de pré-processamento de texto. Foi gerada a *Bag of Words* contendo a frequência das 1000 palavras mais comumente utilizadas nos conjunto dos documentos.

- **(4) Implementação de classificadores** Foram implementados classificadores multi-hierárquicos e binários para a realização da classificação textual. Tais classificadores tiveram como entrada a *Bag of Words* gerada e os vetores de metadados, como saída, foram utilizadas as predições, a fim de classificar as publicações de acordo com a comunidade e subcomunidade corretas e identificar trabalhos multidisciplinares.

  Para implementação dos classificadores foi utilizada a biblioteca *scikit-learn* (codificada em *Python*), que é bastante utilizada na realização das tarefas de *Machine Learning*. Estes classificadores foram implementados utilizando os algoritmos de classificação *Naive Bayes*, *Support Vector Machine* (SVM) e *Decision Tree*.

  - **Implementação de Classificador Multi-Hierárquico** Foi implementado um Classificador Multi-Hierárquico com o intuito de verificar se uma publicação foi depositada em uma comunidade ou subcomunidade de maneira incorreta. Esta situação pode ocorrer nos casos em que o pesquisador possua mais de um vinculo com a instituição, o que favorece que o depósito aconteça erradamente.
  - **Implementação de Classificador Binário** Foi implementado um Classificador Binário, a fim de verificar a possibilidade de sugerir, através de predições, se uma publicação pode ser associada a uma dada subcomunidade, caso se trate de uma publicação de natureza multidisciplinar. Esta situação é identificada nos casos em que predições sejam 1, porém, no conjunto de teste, este valor consta como 0.

- **(5) Avaliação da Classificação Textual** Para avaliação dos classificadores implementados, foi utilizado um experimento de classificação textual. Estes classificadores foram avaliados levando em conta as métricas acurácia, revocação *recall* e o F1-*score*. Com base nisso, foram considerados os resultados das predições, com a comparação das classes reais e das classes preditas no conjunto de teste.

[6] http://nltk.org

- **(6) Extração de palavras-chave** Os textos foram submetidos a métodos de extração de palavras-chave propostos na literatura para identificação de termos representativos em cada documento.
- **(7) Realização de um estudo de caso com as bibliotecárias do SIBI/UFBA** As palavras-chave extraídas passaram pela validação da bibliotecária responsável por cada comunidade.
- **(8) Anotação Semântica dos itens de um RI** Os resultados obtidos através do experimento da classificação textual binária e do estudo de caso com a validação das palavras-chave pelas bibliotecárias, foram anotados com o RDF do padrão de metadados *Dublin Core*.

1.6 ORGANIZAÇÃO DA DISSERTAÇÃO

Este trabalho será dividido em 6 capítulos. O Capítulo 2 apresentará a Fundamentação Teórica acerca dos conceitos e tecnologias relacionados à Classificação Textual (Processamento de Linguagem Natural e Aprendizado de Máquina), à Extração de palavras-chave, às Anotações Semânticas (padrão RDF, ontologias e o padrão de metadados *Dublin Core*), bem como os Repositórios Institucionais.

O Capítulo 3 apresentará trabalhos relacionados aos temas desta dissertação, por sua vez, o Capítulo 4 apresentará uma solução acerca deste trabalho, contendo a descrição da arquitetura e etapas. O Capítulo 5 descreverá os experimentos e estudo de caso realizados, contendo a descrição das atividades executadas em cada um deles e sua avaliação. As Anotações Semânticas obtidas também serão apresentadas neste capítulo.

Por fim, o Capitulo 6 apresentará as considerações finais, bem como as contribuições alcançadas, as publicações realizadas e as sugestões para trabalhos futuros. Após este capítulo, serão listadas as referências utilizadas neste trabalho, que será encerrado com os apêndices.

Capítulo

2

Neste capitulo será apresentada a fundamentação teórica dos temas concernentes a esta dissertação: Classificação Textual (seção 2.1), Anotações Semânticas (seção 2.2), Extração de Palavras-Chave (seção 2.3) e Repositórios Institucionais (seção 2.4).

FUNDAMENTAÇÃO TEÓRICA

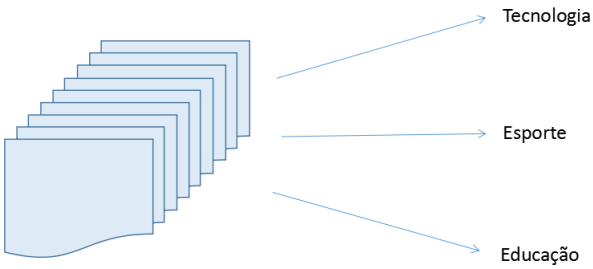
Como este trabalho é de natureza multidisciplinar, neste capítulo será abordada a fundamentação teórica de seus temas.

Inicialmente, será abordada a classificação textual, visto que um dos experimentos executados na solução implementada realiza a classificação de documentos de um RI. Com a apresentação das técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina, em seguida, é abordada a extração de palavras-chave e alguns métodos identificados na literatura; a anotação semântica, bem como suas formas de representação pelos padrões DC e RDF. Por fim, são apresentados os RI's.

2.1 CLASSIFICAÇÃO TEXTUAL

A classificação textual serve para atribuir textos (ou documentos) a uma ou mais categorias previamente definidas levando em conta cada conteúdo (LEWIS, 1992). Para isso, faz-se necessário associar os termos mais relevantes de cada categoria, bem como comparar os termos mais frequentemente utilizados em cada texto (GUTHRIE; WALKER, 1994).

Figura 2.1 Classificação Textual



Fonte: elaborado pela autora (2019).

A Figura 2.1, ilustra a classificação textual, demonstrando que os textos podem ser classificados em categorias pré-definidas, tais como Tecnologia, Esporte e Educação, levando em conta o teor dos conteúdos.

A classificação textual pode ser dividida em multi-classe (*Multi-Class Classification*), quando existe mais de uma classe, mas o texto só pode pertencer a uma delas, multi-label (*Multi-Label Classification*) na qual cada texto pode pertence a mais de uma classe (SEBASTIANI, 2001), multi-hierárquica (*Hierarchical Multi-Label Classification*), na qual um texto pertence a uma hierarquia de classes, evidenciando a necessidade de que ele seja concomitantemente classificado em 2 classes distintas, que obedecem a uma dada hierarquia (VENS et al., 2008) e binária (*Binary Classification*) que identifica se um texto pertence ou não a uma dada categoria (SEBASTIANI, 2001).

A classificação de textos é um dos exemplos existentes no aprendizado supervisionado, no qual um conjunto de documentos e seus rótulos (categorias) são treinados com a finalidade de predizer as categorias de um novo conjunto de documentos, sendo bastante utilizada na Análise de Sentimentos (PANG; LEE; VAITHYANATHAN, 2002), no qual é verificado se um texto possui contexto positivo ou negativo.

Para realizar a classificação textual, é preciso utilizar várias técnicas, dentre as quais se destacam o Processamento de Linguagem Natural e o Aprendizado de Máquina.

2.1.1 Processamento de Linguagem Natural (PLN)

Segundo (OTHERO, 2006), a área de Linguística Computacional estabelece uma relação entre a Linguística e a Computação; em contrapartida, o Processamento de Linguagem Natural (PLN) constrói sistemas focados em processamento textual, de modo que o computador entenda textos escritos em linguagem humana e se comunique de forma automática (RUSSELL; NORVIG, 2009).

O PLN preocupa-se com a fonética (som das palavras), com os aspectos léxicos (*strings + normalização* (palavras)), morfologia (classe das palavras), sintática (estrutura do que se fala), semântica (significado das palavras) e pragmática (intenção sobre o que falar)(OTHERO, 2006).

A área de PLN envolve diversos conceitos, tais como:

- **Corpus e Corpora**
  O *corpus* compreende o conjunto de textos que podem ser rotulados ou não; já o Corpora engloba o conjunto destes *corpus*.

- **Tokenização**
  A tokenização consiste em dividir um texto em pedaços menores, como sentenças ou palavras (RUSSELL; NORVIG, 2009). A sentença "Estou bem mas não tenho certeza se vou conseguir viajar amanhã", quando tokenizada, produz o resultado ilustrado na tabela 2.2 a seguir.

Tabela 2.1 Texto tokenizado

Estou	bem	mas	não	tenho	certeza	se	vou	conseguir	viajar	amanhã

Fonte: elaborado pela autora (2019).

Como é possível observar, o texto foi dividido pelas palavras que o compõe.

- **Normalização**
  A normalização envolve um conjunto de tarefas que visam corrigir problemas de acentuação e possíveis inconsistências.

- **Pré-Processamento** O pré-processamento envolve a eliminação de caracteres de pontuação, de números, palavras com poucos caracteres e das *stopwords*.

- ***Stopwords***
  As *stopwords* são palavras que sem relevância em um texto e que, por não possuírem valor semântico, podem ser descartadas do texto. As *stopwords* geralmente são representadas por conectivos e termos auxiliares. Existe um conjunto de *stopwords* especifico para cada idioma. No cado da Língua Portuguesa, pode-se citar: 'de', 'a', 'o', 'que', 'e', 'do', 'da', 'em', 'um', 'para', 'com', 'não', 'uma', 'os', 'no', 'se', 'na', 'por', 'mais', 'as', 'dos', 'como', 'mas', 'ao', 'ele', 'das', 'à', 'seu', 'sua', 'ou'

- ***N-grams***
  No modelo *n-gram*, cada sequência de "n"termos possui uma probabilidade de ocorrer. Com isso, pode-se prever o próximo termo, através do conhecimento dos termos anteriores (BROWN et al., 1992). Já no modelo 1-*grams* (ou *unigram*), cada elemento do vocabulário representa apenas uma palavra. O modelo 2-*grams*, representa 2 palavras, o modelo 3-*grams* representa 3 palavras, e assim sucessivamente, sendo que "n"indica a quantidade de palavras que aparecem juntas.

Tabela 2.2 Exemplo de *n-gram*

1-gram	Hoje
2-gram	Hoje sabemos
3-gram	Hoje sabemos mais
n-gram	Hoje sabemos mais que ...

Fonte: elaborado pela autora (2019).

- ***Bag of Words***
  Uma *Bag of Words (BOW)* ou "saco de palavras"consiste na representação das palavras presentes em um texto, na qual é informada a quantidade de vezes que ela aparece. Para a representação da *Bag of Words* também podem ser utilizados valores booleanos, com o objetivo de marcar a ocorrência, ou não, destas palavras, na qual a presença é representado por 1 e a ausência representado por 0. Considerando as 3 sentenças a seguir: s1: Aline trabalha na UFBA / s2: Aline é Analista de TI na UFBA / s3: Aline é minha amiga, pode-se dizer que as *Bag of Words* destas sentenças contem as seguintes palavras: 'Aline', 'trabalha', 'na', 'UFBA', 'é', 'Analista', 'de', 'TI' , 'minha', 'amiga', Cada sentença possui a seguinte representação numérica:

'Aline trabalha na UFBA' = [1, 1, 1, 1, 0, 0, 0, 0, 0, 0]
'Aline é Analista de TI na UFBA' = [1, 0, 1, 1, 1, 1, 1, 1, 0, 0]
'Aline é minha amiga' = [1, 0, 0, 0, 1, 0, 0, 0, 1, 1]

Nas sentenças, cada número informa a ocorrência (valor 1) ou não (valor 0) de uma palavra. As palavras encontradas em uma *Bag of Words* são conhecidas como *features*. Enquanto isso, cada sentença ou texto é conhecido como *sample*. A dimensão da matriz indica o número de elementos do vocabulário.

- **Cálculo da frequência**
  A função TF-IDF calcula a frequência das palavras que mais se repetem em cada documento, levando em conta a sua frequência em todo o *corpus*. Nesse caso, a importância de uma palavra é inversamente proporcional à quantidade de vezes que ela aparece nos documentos (SALTON; BUCKLEY, 1988). Palavras que se repetem muitas vezes em todo o *corpus* não são muito significativas para caracterizar um único documento. Destaca-se que o valor TF indica a frequência de uma palavra em cada documento do *corpus* e é representado pela seguinte fórmula: **TF = quantidade de vezes que um termo aparece no texto / quantidade de termos deste documento**. Já o valor IDF calcula o peso de uma palavra rara em todos os documentos do *corpus* e isso leva a crer que quanto mais frequente for uma palavra, menor a pontuação dela. Isso a torna menos importante. Este valor é obtido pela fórmula: **IDF = log (total de documentos/documentos com o termo)**. O valor TF-IDF serve para medir o quão importante é cada uma palavra em um dado documento. O TF-IDF dá um peso maior a n-gram's mais raros, e isso não prioriza, por exemplo, as *stopwords*. O valor TF-IDF é obtido pela fórmula: **TF-IDF = TF * IDF**.

Para a realização da classificação textual é necessária a utilização de técnicas de aprendizado de máquina.

2.1.2 Aprendizado de Máquina

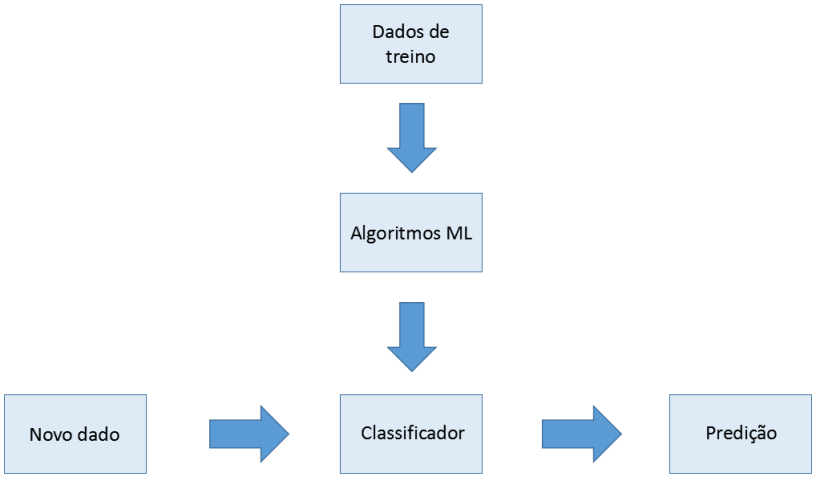
O Aprendizado de Máquina (ou *Machine Learning*) é uma área da IA (Inteligência Artificial), que permite que as máquinas aprendam através de modelos treinados e capazes de identificar padrões e tendências nos dados. A partir disso, estes resultados podem gerar uma base de conhecimento, auxiliando no processo de tomada de decisão (RUSSELL; NORVIG, 2009).

- **Tipos de Aprendizado de Máquina**
  O Aprendizado de Máquina pode ser de 3 (três) tipos: Aprendizado Supervisionado, Aprendizado Não-Supervisionado e Aprendizado por Reforço:

  - **Aprendizado Supervisionado**
    No Aprendizado Supervisionado, existe uma base de conhecimento anterior. Os dados são rotulados (classes), ou seja, os algoritmos são treinados e aprendem vários padrões para cada tipo de classe (AYODELE, 2010).

Como ilustra a Figura 2.2, os dados de treino são recebidos e utilizando-se um algoritmo de *Machine Learning*, treina-se um classificador que recebe novos dados e através do modelo aprendido com os dados anteriores, efetua-se as predições, classificando os novos dados.

Figura 2.2 Aprendizado Supervisionado



Fonte: elaborado pela autora (2019)

– **Aprendizado Não Supervisionado**
No Aprendizado Não Supervisionado, os rótulos são desconhecidos e o objetivo é estabelecer a existência de padrões para, posteriormente, supervisionar os rótulos dos agrupamentos (*clusters*) encontrados. Não existe uma base de conhecimento anterior.

– **Aprendizado por Reforço**
No Aprendizado por Reforço, não existe um conjunto de treinamento, rotulado ou não. Nesse caso, busca-se aprender qual seria a melhor ação a ser tomada, dependendo das circunstâncias nas quais a ação será executada. Ressalta-se que no Aprendizado por Reforço, trabalha-se com variáveis aleatórias e a incerteza.

Este trabalho foca no Aprendizado Supervisionado. Dentre suas técnicas (formas de se resolver uma determinada tarefa de Aprendizado de Máquina) existentes, pode-se citar a classificação, a regressão, o agrupamento *clustering* e as regras de associação.

Nessa seção é dado um destaque à classificação por ser a técnica de Aprendizado de Máquina utilizada neste trabalho.

– **Classificação** Na classificação, o conjunto de treinamento é acompanhado de rótulos (ou *labels*) e o objetivo é treinar esse conjunto a partir do atributo classe (que são nominais ou categóricas), buscando prever a classe de novos

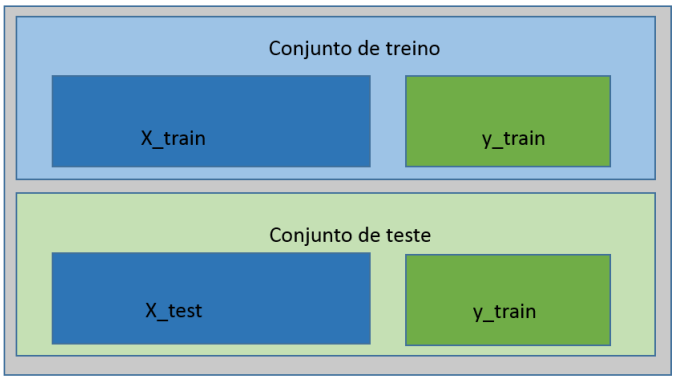
dados não rotulados. A classificação leva em consideração alguns padrões para classificar como, por exemplo, se *e-mail* é *spam* ou não, realização de uma análise de sentimentos, analisar se um cliente tende a ser bom pagador, realização de análise de crédito, entre outros. Também pode-se citar a classificação/categorização de textos, que consiste em organizá-los em categorias levando em conta seus atributos (ou *features*).

- **Regressão** A regressão é semelhante à classificação, no entanto, nesse caso, as classes são numéricas.
- **Agrupamento (*Clustering*)** Os agrupamentos não trabalham com classes. Eles criam grupos e atribuem instâncias a cada grupo, levando em conta as suas características.
- **Regras de Associação** Basicamente, tais regras buscam associar regras.

Os dados podem ser divididos nos conjuntos de treino e teste das seguintes formas:

- *Hold-Out*
  Nesse caso, divide-se o conjunto de treino e teste aleatoriamente. Sendo assim, uma instância só poderá pertencer a um dos 2 conjuntos.
  Como ilustra a figura 2.3, um *dataset* é dividido em 2 conjuntos: treino e teste, levando em conta que um dado só poderá pertencer a um destes *datasets*. Essa divisão pode ser feita nas proporções 70/30, 80/20, por exemplo. Dentro de cada *dataset*, o valor de X contem as informações das *features* e o y os *labels* do *dataset*.

Figura 2.3 Distribuição *hold-out*

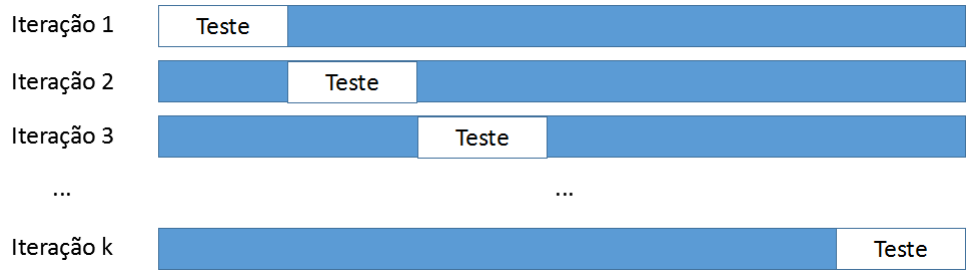


Fonte: elaborada pela autora (2019)

- *Cross-Validation* (Validação Cruzada)
  Uma mesma instância pode ser utilizada para treino e teste. O conjunto é dividido em *folds* (partições), levando em conta o número de iterações. Vale dizer que em uma das iterações, a instância fará parte do conjunto de teste.
  De acordo com a figura 2.4, o *dataset* é particionados em k *datasets* e, a cada iteração, um *dataset* é escolhido para teste e os demais são utilizados como

um conjunto de treino, de modo que todos os *datasets* sejam utilizados para treino quanto para teste quando finalizadas as iterações.

Figura 2.4 Estrutura do *k-fold*



Fonte: elaborada pela autora (2019)

O *cross-validation* tem bons resultados principalmente quando trabalha com bases de dados pequenas, já que a abordagem *hold-out* nem sempre fornece conjunto de treino e teste bons o suficiente para o modelo.

- **Modelo**
  O Aprendizado de Máquina busca predizer (prever) alvos (*target*), levando em conta um conjunto de atributos (*features*) recebidos como entrada.

  Um modelo consiste em treinar um algoritmo com um conjunto de dados, a fim de aprender com uma base já rotulada, para identificar padrões e realizar novas predições.

  Um modelo trabalha com 3 (três) sub-conjuntos de dados, denominados:

  – Treino: que é conhecida pelo método;
  – Validação: que é semi conhecida pelo método (sem querer);
  – Teste: totalmente desconhecida pelo método

  Vale destacar que um modelo deve ser capaz de executar as seguintes tarefas:

  – (1) Treinar uma base já classificada com o objetivo de aprender com ela e identificar padrões (TREINO);
  – (2) Realizar testes, avaliando sua acurácia (quão bom o modelo prever algo) (VALIDAÇÂO);
  – (3) Classificar itens, frases ou textos novos (TESTE).

  Os modelos são treinados para serem capazes de realizar predições, assim como um conjunto conhecido possui a capacidade de prever a qual classe um novo texto ou documento deve pertencer (ser classificado). O objetivo é construir um modelo que generalize para novos dados. Através das métricas, é possível avaliar o desempenho de um algoritmo quando este recebe um novo conjunto de dados.

No decorrer deste trabalho, para a realização do experimento de classificação textual, foi utilizada a técnica de classificação, na qual os modelos são treinados fazendo uso de publicações do RI rotulados com as informações de comunidade e subcomunidade, visando realizar a classificação binária e multi-hierárquica.

- **Algoritmos de Classificação**
  Um algoritmo serve para definir como uma técnica vai ser implementada. Como exemplos de algoritmos de classificação, pode-se citar o *Naive Bayes*, o *Support Vector Machine* (SVM) e as Árvores de Decisão (*Decision Trees*).

  - ***Naive Bayes***
    O algoritmo *Naive Bayes* é um classificador probabilístico simples, baseado no teorema de *Bayes*, que obedece a seguinte fórmula:

$$P(A|B) = \frac{P(B|A)xP(A)}{P(B)}$$

    Sendo que o valor A representa cada classe presente no conjunto de dados e B os atributos (*features*) considerados para a classificação.

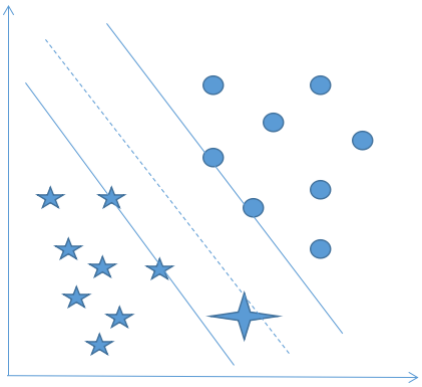
    O algoritmo *Naive Bayes* gera estimativa de probabilidade, ou seja, a estimativa de um objeto pertencer a uma mesma classe. Sendo que essa probabilidade deve ser calculada para cada classe existente no modelo e isso identificará que o dado deve pertencer à classe na qual obtiver o valor de probabilidade mais alto. O *Naive Bayes* supõe que há uma independência entre as *features* do modelo e isso significa que o classificador assume que a presença de uma determinada *feature* não tem nenhuma relação com as demais (HILDEN, 1984).

    O *Naive-Bayes* se baseia em:

    * (1) Construção do classificador;
    * (2) Aplicação da fórmula de *Bayes* para classificar novos objetos e possui duas variações:
      - *Gaussian*: trabalha com valores contínuos;
      - NB multinomial: que considera a contagem de ocorrência de *features* (características);
      - *Bernoulli*: preocupa-se com a ausência/presença de uma *feature*.

  - ***Support Vector Machine* (SVM)**
    No algoritmo SVM, dado um hiperplano com instâncias agrupadas, que representam classes distintas, caso uma nova instância seja criada, como ela deve ser classificada (CORTES; VAPNIK, 1995). O algoritmo SVM encontra um vetor que estabeleça uma fronteira entre as classes e com isso divide o hiperplano em regiões, uma para cada classe, conforme apresentado na figura 2.5:

Figura 2.5 Exemplo SVM



Fonte: Adaptado de (CORTES; VAPNIK, 1995))

Conforme ilustrado na figura, são criados vetores de suporte, que servem para criar as fronteiras que melhor separam esses dados. A partir desses vetores, o objetivo é encontrar o hiperplano (linha tracejada) que melhor separe esses dados.

O algoritmo SVM visa maximizar a margem (distância entre os vetores de suporte e o hiperplano) entre as instâncias mais próximas, criando um vetor para classificá-las.

– **Árvores de Decisão (*Decision Trees*)**
As Árvores de Decisão são uma representação dos resultados em forma de árvore que lembra um gráfico organizacional horizontal (organograma de uma organização). As Árvores de Decisão classificam as instâncias ordenando as árvores acima (ou abaixo) a partir da raiz até alguma folha (classe do *label*)(BREIMAN et al., 1984). Vale dizer que as Árvores de Decisão podem ser representados por regras IF-THEN (SE-ENTÃO) e possuem 3 (três) tipos de nós:

  * *Root node* - nó raiz ou nó de saída;
  * *Branch node* - nó filho ou nó interno, que representam as decisões;
  * *Leaf Node* - nó folha (a classe do *label*), que é quem define a classificação.

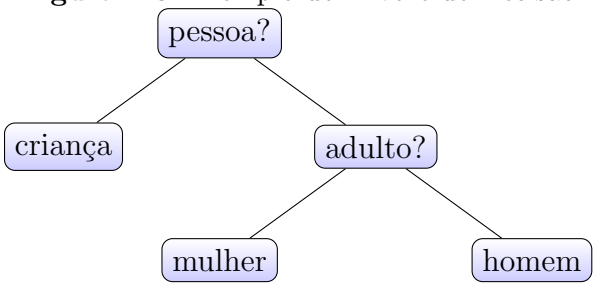
Cada nó da árvore específica o teste de algum atributo da instância e cada ramo, partindo de um nó, corresponde a um dos valores possíveis destes atributos. O número de escolhas (perguntas) define a profundidade (altura) da árvore. O algoritmo só termina de rodar quando todas as perguntas forem respondidas.

As Árvores de Decisão trabalham com os conceitos de entropia e de ganho de informação. A entropia consiste em uma métrica que informa a incerteza ou impureza de um conjunto de dados, ou seja, representa a aleatoriedade em seus valores. Esse valor pode variar entre 0 e 1, sendo que se o valor for 0 indica todos os elementos possuem a mesma classificação; enquanto que o ganho de

informação mede o quanto a entropia é removida a partir de um corte, ou seja, o quanto a pureza do conjunto de dados aumenta. O ganho de Informação (*information gain*) mede quão bem um dado atributo separa os exemplos de treinamento, levando em conta a classificação alvo.

Como ilustra a figura 2.6, uma árvore de decisão busca identificar se uma entidade se trata de uma pessoa. Em seguida, é perguntado se a entidade é uma criança ou adulto. Caso seja uma criança, o processo termina; caso se trate de um adulto, é perguntado qual o sexo dele.

Figura 2.6 Exemplo de Árvore de Decisão



Fonte: Adaptado de (BREIMAN et al., 1984).

Uma instância é classificada inicialmente pelo nó raiz, testando o atributo especificado por este nó, e em seguida, movendo-se através do ramo que corresponde ao valor do atributo de acordo com o exemplo dado. Tal processo é repetido para a sub-árvore que se origina de um novo nó.

- **Avaliação de algoritmos de *Machine Learning*** Para avaliar os modelos implementados, são utilizados diversas métricas. No decorrer deste trabalho, serão abordadas a matriz de confusão, a acurácia, a precisão, o *recall* e f1-*score*. Uma matriz de confusão é um quadro utilizado para a avaliação de um modelo de classificação. Para isso, compara os valores reais das classes com os valores preditos pelo modelo (VISA et al., 2011).

  A Tabela 2.3 a seguir ilustra um exemplo de uma matriz de confusão composta por 2 classes:

Tabela 2.3 Matriz de Confusão

		Valores das Predições	
		0	1
Valores Reais	0	Verdadeiro Positivo	Falso Negativo
	1	Falso Positivo	Verdadeiro Negativo

Fonte: elaborado pela autora (2019).

Na referida matriz acima apresentada, são mostrados os seguintes valores:

- Verdadeiro Positivo ou *True positive* (TP): indica uma classificação correta da classe positiva. Exemplificando, a classe real é Positivo e o modelo, portanto, classificou-a como Positivo;
- Verdadeiro Negativo ou *True negative* (TN): indica uma classificação correta da classe negativa. Como exemplo, a classe real é Negativo e o modelo classificado como Negativo;
- Falso Positivo ou *False positive* (FP): indica uma classificação errada da classe positiva. Por exemplo, a classe real é Negativo e o modelo classificado como Positivo;
- Falso Negativo ou *False negative* (FN): indica uma classificação errada da classe negativa. Exemplificando, a classe real é Positivo e o modelo classificou-a como Negativo.

Na tabela 2.4 é ilustrada um exemplo de classificação para avaliar o risco de crédito:

Tabela 2.4 Exemplo de classificação Matriz de Confusão

Classe real	Classe predita	Tipo
Bom pagador	Bom pagador	VP
Bom pagador	Mau pagador	FN
Mau pagador	Bom pagador	FP
Mau pagador	Mau pagador	VN

Fonte: elaborado pela autora (2019)

Neste exemplo, se o modelo classificar uma instância que possui a classe "Bom pagador"como "Bom pagador", gerará um valor "Verdadeiro Positivo". Porém, quando classificar uma instância com a mesma classe como "Mau pagador", gerará um valor "Falso Negativo". Por outro lado, caso uma instância possua a classe "Mau pagador"e for classificada como "Bom pagador", trata-se de um "Falso (Positivo", enquanto que se ela tiver a mesma classe e for classificada como "Mau pagador", será um "Verdadeiro Negativo".

Os Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN) indicam as taxas de acerto e os Falso Negativos (FN) e Falso Positivos (FP) indicam os erros do modelo.

Caso o modelo possua "N"classes, a matriz de confusão apresentará o formato N x N, ou seja, se o modelo possuir 4 classes, a matriz terá o formato 4 x 4.

O modelo divide os dados pré-rotulados em conjunto de treino e teste. A classe dos dados é ocultada no conjunto de teste, as métricas vão utilizar as predições realizadas para a comparar a classe real de cada instância com a classe predita. Caso o modelo seja bem avaliado, pode ser colocado em produção.

Para avaliar o desempenho de um modelo podem ser utilizadas várias métricas, dentre as quais se destacam a acurácia, a precisão, a revocação *recall* e o F1-*score*, que são calculadas a partir dos valores apresentados na matriz de confusão.

A **acurácia** indica a capacidade do modelo prever algo e mede a quantidade de acertos sobre um todo. Para calcular a acurácia, utiliza-se a seguinte formula:

$$Acurácia = \frac{TP + TN}{TP + FP + TN + FN}$$

Com base nisso, pode-se dizer qual o percentual das instâncias classificadas corretamente. É importante que a acurácia seja calculada em um conjunto com uma boa variedade de resultados para o que modelo não tenha *overfitting*, que é quando um modelo se ajusta bem para um conjunto de dados conhecido, mas não alcança bons resultados em um novo conjunto de dados.

A **precisão** indica quantos dos itens classificados como positivos realmente são e é calculada da seguinte forma:

$$Precisão = \frac{TP}{TP + FP}$$

Isso significa o número de vezes que uma classe foi predita corretamente e dividida pelo número de vezes que a classe foi predita.

Já a **revocação (recall)** indica a relação entre os resultados positivas que foram preditos corretamente e todas as previsões que realmente são positivas, ou seja, os Verdadeiros Positivos e os Falso Negativos) e é calculada da seguinte forma:

$$Recall = \frac{TP}{TP + FN}$$

Isso significa o número de vezes que uma classe foi predita corretamente (TP) dividido pelo número de vezes que a classe aparece no dado de teste (TP+FN). O *recall* indica o quanto o modelo está indicando os casos positivos corretamente.

O **F1-score** é calculada da seguinte forma:

$$f1 - score = \frac{2 * (precision * recall)}{precision + recall}$$

Esta medida é a média harmônica[1] entre a precisão e a revocação. A partir dessa informação, pode-se dizer qual a *performance* do classificador através apenas de um

[1]representada pela quantidade valores dividido pela soma dos inversos de cada valor

indicador. De um modo geral, quanto maior for o valor do f1-*score* (próximo de 1) melhor é o modelo.

2.2 EXTRAÇÃO DE PALAVRAS-CHAVE

As palavras-chave são compostas por uma ou mais palavras (*keyphrase*) e ajudam a descrever um documento. Devem ser bem definidas para possibilitar que este documento seja facilmente recuperado na busca (retornando resultados relevantes) e para enriquecer semanticamente seus metadados (EDMUNDSON, 1969).

A tarefa de extração de palavras-chave serve para identificar automaticamente os termos capazes de descrever um documento. Para tanto, são identificados os termos mais representativos de cada documento, montando-se um *ranking* com o *score* de cada termo e atribuindo um peso - o que define a importância deste termo no documento, por fim, são listados os k termos mais frequentes (SIDDIQI; SHARAN, 2015), como ilustra a figura 2.7:

Figura 2.7 Fluxo da Extração de Palavras-Chave

Fonte: elaborada pela autora[2]

.

Como métodos de Extração de palavras-chave, pode-se citar:

- **YAKE** Método não supervisionado, que independe de *corpus* e de idioma. Trabalha com pré-processamento do texto com a finalidade de descobrir termos candidatos e detecta palavras-chave relevantes com base em análises estatísticas extraídas dos documentos. Retorna uma lista de palavras-chave potencialmente relevantes juntamente com seu grau de relevância(CAMPOS et al., 2018)

- **TF-IDF (*Term Frequency - Inverse Document Frequency*)** Calcula o valor TF-IDF (frequência da palavra em cada documento, comparado com sua frequência em todo o *corpus*) (SALTON; BUCKLEY, 1988). As palavras com maiores valores de TF-IDF são escolhidas como palavras-chave.

- **RAKE (*Rapid Automatic Keyword Extraction*)** Determina as palavras-chave que devem ser retornadas com base na frequência das palavras e sua co-ocorrência com outras palavras do texto (ROSE et al., 2010).

- ***n-gram*** Nesse modelo, cada sequência de n termos possui uma probabilidade de ocorrer no texto. Com base nisso, torna-se possível prever o próximo termo com

[2]Ícones obtidos gratuitamente no site https://www.flaticon.com/

base em termos utilizados anteriormente. O valor de n indica a quantidade de termos que devem ser considerados (BROWN et al., 1992).

- ***most_common*** Função da biblioteca *Natural Language Toolkit* (NLTK)[3](LOPER; BIRD, 2002), implementada em *Python* e que identifica os termos mais frequentes em um documento.

- ***Gensim*** Trabalha com sumarização de texto e identificação de palavras-chave; Utiliza o algoritmo *Text rank*(MIHALCEA; TARAU, 2004) e aprendizado não supervisionado.

Para avaliar as métricas de extração de palavras-chave, pode ser utilizado um conjunto de texto e realizar a extração das palavras-chave de cada texto. Em seguida, deve-se efetuar a verificação manual das sugestões para refinar as sugestões retornadas.

Alguns métodos calculam os termos mais frequentes, enquanto outros atribuem um peso de acordo com a relevância identificada de cada palavra. De posse das palavras-chave refinadas retornadas por cada método, pode-se realizar uma análise/avaliação de quais métodos retornaram os melhores resultados.

2.3 ANOTAÇÃO SEMÂNTICA

A Web Semântica, uma extensão da Web atual, permite adicionar significado ao conteúdo disponível. Com base nisso, tanto as pessoas quanto as máquinas tornam-se capazes de interpretá-lo (BERNERS-LEE; HENDLER; LASSILA, 2001).

Um metadado serve para descrever melhor um documento, facilitando a sua recuperação através de mecanismos de busca. Dependendo do tipo de documento, torna-se possível definir um conjunto de metadados para descrevê-lo.

As Anotações Semânticas favorecem que os metadados de um documento sejam enriquecidos semanticamente, facilitando a sua recuperação, que pode ser realizada de 3 (três) formas (OREN et al., 2006):

- manual: quando a anotação é realizada manualmente;

- semiautomática: quando os termos são extraídos de forma automática e depois são sugeridos a um especialista para serem anotados;

- automática: quando os termos são extraídos e anotados de forma automática.

Destaca-se que a anotação manual é um processo lento, custoso e passível de erros. Em contrapartida, a automatização da anotação permite que o processo seja realizado de forma rápida, destacando os termos relevantes e favorecendo o enriquecimento semântico dos metadados.

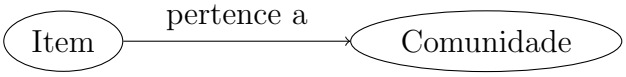
As Anotações Semânticas podem ser tanto intrusivas (quando são feitas no próprio documento) ou não-intrusivas (quando são feitas em um arquivo à parte). Para representar uma anotação semântica, pode-se utilizar o padrão *Resource Description Framework* (RDF), bem como o padrão de metadados *Dublin Core*.

[3]http://www.nltk.org/

2.3.1 Padrão Dublin Core

A *Web* Semântica pode expressar a adição de significado através do padrão RDF, que consiste em um conjunto de triplas no formato **sujeito - predicado - objeto** (BERNERS-LEE; HENDLER; LASSILA, 2001)(HEATH; BIZER, 2011), conforme o grafo ilustrado na Figura 2.8 abaixo:

Figura 2.8 Grafo de uma tripla RDF



Fonte: Adaptado de (HEATH; BIZER, 2011).

Desse modo, é possível fazer uma afirmação do tipo **Item - pertence a - Comunidade**Tanto o **sujeito**, quanto o **predicado** e o **objeto** são identificados por meio de uma *Uniform Resource Identifier* (URI), que serve para denominar o recurso.

Na Web de Dados, tudo é visto como um recurso (pessoas, coisas, páginas *web*, etc). Assim, o RDF descreve recursos, no formato sujeito - predicado - objeto (recurso - propriedade - valor) através de triplas ou declarações RDF (RDF *statement*) (HEATH; BIZER, 2011).

O RDF é considerado pela W3C[4] como um padrão para a descrição de recurso na *Web*. Sendo assim, é possível gerar triplas que associem um documento a seus metadados, de modo que essa adição de semântica possa contribuir com a recuperação desse documento.

O *Dublin Core* (DC) ou *Dublin Core Element Set* (DCMES) é um padrão utilizado para armazenar metadados (informações sobre os próprios dados) e é composto por 15 (quinze) elementos [5] (HILLMANN, 2008), conforme evidenciado na tabela 2.5 a seguir:

[4] *World Wide Web Consortium* (consórcio responsável por estabelecer padrões na *Web*). Site oficial: https://www.w3.org

[5] O conjunto de todos os 15 termos de metadados recebe o nome de DC *Terms*.

Tabela 2.5 Esquema *Dublin Core.*

Elemento Descritivo	**Descrição do Valor**
Title	Nome dado ao recurso
Creator	Entidade responsável por fazer o conteúdo do recurso.
Subject	Tema do conteúdo do recurso.
Description	Relato do conteúdo do recurso.
Publisher	Entidade responsável por tornar o recurso disponível.
Contributor	Entidade responsável por qualquer contribuição para o conteúdo do recurso.
Date	Data associada a um evento no ciclo de vida do recurso.
Type	Natureza ou gênero do conteúdo do recurso.
Format	Expressão física ou digital do recurso.
Identifier	Referência ambígua ao recurso dentro de um determinado contexto.
Source	Referência a um recurso do qual o presente recurso é derivado.
Language	Língua do conteúdo intelectual do recurso.
Relation	Referência a um recurso relacionado.
Coverage	Extensão ou escopo do conteúdo do recurso.
Rights	Informações sobre os direitos detidos em e sobre o recurso.

Fonte: (HILLMANN, 2008), traduzido.

Os metadados servem para identificar e descrever um determinado documento, tornando mais fácil sua recuperação. Os elementos do *Dublin Core* podem ser estendidos pelo *dcterms* [6]. Como pode ser observado na tabela 2.6, os metadados podem ser representados até dois níveis e, no caso do *Dublin Core*, tem-se o formato DC.Elemento.Qualificador.

Tabela 2.6 Exemplo de elementos *dcterms*

Esquema	Elemento	Qualificador
dc	creator	artist
dc	creator	organization
dc	creator	illustrator

Fonte: Adaptado de (HILLMANN, 2008)

Como exemplo de elemento qualificador, tem-se dc.creator destinado à criação de conteúdos e sem o segundo nível - DC.creator.organization, voltado para a organização criadora com o segundo nível.

O código 2.1 - apresentado abaixo - descreve o RDF dos metadados de uma dissertação de mestrado defendida no ano de 2017, em formato pdf e depositada no RI UFBA:

```
01 | <?xml version="1.0"?>
02 | <rdf:RDF
03 | xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
04 | xmlns:dc= "http://purl.org/dc/elements/1.1/"
05 | xmlns:dcterms="http://purl.org/dc/terms/">
```

[6]Disponível em https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

```
06 | <rdf:Description rdf:about="https://repositorio.ufba.br/ri/handle
       /ri/21758">
07 |   <dc:title>PLATAFORMA COMPUTACIONAL WEB PARA CALIBRAÇÃO DE
       SISTEMAS DE MEDIÇÃO</dc:title>
08 |   <dc:subject> plataforma computacional web, calibração,
       incerteza de medição, NBR ISO IEC/17025.</dc:subject>
09 |   <dc:date>2017-05-27</dc:date>
10 |   <dc:type>Dissertação de Mestrado</dc:type>
11 |   <dc:format>application/pdf</dc:format>
12 |   <dc:language>pt-BR</dc:language>
13 | </rdf:Description>
14 | </rdf:RDF>
```

Código 2.1 RDF do Dublin Core

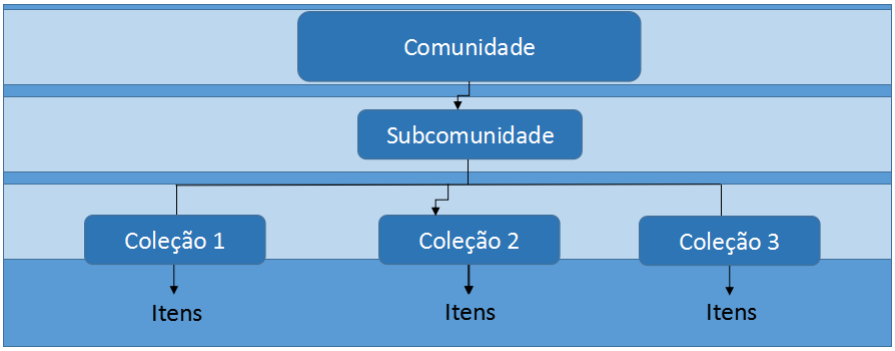
2.4 REPOSITÓRIOS INSTITUCIONAIS

Um RI serve para armazenar as produções científicas e acadêmicas geradas por uma instituição, possibilitando que seus pesquisadores possam divulgar suas pesquisas (GOMES; ROSA, 2010), aumentando a visibilidade das mesmas (LEITE et al., 2012). É necessário que seja realizada uma divulgação dos resultados destas pesquisas, como retorno aos investimentos destinados à ciência (SAYÃO et al., 2009).

Atualmente, existem diversos *softwares* que são utilizados como solução de armazenamento para os RI's, sendo que o *DSpace*[7] é o mais utilizado e foi desenvolvido através de uma parceria firmada entre o *Massachusetts Institute of Technology* (MIT) e a *Hewlett-Packard* (HP) (TANSLEY et al., 2003).

O *DSpace* é organizado em comunidades e subcomunidades, seguindo a estrutura organizacional das instituições. Cada comunidade pode possuir inúmeras coleções que são compostas por itens e permitem armazenar arquivos em diversos formatos, conforme figura 2.9 abaixo:

Figura 2.9 Hierarquia *Dspace*



Fonte: elaborado pela autora (2019)

O *DSpace* permite o autoarquivamento, no qual o próprio pesquisador pode realizar o depósito de sua publicação. Antes que a publicação fique disponível para o usuário final,

[7]http://www.dspace.org

deve passar pela validação dos metadados informados (ROSA; MEIRELLES; PALACIOS, 2011).

Segundo (FARID; KHAN; JAVED, 2013), os RI's frequentemente são desenvolvidos para servir a uma comunidade de usuários de uma dada instituição e possibilita o aumento da visibilidade de suas produções cientificas e acadêmicas.

O *DSpace* possibilita o depósito de arquivos, dando suporte a diversos formatos, dentre os quais se destacam as imagens, os vídeos e os sons (CUEVAS CERVERÓ, 2008) apud (ROSA; MEIRELLES; PALACIOS, 2011).

Considerando que este trabalho tem como foco a anotação semântica, é interessante observar que o enriquecimento semântico dos metadados facilita a interoperabilidade com outros repositórios através do protocolo *Open Archives Initiative – Protocol for Metadata Harvesting* (OAI-PMH). Com isso, as publicações de um RI são disseminadas de maneira eficiente entre os repositórios digitais.

Para estimular o povoamento de um RI, faz-se necessário haver uma política para institucionalizar os depósitos das produções, promovendo ações de incentivo e divulgação das vantagens de se depositar as produções cientificas em um repositório institucional (LEITE et al., 2012).

O *DSpace* serve para armazenar as publicações cientificas de uma instituição. Para descrever as informações de cada documento, é utilizado o padrão de metadados *Dublin Core*. Esses metadados servem para descrever melhor um documento e para auxiliar o compartilhamento das publicações com outros repositórios institucionais.

No próximo capítulo, serão apresentados alguns trabalhos que utilizam as tecnologias abordadas na fundamentação teórica, com destaque para os trabalhos que envolvem a anotação semântica em repositórios institucionais.

Capítulo 3

Neste capitulo Serão descritos alguns trabalhos relacionados aos temas desta dissertação

TRABALHOS RELACIONADOS

Como este trabalho é de natureza multidisciplinar, esse capítulo está dividido em seções de acordo com esses temas.

Nesse capítulo serão apresentados alguns trabalhos identificados na literatura, abordando os temas relacionados a esse trabalho. Cada seção engloba os trabalhos encontrados, começando por trabalhos relacionados a Repositórios Institucionais. Em seguida, são descritos trabalhos que abordam os temas extração de palavras-chave e anotações semânticas. Por fim, são descritos trabalhos que realizam anotação semântica de metadados em repositórios institucionais que utilizam a ferramenta *DSpace* como solução de armazenamento. No final do capítulo, é apresentado o diferencial da solução proposta com outros projetos de anotação semântica em RI.

3.1 REPOSITÓRIOS INSTITUCIONAIS

O trabalho realizado por (CARVALHO, 2018), teve como propósito analisar o uso e a aceitação do RI UFBA pela sua comunidade. O trabalho foi dividido em 4 partes e envolveu desde pesquisa sobre a implantação e as politicas de uso de outros Repositórios Institucionais, passando pela aplicação de um questionário sobre a aceitação e o utilização do RI UFBA, bem como e aplicação da técnica *Think Aloud* para a realização de uma análise da usabilidade do sistema. A terceira etapa consistiu no ajuste de modelos de regressão logística, a fim de verificar algumas variáveis coletadas. Para finalizar, na quarta etapa foi realizada uma interpretação dos resultados obtidos, identificando os fatores positivos para a aceitação e uso do RI-UFBA, como também sua usabilidade e o entendimento da importância de se utilizar as tecnologias abertas. Dentre os fatores apontados pela baixa aceitação, a falta de divulgação do RI por parte da instituição faz com que muitas pessoas desconheçam a sua existência, a sua finalidade e os seus benefícios. Dificuldades na busca e uma *interface* pouco intuitiva também foram apontadas como problema de uso da ferramenta.

(SANTOS, 2019) desenvolveu um trabalho que ressalta o povoamento do RI UFBA, com uma análise das comunidades ligadas à área I. Dada à relevância da divulgação das

produções cientificas das IFES, iniciou-se um estudo da verificação do povoamento das publicações depositadas no RI UFBA, no período 2010 a 2018, no qual a maioria das publicações depositadas no RI UFBA foram teses e dissertações. Isso envolveu uma análise das subcomunidades de Programas de Pós-Graduação, bem como a identificação dos tipos de documentos depositados nelas. Foi realizada uma coleta de dados com os coordenadores destes Programas com a finalidade de identificar quais critérios cada programa utiliza para inserir suas publicações no repositório. O resultado do trabalho aponta para a necessidade de uma institucionalização capaz de alavancar o uso e a disponibilização da produção cientifica da universidade. Mesmo assim, representando a grande maioria das publicações depositadas, não existe da parte dos Programas de Pós-Graduação um critério capaz de estimular os depósitos de publicações, principalmente no que se refere às comunicações orais - trabalhos apresentados em congressos. Foi ressaltada a importância de publicizar o que está sendo produzido na universidade como forma da transparência das informações acadêmicas e científicas para a sociedade, divulgando à comunidade UFBA a finalidade do repositório.

3.2 EXTRAÇÃO DE PALAVRAS-CHAVE E ANOTAÇÕES SEMÂNTICAS

O artigo de (MENDONçA et al., 2012) aborda a extração automática de informações na área médica da hemoterapia utilizando técnicas de PLN. Adotou-se para a extração de termos candidatos a ferramenta *Scketch Engine*, a construção de um *corpus* no domínio de sangue com a anotação *Part-of-Speech* (POS), *Tagger* (na qual cada termo é etiquetado de acordo com seu uso na frase) e o cálculo das frequências dos termos candidatos à ontologia. Nesse caso, considerou-se somente os prefixos dos termos, pois, geralmente estes representam sua semântica. Após este cálculo, os termos foram agrupados de acordo com a classe semântica correspondente; em seguida, foi gerada a taxonomia destas classes semânticas e a validação da ontologia por um especialista da área.

O trabalho de (DIAS; MALHEIROS, 2005) ressaltou que as palavras-chave servem para filtrar o conteúdo. Foi implementada uma nova versão do algoritmo KEA para a extração das palavras-chave em Língua Portuguesa. Uma nova lista de *stopwords* também foi gerada. Para a avaliação do algoritmo, foi realizado um experimento com dissertações e teses de diversas áreas, comparando documentos inalterados e documentos com as palavras-chave extraídas, obtendo-se um resultado compatível com aqueles obtidos na aplicação original com a Língua Inglesa.

O trabalho de (SANCHES, 2018) trata sobre as anotações semânticas em objetos de aprendizagem pois, para sua reutilização, um objeto de aprendizagem deve ser legível às máquinas. Como a anotação manual é inviável, por ser um processo lento e custoso, foi proposto um modelo para anotar os metadados destes objetos de forma automática.

O trabalho de (SILVA, 2016), propõe a anotação semântica automática dos currículos disponíveis na plataforma *Lattes*, que consiste em uma base de dados de currículos de pesquisadores de diversas instituições de pesquisa. Sua implementação envolve a extração de entidades (pessoa, empresa, instituição, pais, etc), utilizando a ferramenta *TextRazor*. Em seguida, os dados obtidos são anotados em uma estrutura RDF *attributes* (RDFa), embutida em um arquivo *Hypertext Markup Language* (HTML). Quando é identificado

que a entidade é do tipo pessoa, ela é mapeada com os vocabulários FOAF ("Friend of a Friend") e SCHEMA:PERSON. As triplas geradas disponibilizam as informações em formato RDF *Turtle*. Para a anotação, são utilizados dados de bases abertas como a *Linked Open Data* (LOD). O sistema implementado disponibiliza um componente de armazenamento e consulta de dados, fato que possibilita utilizar buscadores semânticos nestes currículos.

O trabalho desenvolvido por (MANZATO; GOULARTE, 2012), apresenta uma proposta de uma técnica automática de anotação semântica para conteúdos multimídia. As limitações das técnicas automáticas apresentam uma alta dependência de inferências de bases de conhecimento; as abordagens manuais consumem tempo e apresentam dados incompletos. Diversos serviços possibilitam que os usuários possam adicionar *tags* para anotar o conteúdo produzido. Com isso, foi proposto um anotador no qual é recebido um conjunto de *tags* que são filtradas gerando uma folksonomia (indexação de informação) de *tags*. Estas foram enriquecidas para o caso de uma posterior anotação, seguindo 3 estratégias apresentadas no artigo. Por fim, o conteúdo foi anotado com os conceitos obtidos através destas estratégias.

3.3 ANOTAÇÕES EM METADADOS DE REPOSITÓRIOS INSTITUCIONAIS

No trabalho desenvolvido por (KOUTSOMITROPOULOS; SOLOMOU; KALOU, 2015), é descrito o repositório institucional da Universidade de Patras, na Grécia, que foi implementado no *Dspace*, cujos metadados anotados semanticamente foram publicados no *Linked Data*. Para isso, foi criada uma ontologia para descrever a estrutura do repositório *DSpace* e também para o padrão de metadados *Dublin Core*, no qual foram mapeados alguns de seus elementos e propriedades. Como o *DSpace* permite que seu esquema de metadados seja estendido, este trabalho utilizou o esquema *Learning Object Metadata* (LOM). Os metadados do *Dublin Core* e LOM foram mapeados para *DCTerms* e, em seguida, foram mapeados em triplas RDF. Também foi implementado um sistema de busca semântica, com interface para o usuário buscar por relações mapeadas pela ontologia. Ao listar os resultados retornados, é possível selecionar e visualizar seus metadados, adicionando as referências da *DBpedia* a cada publicação, pela propriedade *foaf:page*.

O trabalho de (KONSTANTINOU et al., 2014), descreve como um repositório no *Dspace* pode ser anotado semanticamente, como um repositório semântico. Baseado no modelo de dados do *DSpace*, foi feito um mapeamento dos metadados dos conteúdos deste repositório (mapeamento R2RML) armazenados em banco de dados, armazenados em banco de dados, gerando um grafo RDF.

(FARID; KHAN; JAVED, 2013) explica a necessidade de se publicar os dados de um repositório na *Web* Semântica com a finalidade de compartilhar as informações e integrar as informações multidisciplinares. O modelo consiste na extração dos metadados, relacioná-los a uma ontologia e por fim, povoar a ontologia com as informações do repositório.

Na Tabela 3.1, são apresentados 5 (cinco) trabalhos sobre anotação semântica, sendo que 3 (três) deles realizam a anotação semântica na ferramenta *Dspace*, um em objetos de aprendizagem e outro na base de dados do Currículo *Lattes*. As anotações realizadas nos

repositórios *Dspace* utilizam o metadado padrão dele, que é o *Dublin Core*, enquanto que os outros trabalhos utilizam o LOM, especifico da área de objetos de aprendizagem e o vocabulário FOAF, que armazena informações de pessoas. Por fim, nenhum dos trabalhos realiza a anotação associando os itens anotados a mais de uma categoria.

Tabela 3.1 Tabela Comparativa de Trabalhos Relacionados

Trabalho	**Objetivo**	**Origem dos Dados**	**Tipos de Metadados**	**É multi-domínio?**
Koutsomitropoulos, Solomou e Kalou (2015)	Anotação semântica em repositório *DSpace*	Repositório no *DSpace*	*Dublin Core*	Não
Konstantinou, Spanos e Mitrou (2013)	Anotação semântica em repositório *DSpace*	Repositório no *DSpace*	*Dublin Core*	Não
Sanches (2018)	Anotação semântica em objetos de aprendizagem	Objetos de aprendizagem	Padrão LOM	Não
Silva (2016)	Anotação semântica do Currículo *Lattes*	Plataforma *Lattes*	Vocabulário FOAF	Não
Farid, Khan e Javed (2013)	Anotação semântica em repositório *DSpace*	Repositório no *DSpace*	*Dublin Core*	Não

Tabela 3.2 Fonte: autoria própria (2019).

A característica de multi-domínio contempla a anotação de itens associados a mais de uma área de conhecimento com a indicação de itens do repositório que podem ser considerados trabalhos multidisciplinares.

3.4 DIFERENCIAL DESTE TRABALHO

Tendo em vista os trabalhos que já foram publicados e que abordam a temática de repositórios institucionais, verifica-se a relevância de buscar formas capazes de estimular o uso e a publicização das produções cientificas das instituições. Para tanto, é necessário oferecer a infraestrutura adequada para que tais soluções funcionem de maneira funcional, apresentando resultados relevantes no sistema de busca, além de contribuir para que os pesquisadores e a comunidade em geral saibam da sua finalidade e tenham sempre o hábito de utilizá-lo em suas pesquisas.

Como diferencial deste trabalho, pode-se citar a identificação de trabalhos multi-domínio que são indicadores de trabalhos multidisciplinares através da classificação textual binária e a implementação de um classificador multi-hierárquico para classificar as publicações ainda não classificadas nas comunidades e subcomunidades, de acordo com os termos mais relevantes. Durante os experimentos, foram utilizadas tanto informações armazenadas na base de dados quanto aquelas extraídas dos textos das publicações.

Além disso, foi realizado um estudo exploratório das técnicas de extração de palavras-chave, com a sugestões de termos relevantes nas publicações do RI UFBA e técnicas de *machine learning* para classificação textual e para identificar e sugerir publicações multidisciplinares. Por fim, a anotação semântica semi-automática recebe as saídas da classificação textual e extração de palavras-chave validadas no estudo de caso realizado com as bibliotecárias do SIBI.

No próximo capítulo, será apresentada a arquitetura da solução implementada para realizar a anotação semântica dos itens de um repositório acadêmico implementado no *DSpace*.

Capítulo

4

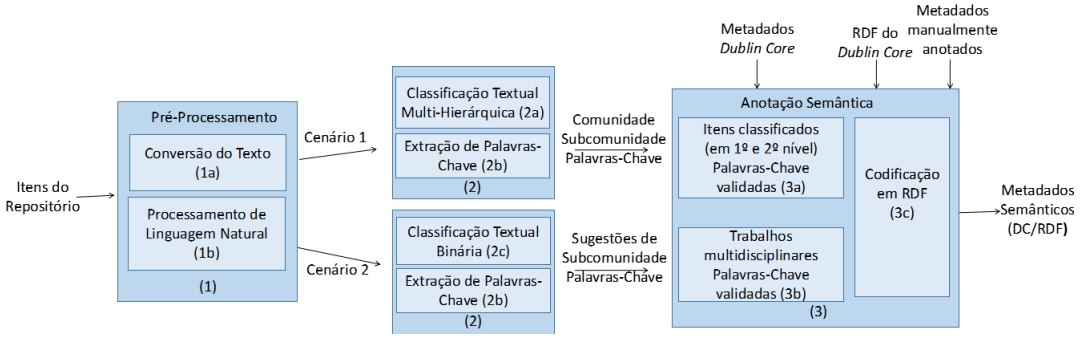
Este capítulo descreve a solução implementada.

SOLUÇÃO IMPLEMENTADA

Neste capítulo será apresentada a arquitetura da solução implementada no decorrer desta dissertação, descrevendo a etapa de pré-processamento e os experimentos de classificação textual e extração de palavras-chave.

A Figura 4.1 ilustra a arquitetura da anotação semântica dos itens do repositório, sendo que cada etapa é descrita em uma seção deste capítulo. O foco desta solução é a anotação de itens textuais, porém, esta pode ser utilizada com outros formatos de arquivos.

Figura 4.1 Arquitetura da Solução



Fonte: elaborado pela autora (2019).

Esta solução trata-se de um trabalho experimental que pode ser implementado futuramente no RI UFBA. Para isso seria necessário a implementação de *plugins* para que o *DSpace* possa incorporar as implementações realizadas para a classificação textual (multi-hierárquica e binária) e a extração de palavras-chave.

Antes disso, é importante realizar uma análise dos requisitos necessários para a implementação desses *plugins.*

Para iniciar o processo, é necessário realizar o pré-processamento, no qual os documentos utilizados durante os experimentos passarão pelo tratamento de texto.

4.1 PRÉ-PROCESSAMENTO

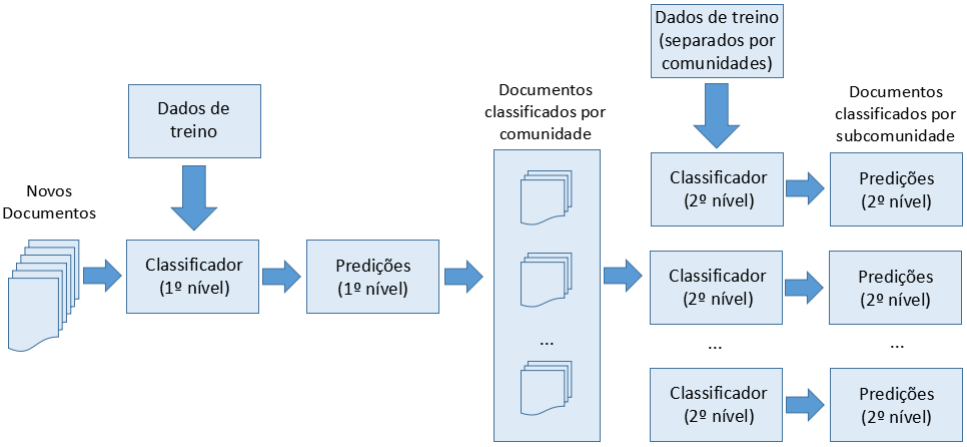
A etapa de pré-processamento (1) objetiva tratar o conteúdo dos itens do repositório, que neste projeto se encontram no formato de texto pdf e que serão convertidos para o formato txt (1a). Em seguida, serão aplicadas as técnicas de PLN (1b). A linguagem utilizada para a implementação dos códigos foi o *Python*[1]. por esta apresentar uma vasta documentação, boa curva de aprendizado e uma biblioteca específica para tarefas de PLN, a biblioteca *Natural Language Toolkit* (NLTK)[2]. O NLTK conta com uma série de comandos para executar as tarefas do PLN, como por exemplo, a tokenização de sentenças, remoção de *stopwords*, etc.

As tarefas de PLN que serão executadas nestes textos envolvem a correção da acentuação, a conversão do texto em letras minúsculas, a eliminação de *stopwords*, a eliminação de de caracteres numéricos e a eliminação de caracteres de pontuação.

4.2 CLASSIFICAÇÃO TEXTUAL

A etapa de classificação textual envolve a realização da classificação textual (multi-hierárquica (2a) e binária (2b)). Esta classificação recebe como entrada os textos gerados na etapa de pré-processamento. A Figura 4.2, apresenta a estrutura de um classificador multi-hierárquico:

Figura 4.2 Classificador Multi-Hierárquico.



Fonte: elaborado pela autora (2019).

O processo inicia-se com a implementação de um classificador para classificar os documentos em 1º nível (informação da comunidade). Finalizadas as predições, com os documentos classificados por comunidades, são implementados classificadores em 2º nível (informação de subcomunidade), sendo um classificador para cada comunidade do conjunto de documentos. O classificador de cada comunidade, deve receber os documentos

[1]https://www.python.org/
[2]http://nltk.org

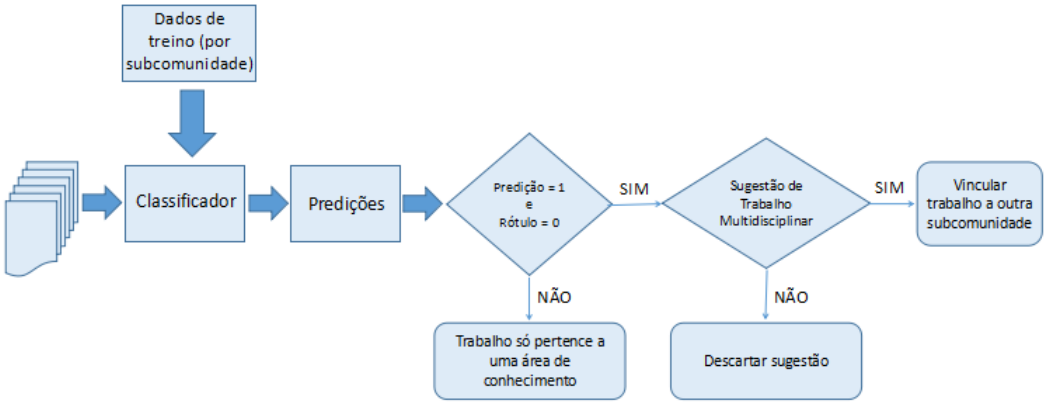
classificados como pertencentes a essa comunidade, a fim de identificar a qual subcomunidade ele pertence. Por fim, são realizadas as predições em 2º nível.

Um classificador multi-hierárquico serviria para classificar publicações que ainda não foram depositadas no RI e que não estejam organizadas por subcomunidade, mas sim, por outro critério, como a data de entrega ao colegiado ou data de defesa, por exemplo.

Dessa forma, com base nos termos representativos de cada comunidade e subcomunidade, torna-se possível treinar um classificador multi-hierárquico para identificar a qual comunidade (1º nível) e subcomunidade (2º nível) cada publicação deve pertencer.

A Figura 4.3 ilustra a estrutura de um classificador binário:

Figura 4.3 Classificador Binário.



Fonte: elaborado pela autora (2019).

O classificador binário recebe documentos classificados por subcomunidade e preenche uma matriz binária, conforme descrito no Apêndice B. Em seguida, realiza as predições e verifica se o valor da predição for 1, sendo que o valor real é 0, a predição é identificada como sugestão de trabalho multidisciplinar. Caso contrário, o trabalho pertence a uma única área de conhecimento. Essa sugestão é avaliada com a finalidade de verificar se existe pertinência ou não. Caso o trabalho seja identificado como multidisciplinar, ele é vinculado também à subcomunidade sugerida. Caso contrário, a sugestão é descartada.

Um classificador binário serviria para identificar e sugerir publicações multidisciplinares, em casos nos quais as predições é indicada como 1 (pertence a subcomunidade) quando na verdade o documento não pertence à aquela subcomunidade (identificada como 0

4.2.1 Cenário Classificação Multi-Hierárquica

Em um RI, o povoamento das comunidades nem sempre ocorre de maneira homogênea, pois, geralmente ele fica a critério de cada curso, que decide se será utilizado o repositório como solução de armazenamento ou não. Com isso, pode ocorrer de algumas comunidades serem bem povoadas e outras não.

Caso haja uma demanda institucional de estímulo à utilização deste RI, como forma de publicizar a produção científica da universidade e se tenha as publicações que serão

depositadas, organizadas por um outro critério, como data da defesa ou data de entrega no colegiado e não por subcomunidade ou coleção, pode-se, baseado nos termos representativos de cada comunidade e subcomunidade, treinar um classificador multi-hierárquico para identificar a qual comunidade e subcomunidade cada trabalho deve pertencer.

Considerando o cenário do trabalho desenvolvido por (SANTOS, 2019), no qual foi relatado que para realizar o povoamento de algumas subcomunidades do RI UFBA, foi necessário realizar mutirões; um classificador multi-hierárquico poderia auxiliar na classificação dessas publicações em suas respectivas subcomunidades.

Para a avaliação da classificação multi-hierárquica, serão avaliadas as métricas acurácia, precisão, *recall* e f1-score.

4.2.2 Cenário Classificação Binária

Levando-se em consideração que um RI é multidisciplinar - quando uma publicação pode pertencer a mais de uma área de conhecimento - pode ser interessante que esta publicação seja associada às subcomunidades relacionadas ao seu tema. Um classificador binário (2a) permite identificar em suas predições que a publicação pode também pertencer a outra área de interesse, bem como sugerir que ela também seja associada à uma subcomunidade correspondente.

As publicações em um RI envolvem desde artigos e capítulos de livros até Trabalhos de Conclusão de Curso (TCC) de graduação, dissertações e teses). Muitos desses trabalhos são multidisciplinares, por tratarem de temas pertencentes a mais de uma área de conhecimento. Pode ser interessante, que caso seja identificada uma situação desta, que este trabalho também seja associado a uma outra subcomunidade.

A classificação binária permite identificar sugestões de publicações quando na predição aparece que a publicação pertence a uma determinada subcomunidade e, na verdade, ela pertence a outra. Sendo assim, o classificador indica as situações em que a classificação retornou uma predição indicada como 1 (pertence a subcomunidade) quando a publicação não pertencer na verdade àquela subcomunidade (identificada como 0).

Para a implementação destes classificadores serão utilizados os algoritmos *Naive Bayes*, SVM e de Árvores de Decisão, por serem algoritmos de classificação, que trabalham com aprendizado supervisionado e com dados rotulados. A implementação será realizada com o auxilio da biblioteca *scikit-learn*[3], implementada em *Python*.

Para a avaliação da classificação binária, as métricas acurácia, precisão, *recall* e f1-score serão avaliadas juntamente com a validação das sugestões de subcomunidades.

4.3 EXTRAÇÃO DE PALAVRAS-CHAVE

Na extração de palavras-chave(2b), serão identificados os termos que sejam representativos em cada publicação, conforme ilustra a Figura 4.4:

[3]https://scikit-learn.org/

Figura 4.4 Extração de Palavras-Chave.

Fonte: elaborada pela autora (2019)[4]

Dado um documento, é gerada uma matriz com seus termos mais frequentes. Em seguida, é atribuído, um valor *score* para cada um destes termos, é montado um *ranking* (ordenando os termos pelo valor do seus *scores*. Por fim, são listados os "k"termos mais frequentes e geradas as sugestões de palavras-chave. Para a implementação deste experimento, foram utilizados os métodos de extração de palavras-chave, implementados na linguagem *Python*, conforme mostra a Tabela 4.1:

Tabela 4.1 Métodos de Extração de Palavras-Chave

Algoritmo	**Proposta**
YAKE	Disponível no endereço https://pypi.org/project/yake/. Permite configurar a quantidade de n-gram, a quantidade de palavras-chave extraídas e o idioma utilizado.
RAKE	Disponível no endereço https://pypi.org/project/rake/. Determina as palavras-chave que devem ser retornadas com base na frequência das palavras e sua co-ocorrência com outras palavras do texto.
n-gram	Implementado a partir da função *CountVectorizer* da biblioteca *scikit-learn*.
TF-IDF	Implementado a partir da função *TfidfVectorizer* da biblioteca *scikit-learn*.
most_common	Utiliza a função most_common() da biblioteca NLTK[5]. É possível informar a quantidade de termos a serem retornados.
Gensim	Disponível no endereço https://pypi.org/project/gensim/. Possibilita definir a quantidade de palavras-chave a ser extraída e seu *score*. Porém não permite que se defina a quantidade de *n-gram*.

Fonte: elaborado pela autora (2019).

As palavras-chave podem ser compostas por um único termo ou por um conjunto de termos (*n-gram*). Vale dizer que os termos *n-gram* se referem à sequências de *tokens* e o valor de "n"indica quantos termos serão considerados nesta sequência, como por exemplo 2-*gram*, 3-*gram*, etc. Dependendo do método utilizado, torna-se possível definir a quantidade *n-gram* que será considerada e de palavras-chave que devem ser retornadas no decorrer da extração.

[4]Ícones obtidos gratuitamente no site https://www.flaticon.com/

Ressalta-se que para conservar a semântica dos termos extraídos nos 3-*gram*, as *stopwords* foram conservadas. As palavras-chave obtidas pelo métodos listados na Tabela 4.1, passaram por um refinamento manual, com a finalidade de remover as ocorrências duplicadas (formas singular/plural) ou sem relevância semântica.

A avaliação da extração de palavras-chave, no caso dessa solução, será realizada por meio de um estudo de caso realizado junto ao RI UFBA.

4.4 MÉTODOS DE GERAÇÃO DE METADADOS

A anotação semântica da classificação textual foi obtida através dos resultados das predições do classificador binário. Como os algoritmos de *Machine Learning* só trabalham com dados numéricos, as predições são apresentadas neste formato e isto exige, para a anotação, a conversão de tais saídas em formato de texto novamente. Para tanto, foi necessário utilizar a função inversa das funções *Label Encoder* e *Label Binarizer*, que converteram o texto em formato numérico (função *inverse_transform*).

Para isso, foi utilizado o RDF do padrão de metadado *Dublin Core*, com a propriedade *<dcterms:isPartOf>* para associar uma publicação a sua comunidade e subcomunidade. Na classificação multi-hierárquica, a anotação foi feita no formato item - *isPartOf* - comunidade e item - *isPartOf* - subcomunidade. Na classificação binária, a anotação foi feita no formato item - *isPartOf* - subcomunidade1 e item - *isPartOf* - subcomunidade2.

A anotação semântica das palavras-chave de cada documento, por envolver a saída manual da validação das sugestões de palavras-chave (anteriormente refinadas) e as palavras-chave já preenchidas pelo pesquisador, devem ser transcritas para um arquivo de texto ou para uma base de dados para depois alimentar uma variável na implementação em *Python*, para a anotação semântica em formato RDF. Para esta anotação, foi utilizado o metadado *<dc:subject>*.

4.4.1 Anotação da Classificação Textual

O código 4.1, apresentado a seguir, mostra o RDF da anotação gerada pela classificação multi-hierárquica de uma publicação, ressalta-se que os metadados gerados estão sublinhados no código:

```
01 | <?xml version="1.0"?>
02 | <metadata
03 |   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
04 |   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
05 |   xmlns:dc="http://purl.org/dc/elements/1.1/"
06 |   xmlns:dcterms="http://purl.org/dc/terms/">
07 |   <dc:title>
08 |     Plataforma computacional web para calibração de sistemas de
       medição
09 |   </dc:title>
10 |   <dcterms:isPartOf xsi:type="dcterms:title">
11 |     Escola Politécnica
12 |   </dcterms:isPartOf>
```

```
13 |     <dcterms:isPartOf xsi:type="dcterms:title">
14 |       Programa de Pós-Graduação em Engenharia Industrial (PEI)
15 |     </dcterms:isPartOf>
16 |   </metadata>
```

Código 4.1 RDF da classificação multi-hierárquica

A anotação da relação entre o item, a comunidade e a subcomunidade é identificado pela *tag* <*dcterms:isPartOf*>. Neste exemplo, a publicação pertence à comunidade POLI e à subcomunidade PEI:

O código 4.2, apresentado em seguida, evidencia o RDF da anotação gerada pela classificação binária, na qual uma publicação pertencente à subcomunidade PGCOMP foi sugerida para a subcomunidade PGMAT:

```
01 |  <?xml version="1.0"?>
02 |  <metadata
03 |    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
04 |    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
05 |    xmlns:dc="http://purl.org/dc/elements/1.1/"
06 |    xmlns:dcterms="http://purl.org/dc/terms/">
07 |    <dc:title>
08 |      Uso de Sistemas de Transições Modais de Kripke para
        Representacão de Comportamento Parcial no Desenvolvimento
        incremental e interativo de software
09 |    </dc:title>
10 |    <dcterms:isPartOf xsi:type="dcterms:title">
11 |      Programa de Pós-Graduação em Ciência da Computação (PGCOMP)
12 |    </dcterms:isPartOf>
13 |    <dcterms:isPartOf xsi:type="dcterms:title">
14 |      Programa de Pós-Graduação em Matemática (PGMAT)
15 |    </dcterms:isPartOf>
16 |  </metadata>
```

Código 4.2 RDF da classificação binária

A anotação da relação entre o item e subcomunidade é identificada pela *tag* <*dcterms:isPartOf*>. A publicação que pertence à subcomunidade PGCOMP na classificação binária foi sugerida para a subcomunidade PGMAT.

4.4.2 Anotação da Extração de Palavras-Chave

O código 4.3 mostra o RDF da anotação gerada pela validação de palavras-chave de uma publicação. As palavras-chave são armazenadas na *tag* <*dc:subject*>. No estudo de caso, foram aceitas as sugestões de palavras-chave: "Estágio", "Cursos técnicos", "Formação profissional":

```
01 |  <?xml version="1.0"?>
02 |  <rdf:RDF
```

```
03 |  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
04 |  xmlns:dc= "http://purl.org/dc/elements/1.1/"
05 |  xmlns:dcterms="http://purl.org/dc/terms/">
06 |  <rdf:Description rdf:about="https://repositorio.ufba.br/ri/handle
       /ri/18283">
07 |    <dc:title>A recontextualização dos saberes profissionais de
       alunos de cursos técnicos profissionalizantes em experiência
       de estágio</dc:title>
08 |  <dc:subject>
       Saber profissional, Conhecimento profissional, Recontextualização, Estágio,
       Cursos técnicos, Formação profissional </dc:subject>
09 |  </rdf:Description>
```

Código 4.3 RDF da extração de palavras-chave

No caso do <*dc:subject*>, como cada item contém mais de uma palavra-chave, existe a possibilidade de incluir todos os termos em uma única *tag* <*dc:subject*>, como evidenciado no código 4.4:

```
01 |  <dc:subject>}Semantic Annotation; Institutional Repositories;
       Machine Learning</dc:subject>
```

Código 4.4 Exemplo de declaração <dc:subject>

Também é possível utilizar uma *tag* <*dc:subject*> para cada palavra-chave, conforme apresentado no código 4.5:

```
01 |  <dc:subject> Semantic Annotation </dc:subject>
02 |  <dc:subject> Institutional Repositories </dc:subject>
03 |  <dc:subject> Machine Learning </dc:subject>
```

Código 4.5 Exemplo de declaração com várias *tags* <dc:subject>

4.5 ANOTAÇÃO SEMÂNTICA

Esta seção envolve a descrição dos metadados, a codificação do padrão de metadados *Dublin Core* para RDF e apresentará um exemplo de anotação semântica.

4.5.1 Metadados

A etapa de Metadados (3) recebe como entrada os itens classificados pelo classificador multi-hierárquico, as sugestões da classificação binária e as sugestões de palavras-chave. O padrão de metadados que será utilizado é o *Dublin Core*, que já vem como padrão no *DSpace*. O *Dublin Core* possui 15 elementos que podem ser estendidos para elementos qualificadores no formato DC.Elemento.Qualificador, através dos *DCTerms*, con evidenciado na Tabela 4.2 a seguir:

Tabela 4.2 Metadados *DCTerms*

Metadado	**Atributo Correspondente**
dc.title	Título
dc.subject	Palavras-chave
dc.description	Resumo
dc.contributor.author	Autor
dc.date.issued	Data de Publicação
dc.date.submitted	Data de Submissão
dc.identifier.uri	Url Handle do Item
dc.rights	Informações da Licença
dc.type	Tipo do Documento
dc.date.available	Data de disponibilidade
dc.description	Informações adicionais
dc.language.iso	Língua
dc.description.localpub	Local de publicação
dc.publisher	Editora
dc.relation.ispartof	Referência a outro recurso relacionado

Fonte: elaborado pela autora (2019).

Com base no padrão *Dublin Core*, para realizar as anotações semânticas, foram escolhidos os metadados destacados com a cor cinza, conforme tabela acima: *dc.subject* e *dc.relation.ispartof* (ou *dcterms:isPartOf*). O *dc.subject* armazena as palavras-chave de cada publicação e é escolhido para anotar o resultado da validação de palavras-chave enquanto o metadado *dc.relation.ispartof* armazena a referência a outro recurso relacionado, servindo para anotar o resultado da classificação textual.

No caso da classificação binária, são criadas duas instâncias do metadado *dc.relation.ispartof*.

As sugestões de subcomunidade (3a) identificadas na classificação binária devem passar por uma verificação manual para averiguar se são pertinentes ou não. Caso seja identificado que a publicação precisa ser vinculada a outra subcomunidade, por se tratar de uma publicação multidisciplinar, tal sugestão deve ser aceita (3a), pois, caso contrário, a mesma deve ser desconsiderada.

Na Tabela 4.3 demonstra as informações de metadados de uma Dissertação de Mestrado da subcomunidade PGMAT, depositada no RI UFBA:

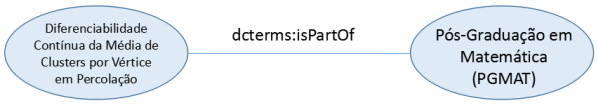
Tabela 4.3 Metadados de uma publicação no RI UFBA

Metadado	Informação Preenchida
dc.title	Diferenciabilidade Contínua da Média de Clusters por Vértice em Percolação
dc.subject	percolação de elos; animais no grafo; diferenciabilidade contínua de k(p).
dc.description	O objetivo desta dissertação é estudar e apresentar alguns resultados em Percolação homogênea de elos para grafos hipercúbicos. O resultado principal deste trabalho é a diferenciabilidade contínua de k(p), o número médio de aglomerados por vértice no grafo. Para obtermos este resultado, serão utilizados argumentos em áreas distintas da Matemática, principalmente em Probabilidade, Combinatória e Teoria dos Grafos.
dc.contributor.author	Diogo Soares Dórea da Silva
dc.date.issued	Maio-2015
dc.date.submitted	13-06-2106
dc.identifier.uri	https://repositorio.ufba.br/ri/handle/ri/19460
dc.type	Dissertação de Mestrado
dc.language.iso	pt-br
dc.description.localpub	Salvador
dc.relation.ispartof	Pós-Graduação em Matemática (PGMAT)

Fonte: elaborado pela autora (2019).

Um exemplo de tripla RDF gerada para o metadado *dcterms:isPartOf* é ilustrado na Figura 4.5, que faz referência a um item faz parte de uma das subcomunidades do RI UFBA, seguindo o formato sujeito-predicado-objeto:

Figura 4.5 Exemplo Tripla RDF de uma publicação.



Fonte: elaborado pela autora (2019).

4.5.2 Codificação

A etapa de Codificação (3d) tem como objetivo gerar o código RDF das triplas geradas na etapa anterior, e para isso, serão utilizadas as triplas obtidas nas etapas anteriores e os metadados escolhidos para a anotação.

O código 4.6, apresentado a seguir, ilustra a anotação do metadado *dcterms:isPartOf*, identificado pela *tag* <*dcterms:isPartOf*>:

```
01 | <?xml version="1.0"?>
02 | <metadata
03 |   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
04 |   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
05 |   xmlns:dc="http://purl.org/dc/elements/1.1/"
06 |   xmlns:dcterms="http://purl.org/dc/terms/">
07 |   <dc:title>
```

```
08 |       título da publicação
09 |     </dc:title>
10 |     <dcterms:isPartOf xsi:type="dcterms:title">
11 |       subcomunidade da publicação
12 |     </dcterms:isPartOf>
13 |     <dcterms:isPartOf xsi:type="dcterms:title">
14 |       sugestão de subcomunidade
15 |     </dcterms:isPartOf>
16 |   </metadata>
```

Código 4.6 RDF da classificação binária

Pelo código, percebe-se que 3 informações de cada item devem ser preenchidas para anotação: o título do trabalho, a subcomunidade do item e a sugestão de subcomunidade (classificação binária). Essas informações podem ser obtidas durante a classificação binária.

O código 4.7 ilustra a anotação do metadado *dc:subject*, identificado pela *tag* <dc:subject>:

```
01 |  <?xml version="1.0"?>
02 |  <rdf:RDF
03 |  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
04 |  xmlns:dc= "http://purl.org/dc/elements/1.1/"
05 |  xmlns:dcterms="http://purl.org/dc/terms/">
06 |  <rdf:Description rdf:about="url do handle da publicaçao">
07 |    <dc:title> título da publicação </dc:title>
08 |  <dc:subject> palavras-chave validadas </dc:subject>
09 |  </rdf:Description>
```

Código 4.7 RDF da extração de palavras-chave

Pelo fato da extração de palavras-chave se tratar de uma tarefa semiautomática (extração de termos mais validação de um especialista), essa anotação deve ser feita manualmente (as palavras-chave são obtidas automaticamente por meio de métodos de extração de palavras-chave e depois são refinadas com a finalidade de simplificar os resultados gerados). Após a etapa de codificação em RDF, os metadados semânticos podem ser armazenados no RI juntamente com a publicação a qual se refere.

4.5.3 Exemplo de Anotação Semântica

Nesta seção será apresentado um exemplo de anotação semântica de itens de um RI. Neste exemplo, um documento de texto em formato pdf (1a) passa por um pré-processamento (1) para conversão do arquivo em formato txt e remoção de caracteres de pontuação, numéricos e de *stopwords*(1b), como ilustrado na Figura 4.6:

Figura 4.6 Exemplo de texto pré-processado

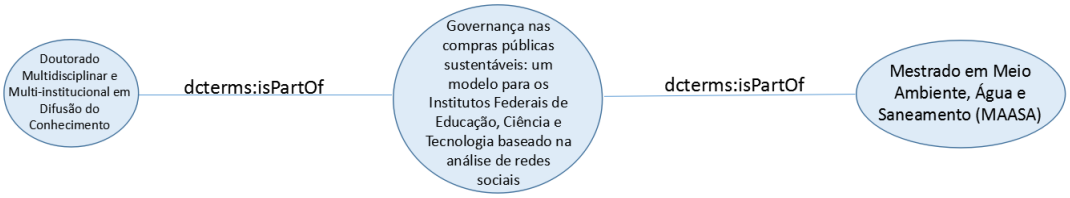
universidade federal bahia ufba universidade estado bahia uneb universidade estadual feira santana uefs instituto federal ciencia tecnologia bahia ifba centro universitário senai cimatec laboratório nacional computação científica programa doutorado multidisciplinar difusão conhecimento eduardo souza seixas governança compras públicas modelo institutos federais ciência tecnologia baseado análise redes sociais salvador eduardo souza seixas governança compras públicas modelo institutos federais ciência tecnologia baseado análise redes sociais tese apresentada programa doutorado institucional multidisciplinar difusão requisito parcial obtenção título doutor difusão professor renelson professor roberto salvador dedico doutoramento deus rogo pesquisa sirva plano agradecimentos professor renelson orientou corrigiu rumo deste barco sempre respeito carinho professor roberto embarcou nesta viagem forma incondicional impulsionou ainda professores indicarem novo norte vida professor luciel ainda sempre pronto professora liliane queiroz orientadora dissertação chegaria professor thyrso permitiu desistisse colegas especial professora lívia pesquisa terminaria tempo irmãos orações apoio especial plínio irmãos apoio compreensão especial sobrinhos anderson fábio matheus dividir demandas amada companheira todas amor sabedoria felicidade presente amor estímulo saudoso amado desembarcou durante jornada tudo sempre meio dificuldade albert einstein eduardo governança compras públicas modelo instituto federal ciência tecnologia meio análise redes tese programa doutorado multidisciplinar difusão universidade federal bahia ufba resumo compras públicas sustentáveis integram

Fonte: elaborado pela autora (2019)

A classificação textual binária (2a) identificou esta publicação como multidisciplinar, sugerindo que ela seja vinculada a outra subcomunidade. Esta classificação leva em consideração situações nas quais a predição seja 1, que indica que a publicação pode pertencer a uma determina subcomunidade, quando na verdade, o rótulo da publicação nesta subcomunidade é 0, que indica que ela não pertence à subcomunidade. Com isso, é sugerido que ela também vinculada à outra subcomunidade.

Esta sugestão de classificação foi avaliada manualmente para verificar sua aceitação ou não. A publicação do exemplo pertence à subcomunidade "Teses de Doutorado (DMMDC)"e foi sugerida para à subcomunidade "Dissertações de Mestrado (MAASA)". Para a avaliação da sugestão foi realizada a leitura do titulo, do resumo e do sumário do trabalho, concluindo-se que a sugestão pode ser aceita (3a). O metadado que deve ser anotado para a classificação binária é o *dcterms:isPartOf*, identificado pela *tag* <*dcterms:isPartOf*>, como ilustra a tripla da Figura 4.7.

Figura 4.7 Exemplo Tripla RDF de uma publicação.



Fonte: elaborado pela autora (2019).

Nesta tripla, a publicação é associada às subcomunidades DIFUSÃO (subcomunidade original) e MAASA (subcomunidade sugerida). De posse dessa informação, é possível identificar que se trata de uma publicação multidisciplinar, de modo que essa descoberta possa contribuir para a diminuição do isolamento da pesquisa, já que identificaria áreas de interesse em que os pesquisadores possam trabalhar em conjunto, já que estão realizando pesquisas na mesma área de conhecimento.

Conforme evidenciado no código 4.8, foi gerado o RDF do metadado representado pela esta tripla, sendo que as informações de metadados obtidas durante a classificação textual se encontram sublinhadas no código:

```
<?xml version="1.0"?>
<metadata
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/">
  <dc:title>
Governança nas compras públicas sustentáveis: um modelo para os
    Institutos Federais de Educação, Ciência e Tecnologia baseado
    na análise de redes sociais
  </dc:title>
  <dcterms:isPartOf xsi:type="dcterms:title">
 Programa de Pós-Graduação Multidisciplinar e Multi-institucional em Difusão
    do Conhecimento (DMMDC)  </dcterms:isPartOf>
    <dcterms:isPartOf xsi:type="dcterms:title">
 Mestrado em Meio Ambiente, Água e Saneamento (MAASA)
  </dcterms:isPartOf>

</metadata>
```

Código 4.8 Exemplo de anotação semântica

Já o código 4.9 ilustra a anotação das palavras-chave desta mesma publicação através do metadado <*dc:subject*>:

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc= "http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/">
<rdf:Description rdf:about="https://repositorio.ufba.br/ri/handle
    /ri/18283">
  <dc:title>Governança nas compras públicas sustentáveis: um
    modelo para os Institutos Federais de Educação, Ciência e
    Tecnologia baseado na análise de redes sociais</dc:title>
<dc:subject>
    Compras públicas sustentáveis;Análise de redes sociais;Governança pública;
    Medição de desempenho;Desenvolvimento sustentável </dc:subject>
</rdf:Description>
```

Código 4.9 RDF da extração de palavras-chave

O metadado <*dc:subject*> conterá as palavras-chave preenchidas pelo autor da publicação, que foram validadas pelo bibliotecário responsável e complementadas pela sugestão de palavras-chave obtidas no experimento da extração. Esse enriquecimento semântico dos metadados, permitirá identificar melhor as publicações e favorecerá à recuperação das mesmas na busca do RI.

Como destacado no capítulo anterior, os principais diferenciais dessa solução são a identificação de trabalhos multidisciplinares através da classificação binária e imple-

mentação de um classificador multi-hierárquico para classificar itens não depositados em um RI. Além disso, a sugestão de palavras-chave auxiliou na complementação de palavras-chave que ocorreu durante a validação dos metadados realizada pelos bibliotecários. Os resultados obtidos da classificação textual e da validação de palavras-chave foram anotados semanticamente no formato RDF do padrão *Dublin Core*, isso permitiu que o enriquecimento semântico do RI, contribuindo para a melhoria da busca a suas publicações.

No próximo capítulo, serão descritos os experimentos e o estudo de caso realizado no decorrer deste trabalho.

Capítulo 5

Neste capítulo serão descritos os experimentos e o estudo de caso realizados e os resultados obtidos.

EXPERIMENTOS E ESTUDO DE CASO

A seção 5.1 descreve a etapa de pré-processamento de texto. As seções 5.2 e 5.3 detalham os experimentos realizados. Já a seção 5.4 descreve o estudo de caso com o RI UFBA. Foram realizados 2 (dois) experimentos no decorrer do projeto, conforme descrito na Tabela 5.1:

Tabela 5.1 Experimentos Realizados

Experimento	**Objetivo**
Classificação textual	Classificação textual multi-hierárquica e binária dos documentos utilizando os algoritmos de classificação: *Naive Bayes*, SVM e *Decision Tree*.
Extração de Palavras-Chave	Extração de palavras-chave de cada documento para sugerir às bibliotecárias durante a validação dos metadados.

Fonte: elaborado pela autora (2019).

Também foi elaborado um estudo de caso, conforme descrito na tabela 5.2, a partir do refinamento das palavras-chave obtidas no experimento da Extração de palavras-chave:

Tabela 5.2 Estudo de Caso Realizado

Estudo de Caso	**Objetivo**
Simulação com o pessoal do SIBI	Bibliotecárias devem validar as sugestões de palavras-chave obtidas em cada documento.

Fonte: elaborado pela autora (2019).

As configurações da atual instalação do RI UFBA e do ambiente utilizado para a realização dos experimentos são descritas no apêndice A.

5.1 PRÉ-PROCESSAMENTO DE TEXTO

A etapa de pré-processamento de texto teve como objetivo de extrair o texto de documentos em formato pdf para o formato txt. A linguagem utilizada para a implementação dos códigos foi o *Python*[1] por esta apresentar uma vasta documentação, boa curva de aprendizado e uma biblioteca disponível para as tarefas de PLN, a biblioteca *Natural Language Toolkit* (NLTK)[2], também codificada em *Python*. Esta biblioteca possui uma série de comandos voltados à execução de de tarefas do PLN como, por exemplo, a a tokenização de sentenças, a identificação de entidades nomeadas, a remoção de *stopwords*, entre outros.

Os textos extraídos foram organizados em pastas, de acordo com cada tarefa realizada durante o pré-processamento e obedecendo à padronização adotada para conservar o nome do arquivo original acrescido da extensão txt, que identifica arquivos de textos. À medida que as etapas foram executadas, o diretório em que os arquivos estavam armazenados foi percorrido com a finalidade de localizar arquivos gerados na etapa anterior, bem como executar os comandos da etapa atual.

A ferramenta *DSpace* possibilita o depósito de publicações em diversos formatos e extensões (textos, imagens, vídeos, áudios, dentre outros). Para este projeto, levou-se em consideração somente as publicações em formato pdf, por este ser o formato mais utilizado no repositório e também por ser de fácil extração textual. Para a realização dos experimentos, foram escolhidos arquivos pdf depositados em diferentes comunidades registradas no repositório.

Para os experimentos, levou-se em consideração somente uma amostra deste conteúdo, e com isso foram selecionados 153 (cento e cinquenta três) itens para compor os conjuntos de treino e teste para a classificação textual e para a extração de palavras-chave. Como não existe uma institucionalização do uso do RI UFBA, algumas comunidades são bem povoadas e outras não. Sendo assim, para a escolha dos documentos, alguns critérios foram levados em consideração, tais como:

- Comunidades de unidade acadêmicas com mais de 1 curso;
- Coleções com mais de 10 documentos depositados;
- Textos em Língua Portuguesa;
- Documentos depositados em subcomunidades de Programas de Pós-Graduação.

Como o algoritmo SVM funciona melhor com bases balanceadas optou-se pela estratégia de balancear a quantidade de documentos nas comunidades, levando em consideração a subcomunidade que possui mais documentos, como no caso do PEI, que possui 28 documentos. Sendo assim, foi necessário replicar os documentos em cada subcomunidade até atingir a quantidade de 28 documentos, totalizando 196 documentos.

[1] https://www.python.org/
[2] http://nltk.org

A tabela 5.3 fornece uma visão geral da quantidade de documentos por comunidade e subcomunidade utilizados nos experimentos, bem como quando as bases foram balanceadas:

Tabela 5.3 Documentos utilizados nos experimentos

Comunidade	Subcomunidade	Total de Itens	Base balanceada
FACED[3]	PPGE[6]	27	28
FACED	DIFUSÃO[7]	25	28
IME[4]	PGCOMP[8]	11	28
IME	PGMAT[9]	16	28
POLI[5]	MAASA[10]	22	28
POLI	PEI[11]	28	28
POLI	PPEQ[12]	24	28
	Total	153	196

Fonte: elaborado pela autora (2019).

- **Conversão de arquivos pdf para o formato txt** Com o auxilio da biblioteca pdfminer.six, codificada em *Python*, foi realizada a conversão dos textos do formato original, extensão pdf para o formato txt, extensão de arquivos de textos.

  Durante a conversão dos documentos do formato PDF para txt, houve perda de informação, tendo em vista que as imagens foram desconsideradas na extração e as tabelas foram desformatadas durante a conversão.

- **Correção da acentuação** Os arquivos das coleções PGCOMP e PGMAT, talvez por alguma inconsistência do *template* Latex, apresentaram problemas de acentuação de caracteres. Sendo assim, para corrigir os problemas de acentuação, foi necessário implementar um código em *Python* visando identificar estas ocorrências e efetuar as correções nos textos. Também foram corrigidos problemas de quebra de linha.

- **Conversão do texto em minúsculas** Depois da correção da acentuação, o próximo passo foi realizar a conversão dos textos em caracteres minúsculos, a fim de facilitar a identificação de *stopwords*.

[3]Faculdade de Educação
[4]Instituto de Matemática e Estatística
[5]Escola Politécnica
[6]Programa de Pós-Graduação em Educação
[7]Programa de Pós-Graduação Multidisciplinar e Multi-Institucional em Difusão do Conhecimento
[8]Programa de Pós-Graduação em Ciência da Computação
[9]Programa de Pós-Graduação em Matemática
[10]Mestrado em Meio Ambiente, Água e Saneamento
[11]Programa de Pós-Graduação em Engenharia Industrial
[12]Programa de Pós-Graduação em Engenharia Química

- **Eliminação de *Stopwords*** Em um texto, algumas palavras não possuem valor semântico e, com base nisso, torna-se interessante que elas sejam eliminadas dos textos. Estas palavras são conhecidas como *stopwords* e gramaticalmente são classificadas como preposições, conjunções, pronomes possessivos e demonstrativos. A biblioteca NLTK permite identificar palavras que sejam *stopwords*. Durante o tratamento do texto, foi utilizada a lista de *stopwords* para a Língua Portuguesa e aquelas *stopwords* identificadas foram removidas do texto.
- **Eliminação de caracteres Numéricos** Durante a etapa de pré-processamento, foi realizada a eliminação de caracteres numéricos, levando-se em consideração que tais caracteres não possuem relevância semântica para o documento extraído.
- **Eliminação de caracteres de Pontuação** Também foi realizada a eliminação de caracteres de pontuação, já que estes serão desconsiderados nas etapas posteriores.

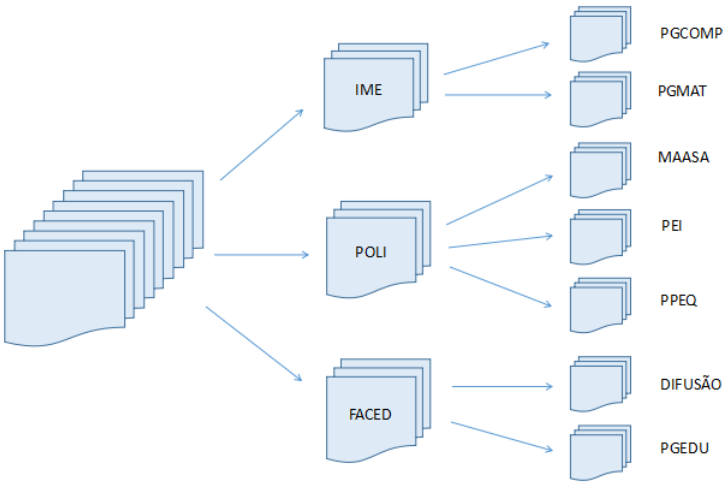
Nas próximas seções será descrito mais detalhadamente os experimentos e o estudo de caso realizado.

5.2 EXPERIMENTO 1: CLASSIFICAÇÃO TEXTUAL

A Classificação Textual permite agrupar textos em categorias pré-definidas de acordo com o seu conteúdo. Para isso, é necessário associar os termos mais relevantes a cada categoria, treinar um classificador para aprender com esse conjunto de textos e a partir de um novo conjunto, predizer a qual categoria cada texto pertence.

Este experimento visa auxiliar as seguintes tarefas: "Dado um documento e seus termos mais relevantes, a qual comunidade e subcomunidade ele deve pertencer?" (classificação multi-hierárquica) e "Existe algum documento que pode ser classificado em mais de uma subcomunidade?"(classificação binária). Conforme a Figura 5.1, um conjunto de documentos pode ser classificado em um repositório multidisciplinar, com base nos termos relevantes identificados em cada documento:

Figura 5.1 Classificação Textual.



Fonte: elaborado pela autora (2019).

A classificação textual tem o objetivo de classificar os documentos de acordo com os termos mais relevantes de cada comunidade e subcomunidade (classificação multi-hierárquica) e identificar trabalhos que possam ser associados a mais de uma área de conhecimento (classificação binária).

A classificação se inicia com a classificação dos documentos nas macro-áreas identificadas, no caso IME, POLI e FACED. Em seguida, dentro de cada macro-área, estes documentos são divididos em micro-áreas, à medida em que esta área se ramifica em outras áreas mais especificas. Para a realização deste experimento, foram instaladas as bibliotecas *numPy* (para trabalhar com *arrays)*, *Pandas* (para visualização de dados) e *scikit-learn* (utilizada para trabalhar com *Machine Learning*), todas implementadas em *Python*.

De posse dos textos obtidos ao fim da etapa anterior, foi gerada uma matriz de frequência dos termos (conhecida como *Bag of Words*). Esta matriz contem os 1000 (mil) termos mais frequentes no conjunto de textos, sendo que cada linha representa um dado documento e cada coluna representa um dos termos (conhecidos como *features*) mais frequentes e sua frequência em cada documento, como ilustra a figura 5.4:

Tabela 5.4 *Bag of Words* dos documentos

	t1	t2	t3	...	t1000
doc1	2	1	1	...	1
doc2	3	0	1	...	1
...	...	...	...	...	...
doc196	1	0	0	...	1

Fonte: elaborado pela autora (2019).

Para gerar essa matriz de termos, foi necessário percorrer o diretório que contém os textos resultantes da etapa de tratamento de texto, adicionando-os em um vetor, em seguida, utiliza-se uma função do *scikit-learn* para gerar a BOW, que ficou com as dimensões 196 x 1000.

5.2.1 Identificação dos metadados para a Classificação Textual

Para identificação dos metadados a serem utilizados para a classificação textual, levou-se em consideração os metadados que apresentassem um conjunto de valores conhecidos. Esta análise baseou-se nas informações que são solicitadas durante o depósito de cada publicação no RI UFBA.

Durante a análise, identificou-se que os metadados comunidade e subcomunidade atendiam a este critério.

- Comunidades: no caso do RI UFBA, as comunidades refletem a estrutura organizacional das unidades acadêmicas (Escolas, Faculdades, Institutos). Porém o RI UFBA, permite abrigar também publicações da Administração Central, como Pró-Reitorias e Superintendências.

- Subcomunidades: são sub-divisões das comunidades e representam as divisões existentes dentro de cada unidade acadêmica como, por exemplo, os Departamentos, os Núcleos e os Programas de Pós-Graduação.

Para o experimento da classificação textual, foram treinados 2 classificadores: um multi-hierárquico para a hierarquia comunidade/subcomunidade e um binário para o metadado subcomunidade, no qual cada classificador foi replicado nos algoritmos *Naive Bayes*, SVM e *Decision Tree*.

Para extrair as informações dos metadados de cada documento, foi necessário implementar um código em *Python* que executa uma consulta *Structured Query Language* (SQL) para buscar estas informações no Sistema Gerenciador de Banco de Dados (SGBD) *Postgres*, que contém a base de dados do RI UFBA. A implementação da extração das informações de metadados será descrita no Apêndice B.

Os vetores obtidos foram utilizados para treinar os classificadores implementados. Os nomes dos documentos também foram armazenados em um terceiro vetor, que foi adicionado aos demais na etapa de análise dos resultados. Como os algoritmos de *Machine Learning* só trabalham com informações numéricas, realizou-se a conversão dos vetores de metadados para o formato numérico, como descrito no Apêndice B.

5.2.2 Implementação dos Classificadores

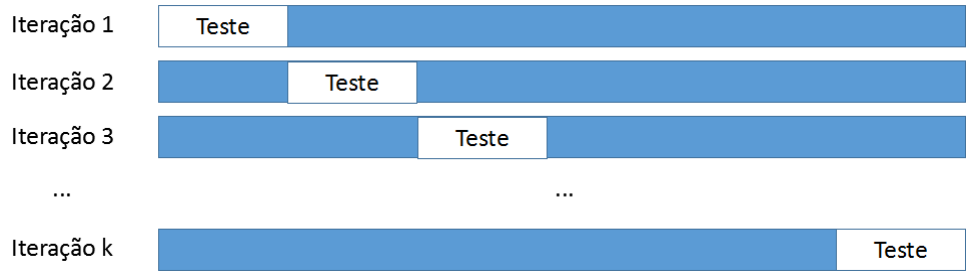
Para a implementação dos classificadores multi-hierárquico e binário, foram escolhidos os algoritmos de *Machine Learning Naive Bayes*, SVM e *Decision Tree*, que são algoritmos de classificação e utilizam o aprendizado supervisionado (quando se deseja prever uma classe/rótulo). A linguagem utilizada para a implementação dos códigos foi o *Python*, pelo fato desta possuir uma biblioteca especifica para *Machine Learning*, a biblioteca *scikit-learn* [3].

Um classificador multi-hierárquico seria útil caso fosse necessário realizar a classificação de um grande volume de documentos, como, por exemplo, um mutirão para povoamento de uma comunidade com publicações antigas - estes documentos não estão organizados por subcomunidade. Para a implementação do classificador multi-hierárquico, foi necessário treinar um classificador em 1º nível, a fim de classificar os documentos de acordo com a comunidade (predizer a comunidade de cada documento). Em seguida, os documentos de cada comunidade foram treinados para predizer a qual subcomunidade cada documento pertencia. Para treinar os itens de 1º nível, foi necessário gerar a BOW de todos os documentos, bem como o vetor de comunidade. Para treinar os itens de 2º nível, foi necessário gerar a BOW e vetor de cada subcomunidade. Ressalta-se que um documento só pode ser associado a uma determinada subcomunidade se antes ele estiver associado a comunidade correspondente.

De posse da matriz *Bag of Words* e dos vetores numéricos dos metadados, o conjunto de documentos foi dividido em *folds*, nos quais, em cada iteração, um conjunto de documentos pertence ao conjunto de treino e o restante pertence a um conjunto de teste, conforme Figura 5.2:

[3] https://scikit-learn.org/

Figura 5.2 Conjunto de treino e teste.



Fonte: elaborado pela autora (2019).

O conjunto de treino serve para que o classificador aprenda com os dados e possa classificar novos itens. Já o conjunto de teste serve para validar e avaliar estes classificadores, sendo que no decorrer das iterações, cada documento tem a possibilidade de pertencer tanto ao conjunto de treino quanto de teste.

Os experimentos foram configurados para o uso de 4-*fold*. Como o conjunto total contava com 196 documentos, cada *fold* ficaria com a mesma quantidade de itens, ou seja, 49 documentos. Com isso em cada iteração, o conjunto de treino contou com 147 documentos e o conjunto de teste com 49 documentos, além do que ao realizar 4 iterações, o tempo de execução da classificação não ficaria tão extenso.

Enquanto isso, um classificador binário serviria para verificar se existem casos em que uma publicação pode ser associada a mais de uma subcomunidade (publicação multidisciplinar).

Como no RI UFBA, uma publicação pode ser considerada multidisciplinar pelo fato de se referir a mais de uma área de conhecimento, um classificador binário pode sugerir/-recomendar que esta publicação seja associada a mais de uma subcomunidade de acordo com o resultado destas predições. Nos casos em que a predição seja diferente do conjunto de teste, pode ser uma sugestão que a publicação também seja associada a outra subcomunidade.

Os classificadores foram implementados com os algoritmos *Naive Bayes*, SVM e Árvores de Decisão, ajustados na parametrização padrão.

5.2.3 Avaliação da Classificação Textual

Finalizadas as predições dos classificadores implementados, pode-se dizer que a avaliação foi realizada de acordo com as principais métricas: acurácia, precisão, *recall* e f1-*score*. Destaca-se que os valores retornados para cada métrica representam as médias aritméticas dos valores obtidas durante as iterações, considerando somente quatro casas decimais após a virgula.

5.2.3.1 Classificador Multi-Hierárquico A Tabela 5.5 apresenta os resultados obtidos na avaliação da classificação multi-hierárquica com o algoritmo *Naive Bayes*:

Tabela 5.5 Avaliação Classificação multi-hierárquica - algoritmo *Naive Bayes*

	acurácia	precisão	recall	f1-score
1º nível	0.8775	0.9081	0.8670	0.8578
POLI	0.8095	0.8509	0.8095	0.8096
FACED	0.8571	0.8767	0.8571	0.8552
IME	**0.9642**	**0.9722**	**0.9642**	**0.9635**

Fonte: elaborado pela autora (2019).

O classificador *Naive Bayes* apresentou ótimos resultados, sobretudo na comunidade IME. que apresentou um valor alto em todas as métricas.

A Tabela 5.6 apresenta os resultados obtidos na avaliação da classificação com o algoritmo SVM:

Tabela 5.6 Avaliação Classificação multi-hierárquica - algoritmo SVM

	acurácia	precisão	recall	f1-score
1º nível	**0.8622**	**0.8186**	**0.8452**	**0.8242**
POLI	0.3809	0.4500	0.3809	0.2561
FACED	0.6071	0.5315	0.6071	0.4997
IME	0.5535	0.5152	0.5535	0.4330

Fonte: elaborado pela autora (2019).

O algoritmo SVM apresentou bons resultados nas métricas em 1º nível, porém mesmo com a base balanceada, ou seja, com a mesma quantidade de documentos em todas as comunidades e a parametrização com *kernel* linear, os resultados em 2º nível foram baixos, provavelmente porque o classificador não deve ter definido um bom hiperplano para identificar os documentos pertencentes a cada subcomunidade.

A Tabela 5.7 apresenta os resultados obtidos na avaliação com o algoritmo de Árvores de Decisão:

Tabela 5.7 Avaliação Classificação multi-hierárquica - algoritmo Árvore de Decisão

	acurácia	precisão	recall	f1-score
1º nível	0.7551	0.7064	0.7678	0.7546
POLI	0.6666	0.6666	0.6309	0.6338
FACED	**0.9464**	0.9531	**0.9464**	**0.9461**
IME	0.8571	**0.9722**	0.8214	0.8857

Fonte: elaborado pela autora (2019).

No algoritmo Árvores de Decisão, os melhores resultados forem obtidos nas subcomunidades da FACED, provavelmente pelo fato dos documentos dessa comunidade apesar de

pertenceram a 2 cursos diferentes, apresentarem uma maior homogeneidade em relação às demais comunidades.

Em relação à acurácia, que indica a quantidade de acertos (TP e TN), considerando o total das predições, o classificador que apresentou os melhores valores foi o *Naive Bayes*, apesar do maior valor de acurácia obtido ter sido na comunidade FACED do classificador Árvores de Decisão.

Considerando a precisão, que calcula a quantidade de itens classificados como positivo que realmente são (TP e FP), o algoritmo *Naive Bayes* também apresentou os melhores resultados, juntamente com os obtidos pelas comunidades FACED e IME no classificador Árvores de Decisão.

Levando-se em conta o *recall*, que calcula a relação entre os resultados positivos e os os que realmente são (TP e FN), os resultados obtidos no classificador *Naive Bayes* foram altos, juntamente com os resultados obtidos nas comunidades FACED e IME no classificador Árvores de Decisão.

Por fim, o *f1-score*, representado pela média harmônica entre os valores obtidos nas métricas precisão e *recall*, apresentou valores altos no classificador *Naive Bayes*, no 1º nível do classificador SVM e nas subcomunidades FACED e IME no classificador de Árvores de Decisão,

Para o cenário do povoamento, indica-se o algoritmo *Naive Bayes*, pois apresentou os melhores valores no geral, provavelmente por trabalhar com a independência de *features*, que nesse caso são representadas pelas palavras que compõem o conjunto de documentos utilizados neste experimento.

Esse experimento validou a hipótese H4 apresentada na seção 1.3 deste trabalho.

5.2.3.2 Classificador Binário No que concerne a classificação textual binária, considerando um conjunto de 153 publicações utilizadas nos experimentos, quando estas são replicadas com a finalidade de balancear as bases, depois de uma verificação manual (leitura de título, resumo e sumário), foram identificados 28 trabalhos multidisciplinares, conforme listado no Apêndice D.

Na implementação para a geração dos conjuntos de treino e teste foi utilizado o método *k-fold*, que se mostrou mais eficiente por possibilitar que todas as instâncias possam ser utilizadas nos dois conjuntos no decorrer das iterações.

Para identificar estas predições, foram verificados os casos em que a predição retornou o valor 1 (pertence à subcomunidade) porém no conjunto de teste esse valor é 0 (não pertence à subcomunidade), a fim de identificar os casos em um item pode ser sugerido para outra subcomunidade.

As sugestões obtidas com o classificador foram conferidas manualmente com a leitura do título, do resumo e do sumário de cada trabalho, com a finalidade de verificar se estas eram pertinentes ou não.

A avaliação da classificação binária foi composta da avaliação das métricas retornadas em cada algoritmo e das sugestões de trabalhos multidisciplinares. A Tabela 5.8 mostra os resultados das métricas obtidas na avaliação da classificação binária com o algoritmo *Naive Bayes*:

Tabela 5.8 Avaliação classificação binária - algoritmo *Naive Bayes*

	acurácia	precisão	recall	f1
MAASA	0.9234	0.6809	0.9642	0.7885
PGCOMP	0.9744	0.9166	0.8928	0.9038
DIFUSÃO	0.8010	0.4276	0.8571	0.5662
PPGE	0.8673	0.7333	**1.0**	0.8004
PEI	0.7346	0.3642	0.8928	0.5089
PPEQ	0.9081	0.6548	**1.0**	0.7778
PGMAT	**1.0**	**1.0**	**1.0**	**1.0**

Fonte: elaborado pela autora (2019).

O classificador *Naive Bayes* retornou inúmeras sugestões, sendo que depois da depois da conferência com os trabalhos multidisciplinares identificados no conjunto de documentos (listados no Apêndice D), foram aceitas 13 sugestões que serão listadas na Tabela 5.11.

O classificador *Naive Bayes* retornou valor 1.0 para a subcomunidade PGMAT, fato que é justificado pela ausência de sugestões de trabalhos para essa subcomunidade. Os valores de acurácia foram altos, porém a subcomunidade PEI apresentou um valor mais baixo, explicado pela quantidade de sugestões retornadas por esse classificador. Na precisão, as subcomunidades PEI e DIFUSÂO apresentaram valores bem baixos, justificado pela quantidade de sugestões retornadas de documentos dessas subcomunidades. Apesar dos valores baixos em algumas subcomunidades, os valores de *recall* foram altos, com destaque para os valores 1.0 obtidos pelas subcomunidades PPGE, PPEQ e PGMAT.

Com relação ao *f1-score*, alguns valores foram medianos por conta do valor de precisão baixo obtido em algumas subcomunidades. Grande parte das sugestões retornadas por esse classificador, depois de efetuada a conferência, verificou-se que tratavam-se de Falsos Positivos, pois a sugestão de trabalho multidisciplinar não era pertinente. Ressalta-se que boa parte das predições identificadas como Falso Positivo, possuem de certa forma, alguma lógica, já que sugeriram o vinculo entre trabalhos da mesma área de conhecimento. Nesse caso, muitas predições de trabalhos PEI, por exemplo, apontaram sugestões para outra subcomunidade da POLI.

Já Tabela 5.9 apresenta os resultados das métricas obtidas na avaliação da classificação binária com o algoritmo SVM, na qual não foram retornadas sugestões de trabalhos:

Tabela 5.9 Avaliação classificação binária - algoritmo SVM

	acurácia	precisão	recall	f1
MAASA	0.8673	**0.5**	0.0714	0.125
PGCOMP	0.8724	**0.5**	0.1071	0.1736
DIFUSÃO	0.8673	**0.5**	0.0714	0.125
PPGE	0.8571	0.0	0.0	0.0
PEI	**0.8775**	**0.5**	**0.1428**	**0.2222**
PPEQ	0.8673	**0.5**	0.0714	0.125
PGMAT	0.8673	**0.5**	0.0714	0.125

Fonte: elaborado pela autora (2019).

O classificador SVM retornou muitos valores TP e TN, o que justifica os valores altos obtidos na acurácia e a ausência de sugestões de trabalhos. Nesse classificador, foram identificadas várias ocorrências de Falso Negativo, que ocorrem quando a classe real possui valor 1 mas a predição identificou como 0. Os valores obtidos nas demais métricas foram baixos, provavelmente porque o classificador não conseguiu definir um bom hiperplano para identificar as *features* que definem cada subcomunidade.

Na precisão, excetuando a subcomunidade PPGE, que zerou o valor dessa métrica, as demais subcomunidades obtiveram valor 0.5. No *recall*, os valores obtidos foram baixos, justificada pelo alto número de FN. Os baixos valores obtidos nas métricas precisão e *recall*, impactaram nos valores baixos obtidos no *f1-score*. Não foram apresentados valores FP, o que justifica a ausência de sugestões de trabalhos. A subcomunidade PPGE apresentou valores 0 nas métricas precisão, *recall* e *f1-score*.

A Tabela 5.10 apresenta os resultados das métricas obtidas na avaliação da classificação binária com o algoritmo Árvore de Decisão, na qual também não foram retornadas sugestões de trabalhos:

Tabela 5.10 Avaliação classificação binária - algoritmo Árvore de Decisão

	acurácia	precisão	recall	f1
MAASA	0.9438	0.7980	0.7857	0.7666
PGCOMP	0.8367	0.4077	0.3571	0.3083
DIFUSÃO	0.8928	0.7409	0.7499	0.7223
PPGE	0.8724	0.7455	0.75	0.7091
PEI	0.8469	0.4781	0.5	0.3993
PPEQ	0.9234	0.8854	0.75	0.8031
PGMAT	**0.9897**	**0.9444**	**1.0**	**0.9687**

Fonte: elaborado pela autora (2019).

O classificador com Árvores de Decisão apresentou valores altos na acurácia, o que justifica não ter retornado sugestões de trabalhos. O classificador retornou muitos valores TP e TN, o que justifica os valores altos obtidos na acurácia e a ausência de sugestões de

trabalhos. Os melhores valores obtidos foram na subcomunidade PGMAT, inclusive com um valor 1.0 na métrica *recall*. A subcomunidade PGCOMP e PEI apresentaram valores baixos de precisão, *recall* e consequentemente *f1-score*. Não foram apresentados valores FP, o que justifica a ausência de sugestões de trabalhos.

Com o refinamento das predições obtidas neste experimento, foram identificados 13 trabalhos multidisciplinares, como lista a Tabela 5.11:

Tabela 5.11 Trabalhos Multidisciplinares identificados na classificação binária

TÍTULO DA PUBLICAÇÃO	SUBCOMUNIDADE	SUGESTÃO
ANÁLISE DA EVIDENCIAÇÃO AMBIENTAL DOS RELATÓRIOS DE SUSTENTABILIDADE DAS INDÚSTRIAS CERVEJEIRAS BRASILEIRAS	MAASA	PEI
GOVERNANÇA NAS COMPRAS PÚBLICAS SUSTENTÁVEIS: UM MODELO PARA OS INSTITUTOS FEDERAIS DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA BASEADO NA ANÁLISE DE REDES SOCIAIS	DIFUSÃO	PGCOMP
RPG DIGITAL E SEGURANÇA PÚBLICA: UMA PROPOSTA DE APLICAÇÃO PEDAGÓGICA PARA INSTRUÇÃO POLICIAL MILITAR	DIFUSÃO	PPGE
CONTRADIÇÕES E POSSIBILIDADES DE SUPERAÇÃO NO TRABALHO PEDAGÓGICO PARA A PRÁTICA DA EDUCAÇÃO AMBIENTAL NOS ANOS INICIAIS DO ENSINO FUNDAMENTAL: UM ESTUDO DE CASO	PPGE	MAASA
DESIGN COGNITIVO COLABORATIVO PARA AMBIENTES VIRTUAIS: O CASO DO PORTAL TBC CABULA	DIFUSÃO	PGCOMP
METODOLOGIA PARA DEFINIR UM SISTEMA DE INDICADORES DE DESEMPENHO SOCIOAMBIENTAL: O ESTUDO DE CASO COELBA	PEI	MAASA
APLICAÇÃO DA FILTRAÇÃO INTERMITENTE EM LEITO DE AREIA E DE ESCÓRIA DA METALURGIA DO COBRE NO TRATAMENTO DE ESGOTOS COM ÊNFASE EM REÚSO	PEI	MAASA
ESTIMAÇÃO DE PARÂMETROS, INFERÊNCIA E CONTROLE DE PROPRIEDADES DE QUALIDADE DE UM PROCESSO DE COPOLIMERAZAÇÃO DE ETENO	PEI	PPEQ
DESENVOLVIMENTO DE NANOPARTÍCULAS POLIMÉRICAS CONTENDO ÓLEO ESSENCIAL DE CITRONELA (CYMBOPOGON WINTERIANUS)	PEI	PPEQ
INVESTIGAÇÃO SOBRE A REMOÇÃO BIOLÓGICA SIMULTÂNEA DE AMÔNIO E NITRITO UTILIZANDO BIOMASSA EM SUSPENSÃO ORIUNDA DE SISTEMA DE LODO ATIVADO	PEI	PPEQ
ESTRATÉGIAS DE CONTROLE APLICADAS A REATORES DE POLIMERIZAÇÃO DE ETENO EM SOLUÇÃO	PEI	PPEQ
DESENVOLVIMENTO DE CATALISADORES PARA ABATIMENTO DE FENÓIS EM EFLUENTES INDUSTRIAIS	PPEQ	PEI
UTILIZAÇÃO DA FIBRA DE SISAL TRATADA COM LÍQUIDO IÔNICO COMO SOLVENTE DE ÓLEOS EM ÁGUA	PPEQ	MAASA

Fonte: elaborado pela autora (2019)

Para o cenário das sugestões de trabalhos multidisciplinares, obtidas na classificação binária, considerando as sugestões retornadas e análise dos valores obtidos nas métricas, o que se mostrou mais eficiente na classificação foi o *Naive Bayes*.

O experimento da classificação binária validou a hipótese H3 apresentada na seção 1.3 desta dissertação.

5.3 EXPERIMENTO 2: EXTRAÇÃO DE PALAVRAS-CHAVE

O experimento da Extração de Palavras-Chave tem a finalidade de auxiliar a tarefa "Quais palavras-chave descrevem melhor um determinado documento?". No *Dublin Core*, padrão de metadados utilizado pela ferramenta *DSpace*, o metadado responsável por armazenar

as palavras-chave de cada publicação é o *dc:subject*. As palavras-chave podem ser armazenadas dentro da mesma declaração *dc:subject* ou pode ser gerada uma declaração *dc:subject* para cada palavra-chave identificada na publicação.

Neste experimento, foram extraídas as palavras-chave de 50 publicações pertencentes a 3 comunidades do RI UFBA, utilizando os métodos apresentados na seção 2.2 desta dissertação. O resultado deste experimento será utilizado durante o estudo de caso que consistirá na validação de palavras-chave pelas bibliotecárias do SIBI.

A implementação, envolveu percorrer o diretório no qual os arquivos estavam localizados, bem como executar o código de cada um dos métodos de extração de palavras-chave escolhidos. Além disso, caso o método retorne 1-*gram* e 2-*gram*, as *stopwords* foram desconsideradas. Em alguns métodos, a remoção de palavras-chave já vem como configuração *default*. Para a extração foram considerados os termos 1-*gram*, 2-*gram* e 3-*gram*. Ressalta-se que nos 3-*gram* foram conservadas as *stopwords* para conservar o sentido dos termos identificados.

Para avaliação dos métodos, foi realizada a extração das palavras-chave dos documentos da subcomunidade PGCOMP. Este teste envolveu a extração em si e a avaliação das palavras-chave retornadas. Para tal verificação, foram realizados os seguintes passos, seguindo às recomendações da equipe SIBI:

- Leitura do titulo do trabalho
- Leitura dos resumos
- Leitura do sumário

Como as sugestões de palavras-chave (3c) obtidas na extração muitas vezes retornam n-*gram* repetidos ou sem relevância semântica, será necessário realizar um refinamento manual destas sugestões, Esse refinamento consiste em realizar os seguintes passos em cada um dos textos, a fim de avaliar os resultados obtidos em cada método:

- Leitura dos resumos;
- Em cada resumo, destacar os termos que julgar serem palavras-chave;
- Olhar os resultados da extração, e de acordo com a leitura do resumo, marcar as palavras-chave;
- Considerar somente uma das formas de cada termo: singular ou plural;
- Verificar nos resultados se algum termo se repete com frequência e julgar se trata de uma palavra-chave.

5.3.1 Avaliação dos Métodos de Extração de Palavras-Chave

A Tabela 5.12 apresenta uma análise dos resultados obtidos por cada método, no decorrer da extração:

Tabela 5.12 Metódos utilizados na extração de palavras-chave

Método	**Avaliação**
YAKE	Não foi necessário remover as *stopwords*, a configuração de 1-gram foi foia ajustada para retornar 20 resultados. Os resultados de 2-gram foram bons, porém os resultados de 3-gram não foram bons.
Gensim	Só oferece a opção de retornar 1-gram, além disso retornou uma quantidade divergente da passada como parâmetro e considerou as formas singular e plural de vários termos.
CountVectorizer	Possui um parâmetro para filtrar as *stopwords* e obteve com termos 1-gram os melhores resultados e com termos 2-gram os resultados foram razoáveis. Porém com termos 3-gram não retornou resultados satisfatórios.
RAKE	Retornou resultados pouco relevantes nos termos 1-gram, 2-gram e 3-gram.
most_common	A implementação *default* só considera 1-gram mas existem implementações com 2-gram e 3-gram. Os resultados 1-gram foram bem melhores que os demais.
TF-IDF	Retornou foi configurado para retornar 20 resultados para 1-gram e obteve bons resultados. O retorno 3-gram foi razoável.

Fonte: elaborado pela autora (2019).

Os melhores resultados foram obtidos através dos métodos YAKE, TF-IDF, *CountVectorizer* e *most_common*. Conclui-se, a partir dai, que os resultados com 1-*gram* foram os melhores. A solução final talvez seja extrair os n-*gram* e refinar os resultados para as bibliotecárias.

A remoção das *stopwords* nos termos 1-*gram* e 2-*gram* colaborou para a melhoria dos resultados em relação aos testes anteriores. Observou-se também que muitos resultados desta extração precisaram passar por um refinamento devido às várias ocorrências de uma mesma palavra-chave nos diferentes métodos, além das múltiplas formas (singular e plural), locuções adverbiais, etc. Para melhorar estes resultados, uma sugestão seria utilizar um algoritmo de similaridade de *strings* para corrigir essas redundâncias.

Depois da extração e refinamento das palavras-chave, foi realizado um estudo de caso com as bibliotecárias do SIBI, responsáveis pela validaçao dos metadados das comunidades POLI, FACED e IME.

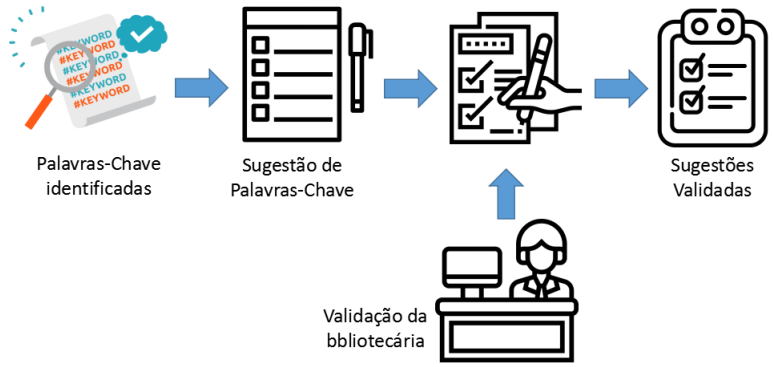
5.4 ESTUDO DE CASO COM A EQUIPE SIBI

Este estudo de caso tem por finalidade auxiliar na seguinte tarefa: "Dado um conjunto de termos sugeridos para identificar um dado documento, quais deles devem ser validados em seu metadado palavra-chave?".

A Figura 5.3 ilustra o estudo de caso para validação das sugestões de palavras-chave pela bibliotecária. Essa validação inicia-se depois da extração das palavras-chaves pelos métodos identificados na literatura. Em seguida, o resultado dessa extração passa por um refinamento após a leitura da ficha catalográfica, do resumo e das palavras-chave (preen-

chidas pelo autor) de cada publicação. As sugestões refinadas passarão pela validação de uma bibliotecária. Por fim, são listadas as sugestões validadas pela bibliotecária.

Figura 5.3 Estudo de Caso: Validação das Palavras-Chave.



Fonte: elaborada pela autora (2019)[4]

.

O estudo de caso realizado no RI UFBA aconteceu em três momentos diferentes pois, as bibliotecárias responsáveis pela validação dos metadados de cada comunidade são lotadas em unidades diferentes - o que inviabilizou a realização do estudo de caso com todas ao mesmo tempo. Sendo assim, cada momento contou com a participação da bibliotecária responsável por cada comunidade.

Cada momento durou, em média 2 (duas) horas, sendo que cerca 20 minutos foram destinados à contextualização do projeto e 1:40h para as demais atividades: assinatura do termo de consentimento livre e esclarecido, validação das sugestões de palavras-chave e preenchimento do questionário.

O estudo foi conduzido com a apresentação da proposta do Mestrado, a descrição dos experimentos realizados e do estudo de caso. Em seguida, foi efetuada a leitura do termo de consentimento livre e esclarecido, no qual todas as bibliotecárias concordaram em participar deste. Para realização do estudo de caso foi necessário extrair as palavras-chave de 50 publicações distribuídas em 7 subcomunidades, conforme apresentado na Tabela 5.13 a seguir:

O momento com a comunidade FACED ocorreu em 2 dias diferentes pois não foi possível finalizar o estudo de caso no mesmo dia, pelo fato dos resumos destes trabalhos serem mais extensos que os das demais comunidades.Em geral, a validação de metadados no RI UFBA envolve a validação de inúmeros metadados, porém, no estudo de caso somente foi considerada a validação das palavras-chave.

A validação das palavras-chave consistiu na leitura da ficha catalográfica, do resumo e das palavras-chave de cada publicação. Este procedimento seguiu a orientação passada pelo SIBI para validação dos metadados no RI UFBA. Estas sugestões são resultado do experimento de extração de palavras-chave descrito na seção 5.3, no qual foram utilizados alguns métodos de identificação de palavras-chave. As palavras-chave identificadas foram

[4]Ícones obtidos gratuitamente no site https://www.flaticon.com/

Tabela 5.13 Documentos utilizados no Estudo de Caso

Subcomunidade	Quantidade de Documentos
PGMAT	8
PGCOMP	8
MAASA	5
PEI	6
PPEQ	6
PPGE	9
DIFUSÃO	8

Fonte: elaborado pela autora (2019)

refinadas pela autora, seguindo o mesmo procedimento orientado pelo SIBI. Esta tarefa permitiu eliminar resultados repetidos bem como ocorrências na forma singular e plural de uma mesma palavra-chave.

Neste experimento foi avaliado se a sugestão de palavras-chave pode favorecer na melhor descrição dos itens deste repositório, bem como contribuir para a validação de metadados pelas bibliotecárias do SIBI. Após a validação dos metadados, pelas bibliotecárias, os *tokens* foram associados ao RDF do padrão de metadados *Dublin Core* durante a anotação semântica que servirá para anotar os itens neste padrão de metadados, transformando o repositório em um repositório semântico.

Durante a leitura do resumo, algumas palavras foram sublinhadas no texto para indicar que poderiam ser eventuais palavras-chave da publicação. Em cada publicação deveria ser marcado um "x"ao lado da sugestão aceita. Caso não fosse aceita nenhuma das sugestões deveria ser escrito um "ok"para informar que a publicação já havia sido validada. Houveram casos em que algumas palavras-chave foram sugeridas pela bibliotecária.

Ressalta-se que as palavras-chave validadas servem apenas para analisar as sugestões aceitas de cada publicação e não implicará em alteração nas informações contidas atualmente nos metadados destas publicações no RI UFBA. Por fim, cada bibliotecária preencheu um questionário no qual foi possível avaliar se as palavras-chave sugeridas em cada publicação poderiam contribuir durante a validação dos metadados.

Este questionário pode ser integralmente visualizado no Apêndice C, sendo composto por questões abertas e fechadas, distribuídas da seguinte forma:

- 4 questões dicotômicas (alternativas Sim/Não)
- 4 questões com Escala *Likert*
- 3 questões abertas

A partir do questionário, foram obtidos as seguintes respostas para as questões dicotômicas (questões de 4 a 7), conforme Tabela 5.14 abaixo:

Tabela 5.14 Respostas das questões fechadas - Sim ou Não.

	POLI	Exatas	FACED
Questão 4	Não	Sim	Não
Questão 5	Não	Sim	Não
Questão 6	Sim	Sim	Não
Questão 7	Não	Sim	Não

Fonte: elaborado pela autora (2019).

Estas questões de referem à perguntas sobre o conhecimento das bibliotecárias dos tópicos abordados nesta dissertação: classificação textual, *machine learning*, extração de palavras-chave e anotações semânticas.

Já as questões que utilizam a escala *Likert* (questões de 8 a 11), obtiveram as respostas evidenciadas na Tabela 5.15:

Tabela 5.15 Respostas das questões fechadas - Escala *Likert*.

	POLI	Exatas	FACED
Questão 8	Discordo Parcialmente	Concordo Parcialmente	Concordo Totalmente
Questão 9	Discordo Parcialmente	Concordo Parcialmente	Concordo Parcialmente
Questão 10	Concordo Parcialmente	Concordo Parcialmente	Concordo Totalmente
Questão 11	Concordo Totalmente	Concordo Parcialmente	Indiferente

Fonte: elaborado pela autora (2019).

Estas questões se referem à condução do estudo de caso, avaliando sobre a quantidade e qualidade das palavras-chave sugeridas, a quantidade documentos validados e a duração do estudo de caso.

Com base nas respostas das questões abertas e entrevistas *in loco*, de um modo geral, foi constatado que as palavras-chave sugeridas foram adequadas e que a quantidade foi suficiente. A quantidade de sugestões apresentadas variou em cada publicação, sendo assim, não houve uma quantidade definida de palavras-chave - até porque isso depende também do refinamento realizado - e, por não existir uma quantidade limite de palavras-chave configurada no RI UFBA para cada publicação.

A quantidade de documentos utilizados no estudo de caso também foi considerado adequado, bem como o tempo de realização da atividade, apesar dele ter sido extrapolado em uma das comunidades, como já informado anteriormente.

Durante a realização do estudo de caso, todas as bibliotecárias chamaram a atenção para a utilização de um vocabulário controlado (conhecidos como tesauros) como forma de complementar a lista de palavras-chave. Existem diversas bases de dados nas mais variadas áreas do conhecimento nas quais é possível consultar quais termos podem complementar a descrição da publicação e facilitar sua recuperação na busca. O questionário também abriu um espaço para que as bibliotecárias pudessem avaliar a credibilidade dos experimentos e se os mesmos poderiam contribuir para a melhoria de sua atividade na validação de metadados:

Avaliou-se que este experimento pode contribuir com o trabalho da validação de metadados apesar de nem sempre os termos apresentados serem os mais consistentes. Sugeriu-se como retorno que o resultado deste trabalho fosse socializado com os bibliotecários do SIBI. Em vários momentos do estudo de caso, foi relatado que os usuários reportam constantemente problemas de navegação no RI UFBA, seja por *links* quebrados ou por problemas de usabilidade da ferramenta.

Existe em andamento um projeto para atualização da versão do *DSpace* do RI UFBA da versão 3.2 para 5.7. Isso visa trazer mais funcionalidades ao RI UFBA e também corrigir as inconsistências apresentadas na versão atual. Esta atualização também envolve mudança do *template* utilizado no RI.

A hipótese H2 foi parcialmente validada com a realização do estudo de caso com as bibliotecárias do SIBI com a validação das palavras-chave porém não houve validação da classificação textual binária.

Os resultados obtidos no experimento da classificação textual e do estudo de caso servem para a anotação semântica dos itens de um repositório. Para tanto, deve ser utilizado o RDF do padrão *Dublin Core*, conforme descreve a seção 4.5 desta dissertação. A solução apresentada trata-se de um trabalho experimental, sendo assim as triplas geradas na anotação semântica não foram publicadas na *Linked Open Data* e, com isso, não foi possível realizar uma busca semântica.

No próximo capítulo, serão apresentadas as considerações finais deste trabalho, com as contribuições alcançadas no decorrer deste trabalho e sugestões de trabalhos futuros.

Capítulo

6

CONSIDERAÇÕES FINAIS

Neste trabalho, buscou-se responder as questões de pesquisa elencadas na seção 1.3, ou seja, este trabalho descreveu uma solução com o objetivo de anotar semanticamente os itens de um repositório acadêmico no *Dspace* de maneira semiautomática, utilizando o padrão RDF do *Dublin Core* a partir dos resultados obtidos na classificação textual e validação das sugestões de palavras-chave.

Para a realização deste trabalho, foi implementado um classificador multi-hierárquico com o objetivo de classificar itens ainda não depositados no repositório e que não estejam organizados por comunidade e subcomunidade, bem como um classificador binário com o objetivo de identificar e sugerir trabalhos multidisciplinares de modo que eles sejam também vinculados a outras subcomunidades, Na implementação dos 2 classificadores, utilizou-se a estratégia do *k-fold* com a finalidade de possibilitar que todos os documentos fossem utilizados durante as iterações nos conjuntos de treino e de teste.

Na avaliação da classificação textual, concluiu-se que o algoritmo *Naive Bayes* apresentou os melhores resultados nos dois classificadores, tanto o multi-hierárquico quanto o binário.

No classificador multi-hierárquico, o classificador implementado com o algoritmo *Naive Bayes* apresentou as métricas com melhores resultados no geral, apesar de que em algumas métricas os melhores valores tenham sido em outros classificadores, como descrito na seção 5.2.3.1. Sendo que o classificador *Naive Bayes* foi o indicado para o cenário de povoamento de um RI. Esse desempenho deve-se provavelmente pelo fato desse algoritmo trabalhar com a independência de *features*, que assume que a presença de uma determinada *feature* não tem relação com as demais.

Para o classificador binário, o algoritmo *Naive Bayes* retornou 13 sugestões das 28 identificadas no conjunto de documentos utilizados no experimento, mesmo tendo apresentado valores baixos em algumas métricas, isso sendo justificado pelo número de sugestões retornadas por esse algoritmo. O classificador implementado com o algoritmo *Naive Bayes* apresentou as métricas com melhores resultados no geral, apesar de que em algumas métricas os melhores valores obtidos tenham sido em outros classificadores. A

avaliação do desempenho desse algoritmo envolveu a análise dos resultados obtidos nas métricas e das sugestões de trabalhos multidisciplinares.

Nesse classificador, alguns resultados obtidos justificam as sugestões retornadas pelo classificador, como por exemplo o valor 1.0 na acurácia para a subcomunidade PGMAT e um valor mais baixo para a mesma métrica para a subcomunidade PEI, que retornou várias sugestões. Na conferência das sugestões, notou-se um número considerável de Falso Positivos, que representam sugestões que foram descartadas, porém boa parte desses Falso Positivos, possuem de certa forma, alguma lógica, já que tratam de trabalhos que possuem a mesma área de conhecimento, como por exemplo a Educação.

Na extração de palavras-chave foram obtidas as sugestões utilizadas no estudo de caso realizado com as bibliotecárias do SIBI responsáveis pelas comunidades dos documentos utilizados nos experimentos. Uma evidência é que para auxiliar o refinamento das sugestões de palavras-chave seja utilizado algoritmo de similaridade de *strings*.

No estudo de caso, foi avaliado que as palavras-chave sugeridas foram adequadas e que a quantidade foi suficiente. Porém, as bibliotecárias chamaram a atenção para a utilização de vocabulário controlado (conhecidos como tesauros) como forma de complementar a lista de palavras-chave.

A anotação semântica de itens de um RI pode favorecer a recuperação das informações, otimizando a busca de itens pesquisados num RI. Esta anotação sendo feita de maneira semiautomática poderá auxiliar o trabalho dos bibliotecários na validação dos metadados de cada item depositado no RI.

As questões de pesquisa presentes na seção 1.3 desta dissertação foram respondidas com a realização dos experimentos e do estudo de caso descritos no Capítulo 5. A implementação do classificador binário mostrou que é possível a identificação de trabalhos multidisciplinares em um repositório acadêmico (Questão P3). A implementação de um classificador multi-hierárquico mostrou que é possível classificar itens não depositados ainda em um RI (Questão P4). O estudo de caso com as bibliotecárias do SIBI respondeu parcialmente à questão P2 pois envolveu somente a validação de palavras-chave das publicações.

É importante salientar que é necessário que haja uma política institucionalizada para estimular os depósitos das publicações no repositório institucional. Isso certamente contribuirá com a publicização dos resultados das pesquisas realizadas e com a possibilidade de formação de parcerias entre pesquisadores que atuem no mesmo tópico de pesquisa.

O esforço para replicar esta solução em outras comunidades implica no povoamento do repositório, ao passo que é necessário uma quantidade de itens depositados para a realização dos experimentos. Sendo também necessário rever os critérios abordados na seção 5.1 para a escolha desses documentos.

6.1 CONTRIBUIÇÕES

Como principais contribuições alcançadas destacam-se:

- Estudo exploratório de métodos de validação e classificação de depósitos das publicações em um RI;

- Implementação de classificadores multi-hierárquico e binário;
- Extração e sugestão de palavras-chave;
- Anotação semântica, de maneira semiautomática, das publicações de um repositório acadêmico;
- Melhoria da recuperação de itens na busca de um repositório acadêmico de forma a publicizar as pesquisas desenvolvidas na instituição;
- Auxilio ao trabalho dos bibliotecários durante a validação dos metadados de cada publicação;
- Método de descrição e enriquecimento semânticos dos itens de um RI.

6.2 LIMITAÇÕES

Durante o projeto foram identificadas algumas limitações, ou seja, foi percebido que houve uma falha metodológica, pois não foi realizada uma entrevista prévia com as bibliotecárias que participariam do estudo de caso. Com isso, a informação de que seria interessante a utilização de vocabulário controlado só foi conhecida durante a realização do estudo de caso.

6.3 TRABALHOS PUBLICADOS

Como resultado dessa dissertação, um artigo foi aprovado e apresentado no *Workshop* de Teses e Dissertações do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: *Webmedia* 2017 [1] e um capítulo de livro foi publicado no livro eletrônico "Comunicação, Mídias e Educação", ISBN 978-85-7247-344-6 e DOI 10.22533/at.ed.446192205 [2]

6.4 TRABALHOS FUTUROS

Como sugestão de trabalhos futuros, pode-se citar:

- Automatizar refinamento das palavras-chave;
- Criar lista com termos frequentemente utilizados nos trabalhos, para filtrar na extração (expressões idiomáticas, advérbios, substantivos (tabela, figura)) - como foi feito com as palavras-chave;
- Ampliar a quantidade de comunidades atendidas pelos experimentos (depende do povoamento do RI);
- Utilização de vocabulário controlado (tesauros) na sugestão de palavras-chave;

[1] Disponível em https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/4831

[2] Disponível em https://www.atenaeditora.com.br/wp-content/uploads/2019/05/e-book-Comunicacao-Midias-e-Educacao-1.pdf

- Refinar as sugestões de palavras-chave com a verificação de similaridade de termos em tesauros (vocabulário controlado);
- Rotular a base com trabalhos multidisciplinares.

REFERÊNCIAS BIBLIOGRÁFICAS

AYODELE, T. O. Types of machine learning algorithms. In: ZHANG, Y. (Ed.). *New Advances in Machine Learning*. Rijeka: IntechOpen, 2010. Disponível em: ⟨https://doi.org/10.5772/9385⟩.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, v. 284, n. 5, p. 1–5, 2001. Disponível em: ⟨http://csis.pace.edu/~marchese/CS835/Lec9/112_SemWeb.pdf⟩.

BREIMAN, L. et al. *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

BROWN, P. F. et al. Class-based n-gram models of natural language. *Computational Linguistics*, v. 18, p. 467–479, 1992.

CAMPOS, R. et al. Yake! collection-independent automatic keyword extractor. In: PASI, G. et al. (Ed.). *Advances in Information Retrieval*. Cham: Springer International Publishing, 2018. p. 806–810. ISBN 978-3-319-76941-7.

CARVALHO, V. H. V. *Análise dos Aspectos de Aceitação e uso do Repositório Institucional da Universidade Federal da Bahia (RI-UFBA) com base no Modelo UTAUT*. 182 p. Dissertação (Mestrado) — Universidade Federal da Bahia, Salvador, 2018.

CORTES, C.; VAPNIK, V. Support-vector networks. *Mach. Learn.*, Kluwer Academic Publishers, Norwell, MA, USA, v. 20, n. 3, p. 273–297, 1995. ISSN 0885-6125. Disponível em: ⟨https://doi.org/10.1023/A:1022627411411⟩.

DIAS, M.; MALHEIROS, M. Extração automática de palavras-chave de textos da língua portuguesa. 2005.

EDMUNDSON, H. P. New methods in automatic extracting. *J. ACM*, ACM, New York, NY, USA, v. 16, n. 2, p. 264–285, 1969. ISSN 0004-5411. Disponível em: ⟨http://doi.acm.org/10.1145/321510.321519⟩.

FARID, H.; KHAN, S.; JAVED, M. Y. Publishing institutional repositories metadata on the semantic web. In: IEEE. *Digital Information Management (ICDIM), 2013 Eighth International Conference on*. 2013. p. 79–84. Disponível em: ⟨http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6694008⟩.

GOMES, M. J.; ROSA, F. G. *Repositórios Institucionais*: democratizando o acesso ao conhecimento. Brasil: EDUFBA, 2010. Disponível em: ⟨https://repositorio.ufba.br/ri/bitstream/ri/616/3/Repositorios\%20institucionais.pdf⟩.

GUTHRIE, L.; WALKER, E. Document classification by machine:theory and practice. In: *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*. [s.n.], 1994. Disponível em: ⟨https://www.aclweb.org/anthology/C94-2172⟩.

HEATH, T.; BIZER, C. *Linked Data: Evolving the Web into a Global Data Space*. 1st. ed. Morgan & Claypool, 2011. ISBN 9781608454303. Disponível em: ⟨http://linkeddatabook.com/⟩.

HILDEN, J. Statistical diagnosis based on conditional independence does not require it. *Computers in Biology and Medicine*, v. 14, p. 429–435, 1984.

HILLMANN, D. Using dublin core. Dublin Core Metadata Initiative, 2008. Disponível em: ⟨http://dublincore.org/documents/usageguide/⟩.

KONSTANTINOU, N. et al. Exposing scholarly information as linked open data: Rdfizing dspace contents. *The Electronic Library*, Emerald Group Publishing Limited, v. 32, n. 6, p. 834–851, 2014.

KONSTANTINOU, N.; SPANOS, D.-E.; MITROU, N. Transient and persistent rdf views over relational databases in the context of digital repositories. In: SPRINGER. *MTSR*. 2013. p. 342–354. Disponível em: ⟨http://rd.springer.com/chapter/10.1007/978-3-319-03437-9_33⟩.

KOUTSOMITROPOULOS, D. A.; SOLOMOU, G. D.; KALOU, A. K. Herding linked data: Semantic search and navigation among scholarly datasets. *International Journal of Semantic Computing*, World Scientific, v. 9, n. 04, p. 459–482, 2015. Disponível em: ⟨http://www.worldscientific.com/doi/pdf/10.1142/S1793351X15500099⟩.

LEITE, F. et al. *RI - Repositórios Institucionais*: Boas práticas para a construção de repositórios institucionais da produção científica. Brasil: IBICT, 2012. Disponível em: ⟨http://livroaberto.ibict.br/handle/1/703⟩.

LEWIS, D. D. Feature selection and feature extract ion for text categorization. In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. [s.n.], 1992. Disponível em: ⟨https://www.aclweb.org/anthology/H92-1041⟩.

LOPER, E.; BIRD, S. Nltk: the natural language toolkit. *CoRR*, 2002.

MANZATO, M. G.; GOULARTE, R. Automatic annotation of tagged content using predefined semantic concepts. In: ACM. *Proceedings of the 18th Brazilian symposium on Multimedia and the web*. [S.l.], 2012. p. 237–244.

MENDONçA, F. et al. Extração automática de termos candidatos às ontologias: um estudo de caso no domínio da hemoterapia. In: . [S.l.: s.n.], 2012.

MIHALCEA, R.; TARAU, P. TextRank: Bringing order into text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411. Disponível em: ⟨https://www.aclweb.org/anthology/W04-3252⟩.

OREN, E. et al. *What are Semantic Annotations?* [S.l.], 2006. Disponível em: ⟨http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf⟩.

OTHERO, G. Lingüística computacional: uma breve introdução. *Letras de Hoje*, v. 41, 2006.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. 2002. p. 79–86. Disponível em: ⟨https://scholar.google.de/scholar.bib?q=info:FPqZqhHpKIAJ:scholar.google.com/&output=citation&hl=de&ct=citation&cd=0⟩.

ROSA, F.; MEIRELLES, R. F.; PALACIOS, M. Repositório institucional da universidade federal da bahia: implantação e acompanhamento. 2011. Disponível em: ⟨https://repositorio.ufba.br/ri/bitstream/ri/1590/1/5603.pdf⟩.

ROSE, S. et al. Automatic keyword extraction from individual documents. In: BERRY, M. W.; KOGAN, J. (Ed.). *Text Mining. Applications and Theory*. John Wiley and Sons, Ltd, 2010. p. 1–20. ISBN 9780470689646. Disponível em: ⟨http://dx.doi.org/10.1002/9780470689646.ch1⟩.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN 0136042597, 9780136042594.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, Pergamon Press, Inc., USA, p. 513–523, 1988. Disponível em: ⟨https://doi.org/10.1016/0306-4573(88)90021-0⟩.

SANCHES, L. M. P. *Anotação Semântica Semiautomática de Objetos de Aprendizagem em Português*. 186 p. Dissertação (Mestrado) — Universidade Federal da Bahia, Salvador, 2018.

SANTOS, D. S. *O Repositório Institucional da Universidade Federal da Bahia: verificação da adesão dos Programas de Pós-Graduação da área I*. 114 p. Dissertação (Mestrado) — Universidade Federal da Bahia, Salvador, 2019.

SAYÃO, L. et al. *Implantação e gestão de repositórios institucionais*: politicas, memória, livre acesso e preservação. Brasil: EDUFBA, 2009. Disponível em: ⟨https://repositorio.ufba.br/ri/bitstream/ufba/473/3/implantacao_repositorio_web.pdf⟩.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys*, v. 34, p. 1–47, 2001.

SHINTAKU, M.; MEIRELLES, R. F. Manual do dspace: administração de repositórios. EDUFBA, 2010. Disponível em: ⟨https://repositorio.ufba.br/ri/bitstream/ri/769/1/Manual\%20do\%20Dspace(2).pdf⟩.

SIDDIQI, S.; SHARAN, A. Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*, v. 109, p. 18–23, 2015.

SILVA, W. D. da. *Anotação Semântica Automática do Currículo Lattes Utilizando Linked Open Data*. 132 p. Dissertação (Mestrado) — Universidade FUMEC, Belo Horizonte, 2016.

TANSLEY, R. et al. The dspace institutional digital repository system: current functionality. In: IEEE COMPUTER SOCIETY. *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. [S.l.], 2003. p. 87–97.

VENS, C. et al. Decision trees for hierarchical multi-label classification. *Machine Learning*, v. 73, n. 2, p. 185, 2008. ISSN 1573-0565. Disponível em: ⟨https://doi.org/10.1007/s10994-008-5077-3⟩.

VISA, S. et al. Confusion matrix-based feature selection. In: . [S.l.: s.n.], 2011. v. 710, p. 120–127.

Apêndice

A

CONFIGURAÇÕES DO AMBIENTE

Neste apêndice são descritas as configurações do RI UFBA e do ambiente utilizado nos experimentos.

A.1 INSTALAÇÃO RI UFBA

A versão do *DSpace* instalada atualmente no RI UFBA é a 3.2, hospedada em um servidor *Linux Debian*, com o SGBD *PostgreSQL* 8.3, *Apache 2.2.16*, *Tomcat* 7.0.47, Java versão JDK (Java SE *Development Kit*) 1.5, com pacotes *Java-Ant* e *Maven* 2 instalados. Essa atualização inclui tanto a correção dos metadados e inconsistências da instalação atual quanto a mudança de *template*.

O projeto de atualização do *DSpace* do RI UFBA para a versão 5.7 encontra-se em andamento no momento.

A.2 CONFIGURAÇÃO DO AMBIENTE PARA OS EXPERIMENTOS

Para os experimentos, foram utilizadas as seguintes configurações no ambiente:

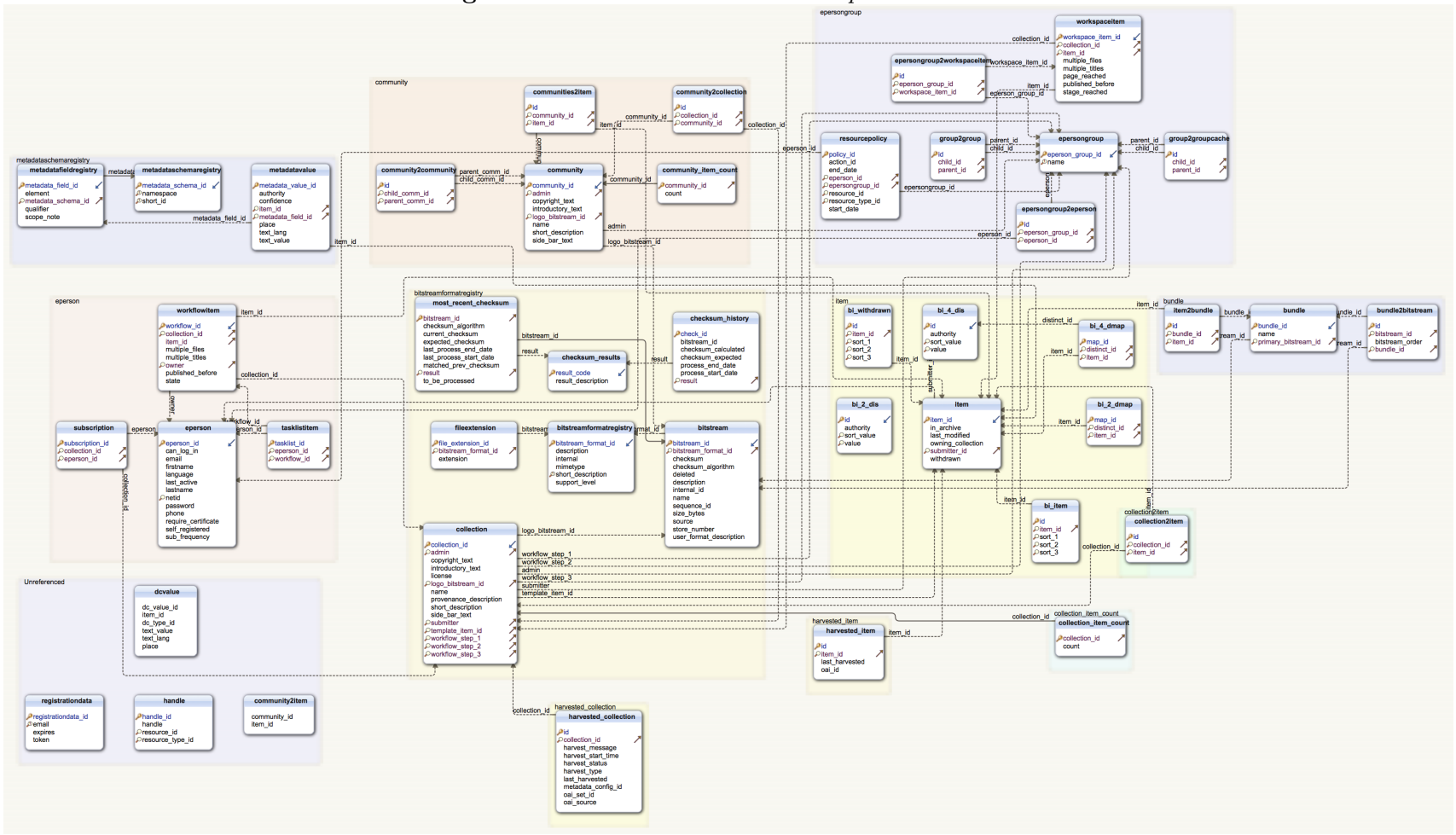
- **Ambiente *Python***
  Para os experimentos foi instalada a versão 3.6 do *Python*.[1] e para execução dos comandos e visualização dos resultados passo a passo, foi utilizada a ferramenta *Jupyter Notebook*.[2] Escolheu-se utilizar a linguagem *Python* durante os experimentos por se tratar de uma linguagem que já possui diversas bibliotecas implementadas e que auxiliam na extração, manipulação, análise de informações textuais, bem como uma biblioteca especifica para aprendizado de máquina, a *scikit-learn*, além de apresentar baixa curva de aprendizado e pela vasta documentação existente sobre a linguagem.

[1] *Download* em https://www.python.org/
[2] *Download* em http://jupyter.org/install.html

- **Criação da base de dados**
  Uma base de dados foi montada contendo o *dump* da base de dados na qual estão armazenadas as informações do RI UFBA. Este *dump* foi obtido através de uma solicitação formal junto à STI-UFBA, que faz parte do Comitê Gestor do RI UFBA. Para a realização dos experimentos foi efetuado o *download* de 153 publicações de 3 comunidades do RI UFBA, utilizando os critérios da seção 5.2.7.

- ***Dump* do banco de dados**
  A instalação do *DSpace* no RI UFBA utiliza o SGBD *PostgreSQL* na versão 8.3. De posse do *dump* do banco de dados obtido, foi montado o ambiente de banco de dados contendo as mesmas especificações do banco original. Para isso, foi instalado o gerenciador *pgAdmin* que fornece um suporte para manipulação do SGBD *Postgres*. O passo seguinte foi estudar o modelo de dados do *DSpace*, apresentado na Figura A.1, para conhecer como as tabelas do banco são relacionadas entre si. Posteriormente, foram executadas consultas para obter as informações de comunidades e coleções de cada publicação.

Figura A.1 Modelo de Dados do *DSpace* versão 3.x.



Fonte: Documentação *DSpace*.

Apêndice B

Neste apêndice são listadas as consultas SQL que foram executadas no banco de dados Postgresql, que contém as informações dos itens depositados no RI UFBA, com a finalidade de gerar os vetores de metadados

VETORES DE METADADOS

O conjunto de publicações utilizado nos experimentos foi rotulado com a extração dos metadados da base de dados do RI UFBA. Para isso, foi necessário executar 2 consultas *Structured Query Languge* (SQL) embutidas no código *Python* para retornar a informação de comunidade e subcomunidade de cada documento e gravá-las em uma matriz. Para obter as comunidades dos documentos extraídos foi utilizada a sub-consulta apresentada no Código B.1:

```
01 | select name from community where community_id in (select
       community_id from communities2item, item2bundle,
       bundle2bitstream where communities2item.item_id = item2bundle.
       item_id and item2bundle.bundle_id = bundle2bitstream.bundle_id
        and communities2item.item_id = [id do item]));
```

Código B.1 Extração de comunidades

Para obter as subcomunidades dos documentos extraídos foi utilizada a sub-consulta apresentada no Código B.2:

```
01 | select * from community2community and select * from
       communities2item and select * from item2bundle and select *
       from bundle2bitstream);
```

Código B.2 Extração de subcomunidades

Para extrair essas informações de cada documento, foi necessário passar como parâmetro o nome de cada arquivo e gravar as informações em vetores, um para cada metadado.

Na qual:

1ª coluna	2ª coluna	3ª coluna
nome da comunidade	nome da subcomunidade	nome do arquivo

De posse destes vetores, as publicações foram rotuladas e estas informações foram convertidas em formato numérico, já que os algoritmos de aprendizado de máquina só trabalham com números e passadas para os classificadores implementados.

Na tabela B.1, é possível visualizar os valores obtidos para o classificador multi-hierárquico, obtidas com a função *LabelEncoder*:

Tabela B.1 Valores Numéricos - Classificador Multi-Hierárquico

Comunidade	Valor numérico	Subcomunidade	Valor numérico
POLI	0	MAASA	0
		PEI	1
		PPEQ	2
FACED	1	DIFUSÃO	0
		PPGE	1
IME	2	PGCOMP	0
		PGMAT	1

Fonte: elaborado pela autora (2019)

Na tabela B.2, foram listados os valores numéricos de cada subcomunidade, para o classificador binário, obtidos com a utilização da função *LabelBinarizer*:

Tabela B.2 Valores Numéricos - Classificador Binário

Subcomunidade	Valor numérico
MAASA	0
PGCOMP	1
DIFUSÃO	2
PPGE	3
PEI	4
PPEQ	5
PGMAT	6

Fonte: elaborado pela autora (2019)

No caso do classificador binário, após a conversão, é criada uma matriz, na qual cada linha corresponde a um documento e cada coluna corresponde a uma subcomunidade, conforme a Tabela B.3:

Tabela B.3 Vetores binários das subcomunidades

	Original	MAASA	PGCOMP	DIFUSAO	PPGE	PEI	PPEQ	PGMAT
i0	MAASA	1	0	0	0	0	0	0
i1	PPEQ	0	0	0	0	0	1	0
i2	MAASA	1	0	0	0	0	0	0
i3	PGCOMP	0	1	0	0	0	0	0
i4	PGCOMP	0	1	0	0	0	0	0
i5	PPGE	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...
i193	MAASA	1	0	0	0	0	0	0
i194	PGCOMP	0	1	0	0	0	0	0
i195	PEI	0	0	0	0	1	0	0

Fonte: elaborado pela autora (2019).

Para efeito de ilustração, a coluna rotulada como Original contém a informação da subcomunidade a qual o documento pertence. Nesse caso, será apresentado o valor 1 na coluna da subcomunidade correspondente e valor 0 nas demais colunas,

Como no vetor original, existem 7 subcomunidades e foram gerados 7 novas colunas, sendo que ela será preenchida com o valor 1 - caso o documento pertença à subcomunidade representada por esta coluna e 0 - caso não pertença.

Apêndice C

Este apêndice contém o Roteiro para condução do Estudo de Caso com as bibliotecárias do SIBI

ROTEIRO DO ESTUDO DE CASO

C.1 CONTEXTUALIZAÇÃO

Neste momento será explicado o projeto de pesquisa, quais experimentos foram realizados e os resultados obtidos, com destaque para a extração de palavras-chaves por ter ligação com o estudo de caso. Também será falado das dificuldades que são relatadas atualmente, buscando encontrar as publicações através da busca no RI UFBA (inconsistência no sistema, descrição dos metadados, etc).

C.2 OBJETIVO DO ESTUDO DE CASO

Neste momento, será explicado qual o objetivo do estudo de caso e quais as contribuições que se espera alcançar ao final. Também será abordado que sugerir palavras-chaves mais representativas de um texto auxilia na validação dos metadados, enriquecendo-os semanticamente e favorecendo que as publicações sejam retornadas nas buscas.

C.3 DESCRIÇÃO DOS EXPERIMENTOS

Tabela C.1 Experimentos Realizados

Experimento	**Objetivo**
Classificação textual	Classificação textual dos documentos utilizando os algoritmos de classificação: SVM, *Naive Bayes* e *Decision Tree*.
Extração de Palavras-Chaves	Extrair as palavras-chaves de cada texto para sugerir aos bibliotecárias durante a validação dos metadados.

Fonte: Autoria Própria (2019).

C.4 DESCRIÇÃO DO EXPERIMENTO EXTRAÇÃO DE PALAVRAS-CHAVES

Este estudo de caso levará em consideração somente a validação do metadado palavras-chaves, representada pelo Dublin Core através do atributo *dc: subject*. Também mostrará como foi feita a verificação manual das palavras-chaves sugeridas nos documentos da subcomunidade PGCOMP, com a leitura do título, das palavras-chaves preenchidas pelo pesquisador e do resumo do trabalho. Esta verificação também servirá para avaliar os métodos de extração de palavras-chaves testados anteriormente.

C.5 DESCRIÇÃO DAS ATIVIDADES

O estudo de caso será conduzido com a participação das bibliotecárias responsáveis pela validação dos metadados nas comunidades POLI, IME e FACED no RI UFBA. Como estas bibliotecárias estão lotadas fisicamente em *campi* diferentes, o estudo de caso será realizado em 3 momentos diferentes, um para cada comunidade, sempre com a participação da(s) bibliotecária(s) responsáveis pela validação de metadados destas comunidades.

Tabela C.2 Estudo de Caso Realizado

Estudo de Caso	**Objetivo**
Simulação com o pessoal do SIBI	Bibliotecárias devem validar as sugestões de palavras-chaves de cada documento.

Fonte: Autoria Própria (2019).

Será explicado às bibliotecárias como a atividade será conduzida. Cada bibliotecária terá 1:30 min para validar as sugestões de palavras-chaves refinadas depois da extração. Cada bibliotecária receberá em torno de 8 documentos. Sendo assim, a validação dos documentos deverá seguir a seguinte distribuição: IME - 16 documentos, POLI - 17 documentos e FACED - 17 documentos, totalizando 50 documentos. Seguindo as recomendações do Núcleo Tecnológico do SIBI, esta validação deve envolver a leitura do título do trabalho, da ficha catalográfica e do resumo do trabalho e validação ou não das palavras-chaves sugeridas. Ao finalizar esta validação, será aplicado um questionário (algumas questões utilizam a Escala *Likert* para avaliação) no qual as bibliotecárias deverão responder à perguntas referentes à execução e possível contribuição do estudo de caso, com espaço para relatar a experiência,as dificuldades encontradas, bem como avaliar os benefícios da sugestão de palavras-chaves durante a validação dos metadados.

O tempo alocado para a atividade será de 2h, dividido entre apresentação da atividade, a validação das palavras-chaves sugeridas e o preenchimento do questionário.

Tempo da atividade = contextualização do projeto + realização do estudo de caso
Estudo de caso = validação das palavras-chaves sugeridas + preenchimento do questionário
Tempo estimado para a atividade: 2h para cada encontro

C.6 VALIDAÇÃO DE PALAVRAS-CHAVES

Vianna, Renata de Moura Issa, 1985
Dualidade no Modelo KMP e a Lei de Fourier / Renata de Moura Issa Vianna. - 2015.
72 f. : il.
Orientador: Prof. Dr. Tertuliano Franco Santos Franco.
Dissertação (mestrado) – Universidade Federal da Bahia, Instituto de Matemática, Programa de Pós-graduação em Matemática, 2015.
1. Modelos Matemáticos. 2. Osciladores Harmônicos. 3. Calor - Condução. 4. Lei de Fourier. I. Franco, Tertuliano Franco Santos. II. Universidade Federal da Bahia, Instituto de Matemática. III. Título.
CDD : 510
CDU : 51

RESUMO DO TRABALHO

O intuito desta dissertação é estudar o modelo KMP. Este é um clássico modelo de interação constituído por uma cadeia de osciladores harmônicos unidimensionais desacoplados que trocam energia por meio de um processo estocástico. Cada elo tem um relógio de Poisson. Sempre que o relógio toca, dois osciladores vizinhos redistribuem energia de maneira uniforme. Além disso, o sistema está em contato com reservatórios nas extremidades, a diferentes temperaturas. Neste trabalho, apresentamos o estudo deste modelo e mostramos a validade da Lei de Fourier.

Palavras-chave: modelo KMP; osciladores harmônicos; Lei de Fourier.

SUGESTÕES DE PALAVRAS-CHAVE

variável aleatória ( )
espaço métrico ( )
processo estocástico ( )
espaço vetorial ( )
classe determinante ( )
osciladores harmônicos ( )
espaço amostral ( )

Instituto de Matemática e Estatística
Departamento de Ciência da Computação
Programa de Pós-Graduação em Ciência da Computação

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Prezada Senhora,

Esta pesquisa é sobre **Anotações Semânticas em Repositórios Acadêmicos: um estudo de caso com o RI UFBA** e está sendo desenvolvida pela pesquisadora **Aline Meira Rocha**, mestranda do Programa de Pós-Graduação em Ciência da Computação (PGCOMP) da Universidade Federal da Bahia (UFBA), sob a orientação da **Profª Dra. Lais do Nascimento Salvado**r e co-orientação do **Prof. Dr. Marlo Vieira dos Santos e Souza**.

O objetivo geral da pesquisa **é identificar como anotar semanticamente os itens de um repositório acadêmico no DSpace de maneira semi-automática, além de realizar um estudo de caso com o RI UFBA**. A finalidade deste trabalho **é auxiliar o trabalho dos bibliotecários durante a validação dos depósitos das publicações e contribuir para a recuperação dessas publicações com a otimização da busca em um RI**.

Solicitamos a sua colaboração para realização de **um estudo de caso** com tempo médio de **02 horas**, como também sua autorização para apresentar os resultados deste estudo em publicações em eventos da área da computação, bem como publicar em revista científica nacional e/ou internacional. Por ocasião da publicação dos resultados, seu nome será mantido em sigilo absoluto. Informamos que os dados coletados durante a realização do estudo de caso serão inseridos na redação final da dissertação de mestrado. Esses procedimentos não oferecem risco algum a integridade física ou moral do participante, bem como despesas, prejuízos ou benefícios diretos.

Esclarecemos que sua participação no estudo é voluntária e, portanto, a senhora não é obrigada a fornecer as informações e/ou colaborar com as atividades solicitadas pela pesquisadora. Caso decida não participar do estudo, ou resolver a qualquer momento desistir do mesmo, não sofrerá nenhum dano, nem haverá modificação na assistência que vem recebendo na Instituição (se for o caso). Os pesquisadores estarão a sua disposição para qualquer esclarecimento que considere necessário em qualquer etapa da pesquisa.

_____

Assinatura da pesquisadora responsável
Aline Meira Rocha

Considerando, que fui informada dos objetivos e da relevância do estudo proposto, de como será minha participação, dos procedimentos e riscos decorrentes deste estudo, declaro o meu consentimento em participar da pesquisa, como também concordo que os dados obtidos na investigação sejam utilizados para fins científicos (divulgação em eventos e publicações). Estou ciente que receberei uma via desse documento.

Salvador, ____ de _____ de 2019

_____

Assinatura do participante

Contato com a Pesquisadora Responsável:

Caso necessite de maiores informações sobre a presente pesquisa, favor encaminhar e-mail para aline.meira@ufba.br para entrar em contato com a pesquisadora Aline Meira Rocha.

Questionário do Estudo de Caso

Esse questionário tem o objetivo de coletar informações sobre o Estudo de Caso realizado com as bibliotecárias do SIBI durante a validação das sugestões de palavras-chaves nas publicações do RI UFBA.

Dados Pessoais

1. **Unidade de Lotação**: _____
2. **Tempo de Serviço como Bibliotecária**: _____ anos.
3. **Sou responsável pela seguinte coleção**: _____

Estudo de Caso

Por favor, responda as seguintes questões sobre a realização do Estudo de Caso.

4. **Já tinha ouvido falar sobre Classificação Textual?** □ Sim □ Não
5. **Já tinha ouvido falar sobre *Machine Learning*?** □ Sim □ Não
6. **Já tinha ouvido falar sobre métodos de Extração de Palavras-Chaves?** □ Sim □ Não
7. **Já tinha ouvido falar sobre Anotações Semânticas?** □ Sim □ Não
8. **As palavras-chaves sugeridas foram adequadas?**
   - □ Discordo Totalmente.
   - □ Discordo Parcialmente
   - □ Indiferente
   - □ Concordo Parcialmente
   - □ Concordo Totalmente
9. **A quantidade de sugestão de palavras-chaves foi suficiente?**
   - □ Discordo Totalmente.
   - □ Discordo Parcialmente
   - □ Indiferente
   - □ Concordo Parcialmente
   - □ Concordo Totalmente
10. **A quantidade de documentos utilizados na atividade foi adequada?**
    - □ Discordo Totalmente.
    - □ Discordo Parcialmente
    - □ Indiferente
    - □ Concordo Parcialmente
    - □ Concordo Totalmente
11. **O tempo para a realização da atividade foi adequada/suficiente?**
    - □ Discordo Totalmente.
    - □ Discordo Parcialmente
    - □ Indiferente
    - □ Concordo Parcialmente
    - □ Concordo Totalmente

Considerações finais

12. Houve algum problema durante a realização do Estudo de Caso? Caso afirmativo, relate a situação.

13. Acredita que os experimentos podem contribuir para a melhoria do seu trabalho na validação dos metadados? Justifique sua resposta.

14. Criticas / Sugestões / Observações:

Apêndice

D

Este apêndice apresenta a lista dos trabalhos multidisciplinares utilizados nos experimentos

TRABALHOS MULTIDISCIPLINARES

D.1 TRABALHOS MULTIDISCIPLINARES

Para a classificação textual binária, considerando as 153 publicações distribuídas conforme a Tabela 5.3, depois de uma verificação manual (leitura de título, resumo e sumário), foram identificados 28 trabalhos multidisciplinares, apresentados nas Tabelas D.1 e D.2 a seguir:

Tabela D.1 Trabalhos Multidisciplinares - Parte 1

TÍTULO DA PUBLICAÇÃO	SUBCOMUNIDADE	SUGESTÃO
DIGESTÃO ANAERÓBIA DA BIOMASSA RESIDUAL DE MICROALGAS PÓSEXTRAÇÃO DE LIPÍDIOS	MAASA	PPEQ
ANÁLISE DA EVIDENCIAÇÃO AMBIENTAL DOS RELATÓRIOS DE SUSTENTABILIDADE DAS INDÚSTRIAS CERVEJEIRAS BRASILEIRAS	MAASA	PEI
ANÁLISE DA ESTRUTURA QUALI-QUANTITATIVA ZOOBENTÔNICA DO MESOLITORAL DA BAÍA DE TODOS OS SANTOS (BA) E RELAÇÕES COM A CONTAMINAÇÃO QUÍMICA DOS SEDIMENTOS	MAASA	PPEQ
Sistemas de Transições Modais de Kripke para Representação de Comportamento Parcial no Desenvolvimento Incremental e Iterativo de Software	PGCOMP	PGMAT
Uma Solução para o Refinamento de Modelos KMTS Baseado em Verificação de Modelos com Jogos	PGCOMP	PGMAT
CARACTERIZAÇÃO DE MODELOS CONCEITUAIS UTILIZANDO ONTOLOGIAS DE DOMÍNIO: APLICAÇÃO DA ONTOLOGIA IMS LD NA CONSTRUÇÃO DE MODELOS CONCEITUAIS PARA E-LEARNING	PGCOMP	PPGE
GOVERNANÇA NAS COMPRAS PÚBLICAS SUSTENTÁVEIS: UM MODELO PARA OS INSTITUTOS FEDERAIS DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA BASEADO NA ANÁLISE DE REDES SOCIAIS	DIFUSÃO	PGCOMP
RPG DIGITAL E SEGURANÇA PÚBLICA: UMA PROPOSTA DE APLICAÇÃO PEDAGÓGICA PARA INSTRUÇÃO POLICIAL MILITAR	DIFUSÃO	PPGE
AS RELAÇÕES DE SOCIABILIDADE E AS (RE)INTERPRETAÇÕES DE GÊNERO E MASCULINIDADES DE JOVENS NO CONTEXTO ESCOLAR	DIFUSÃO	PPGE
ANÁLISE DE DOMÍNIO NA AQUISIÇÃO DE CONHECIMENTOS: ONTOLOGIAS PARA SISTEMAS COMPUTACIONAIS	DIFUSÃO	PGCOMP
O PROCESSO DE EXPANSÃO DA REDE FEDERAL DE EDUCAÇÃO TECNOLÓGICA. UM ESTUDO DE CASO NA BAHIA	DIFUSÃO	PPGE

Fonte: elaborado pela autora (2019)

Tabela D.2 Trabalhos Multidisciplinares - Parte 2

TÍTULO DA PUBLICAÇÃO	SUBCOMUNIDADE	SUGESTÃO
CONTRADIÇÕES E POSSIBILIDADES DE SUPERAÇÃO NO TRABALHO PEDAGÓGICO PARA A PRÁTICA DA EDUCAÇÃO AMBIENTAL NOS ANOS INICIAIS DO ENSINO FUNDAMENTAL: UM ESTUDO DE CASO	PPGE	MAASA
DESIGN COGNITIVO COLABORATIVO PARA AMBIENTES VIRTUAIS: O CASO DO PORTAL TBC CABULA	DIFUSÃO	PGCOMP
HIPERTEXTO E HIPERLEITURA: CONTRIBUIÇÕES PARA UMA TEORIA DO HIPERTEXTO	PPGE	PGCOMP
PLATAFORMA COMPUTACIONAL WEB PARA CALIBRAÇÃO DE SISTEMAS DE MEDIÇÃO	PEI	PGCOMP
METODOLOGIA PARA DEFINIR UM SISTEMA DE INDICADORES DE DESEMPENHO SOCIOAMBIENTAL: O ESTUDO DE CASO COELBA	PEI	MAASA
MODELO PARA A GESTÃO DOS IMPACTOS SOCIOAMBIENTAIS NO SETOR DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA: O ESTUDO DE CASO COELBA	PEI	MAASA
SISTEMA DE MENSURAÇÃO DE DESEMPENHO EM INOVAÇÃO PARA UNIVERSIDADES PÚBLICAS NO BRASIL	PEI	DIFUSÃO
ANÁLISE DE COMPONENTES PRINCIPAIS INTEGRADA A REDES NEURAIS ARTIFICIAIS PARA PREDIÇÃO DE MATÉRIA ORGÂNICA	PEI	PGCOMP
APLICAÇÃO DA FILTRAÇÃO INTERMITENTE EM LEITO DE AREIA E DE ESCÓRIA DA METALURGIA DO COBRE NO TRATAMENTO DE ESGOTOS COM ÊNFASE EM REÚSO	PEI	MAASA
AVALIAÇÃO DA QUALIDADE DO ENSINO DE ENGENHARIA DE PRODUÇÃO NO BRASIL A PARTIR DOS INDICADORES DO SINAES	PEI	PPGE
ESTIMAÇÃO DE PARÂMETROS, INFERÊNCIA E CONTROLE DE PROPRIEDADES DE QUALIDADE DE UM PROCESSO DE COPOLIMERAZAÇÃO DE ETENO	PEI	PPEQ
DESENVOLVIMENTO DE NANOPARTÍCULAS POLIMÉRICAS CONTENDO ÓLEO ESSENCIAL DE CITRONELA (CYMBOPOGON WINTERIANUS)	PEI	PPEQ
INVESTIGAÇÃO SOBRE A REMOÇÃO BIOLÓGICA SIMULTÂNEA DE AMÔNIO E NITRITO UTILIZANDO BIOMASSA EM SUSPENSÃO ORIUNDA DE SISTEMA DE LODO ATIVADO	PEI	PPEQ
ESTRATÉGIAS DE CONTROLE APLICADAS A REATORES DE POLIMERIZAÇÃO DE ETENO EM SOLUÇÃO	PEI	PPEQ
A INFLUÊNCIA DOS FATORES HUMANOS NA UTILIZAÇÃO DOS PROCEDIMENTOS OPERACIONAIS EM UMA EMPRESA PETROQUÍMICA	PEI	PPEQ
DESENVOLVIMENTO DE CATALISADORES PARA ABATIMENTO DE FENÓIS EM EFLUENTES INDUSTRIAIS	PPEQ	PEI
UTILIZAÇÃO DA FIBRA DE SISAL TRATADA COM LÍQUIDO IÔNICO COMO SORVENTE DE ÓLEOS EM ÁGUA	PPEQ	MAASA

Fonte: elaborado pela autora (2019)