UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CHRISTIAN SCHMITZ

ACERPI: Uma abordagem para coleta de documentos, extração de informação e resolução de entidades em Portarias institucionais

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Profª. Drª. Renata Galante
Co-orientador: Prof. Dr. Edimar Manica

Porto Alegre
2020

AGRADECIMENTOS

Gostaria de agradecer à minha família, meu pai Francisco Xavier Schmitz, minha mãe Sandra Winter Schmitz e minha irmã Julia Winter Schmitz pelo incentivo e suporte durante toda a minha jornada acadêmica, pela empatia nas situações onde tive que abrir mão de momentos familiares para dedicar-me aos estudos e pelo apoio emocional necessário durante essa jornada.

Agradeço também à minha namorada Victória Alves pelo suporte que sempre me foi dado, acolhimento nos momentos de angústia, de decisões difíceis e parceria na realização dos meus sonhos.

Agradeço à todos os meus amigos que, de algum modo, participaram dessa incrível etapa da minha vida, da qual levarei lembranças para sempre. Em especial aos amigos Augusto Boranga, Lúcio Franco e Igor Pires. Sem dúvidas, sem vocês passar por esse desafio seria muito mais difícil.

Agradeço também aos professores do Instituto de Informática da Universidade Federal do Rio Grande do Sul que, mesmo diante das maiores dificuldades que sabemos existirem na vida acadêmica no Brasil, seguem fortes no ensino e pesquisa de excelência. Agradecimentos especiais à professora Renata de Matos Galante, orientadora deste trabalho e professora mais preocupada com os alunos que tive o privilégio de ter tido, bem como o professor Edimar Manica, co-orientador deste trabalho.

Muito obrigado.

RESUMO

Portarias são documentos emitidos por órgãos institucionais federais que contém, dentre outras, informações a respeito de servidores de Instituições. Esses documentos estão acessíveis através de repositórios públicos de cada instituição que, em geral, não permitem nenhum tipo de filtro ou busca avançada sobre o conteúdo dos documentos. Através da abordagem ACERPI (**Abordagem** para **Coleta** de documentos, **Extração** de informação e **Resolução** de entidades em **Portarias Institucionais**) desenvolvida neste trabalho, é realizada a criação de um banco de dados orientado a documentos (MongoDB) para consultas avançadas a respeito dos documentos relacionados a um servidor de uma Instituição, bem como quais servidores são referenciados em um dado documento publicado. Para isso, são usadas técnicas de descoberta, obtenção, conversão e estruturação de arquivos, extração de informação e resolução de entidades (servidores, no contexto deste trabalho). Experimentos com dados reais da Universidade Federal do Rio Grande do Sul e do Instituto Federal do Rio Grande do Sul, *Campus* Ibirubá, demonstram e explicam os principais desafios encontrados ao aplicar a abordagem em duas fontes de dados. Por fim, são mencionados pontos de melhoria e continuidade de desenvolvimento da abordagem, considerados possíveis trabalhos futuros.

Palavras-chave: Coleta de documentos. extração de informação. resolução de entidades.

ACERPI: An approach for document collection, information extraction and entity resolution in federal institutions' documents from Brazil

ABSTRACT

Portarias are documents issued by federal institutional organizations that contain, among others, information regarding the staff of institutions. These documents are accessible through public repositories from each institution that, in general, do not allow any type of filter or advanced search on documents' contents. Through the ACERPI approach developed in this work, the creation of a document oriented database (MongoDB) is carried out for advanced queries regarding the documents related to an institution's employee, as well as which employees are referenced in a given published document. In order to do this, techniques are used to discover, obtain, convert and structure documents, extract information and link entities (employees, in the context of this work). Experiments with data from the Federal University of Rio Grande do Sul and the Federal Institute of Rio Grande do Sul, *Campus* Ibirubá, demonstrate and explain the main challenges encountered when applying the approach to two data sources. Finally, improvement points and future work are discussed.

Keywords: documents retrieval, information extraction, entity resolution.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface* - Interface de Programação de Aplicativos
CAPTCHA	*Completely Automated Public Turing test to tell Computers and Humans Apart* - Teste de Turing Público Completamente Automatizado para Diferenciar Computadores de Humanos
ER	*Entity Resolution* - Resolução de Entidades
HTML	*Hypertext Markup Language* - Linguagem de Marcação de Hipertexto
IA	Inteligência Artificial
NER	*Named Entity Recognition* - Reconhecimento de Entidades Nomeadas
PDF	*Portable Document Format* - Formato Portátil de Documento
PNL	Processamento de Linguagem Natural
SGDB	Sistema de Gerenciamento de Banco de Dados
URL	*Uniform Resource Locator* - Localizador Uniforme de Recursos
XML	*Extensible Markup Language* - Linguagem de Marcação Extensível

SUMÁRIO

1 INTRODUÇÃO

Instituições federais realizam a disseminação de alterações em cargos de seus servidores, afastamentos, substituições de função, entre outros, através da publicação de documentos chamados Portarias. Portarias são documentos oficiais emitidos por órgãos das instituições que concretizam a instauração das resoluções nelas contidas. Hoje, devido a Lei nº 12.527 (BRASIL, 2011) que formaliza a divulgação de informações que possam ser de interesse público produzidas por instituições federais, a publicação das Portarias pelas instituições ocorre publicamente, permitindo que qualquer pessoa tenha acesso a essas informações.

Apesar disso, o acesso a essas informações é, muitas vezes, dado através da simples disponibilização dos arquivos PDF dos documentos em repositórios individuais de cada instituição, ou mesmo de diferentes *Campus* dentro das instituições. Esses repositórios de documentos, com pouco ou nenhum filtro para a realização de pesquisas avançadas sobre o conteúdo dos registros, não permitem a busca rápida (quiçá praticável) a respeito de servidores específicos ou tipos de documentos. Um exemplo de repositório pode ser visto na Figura 1.1.

Figura 1.1 – Repositório de documentos do Instituto Federal do Rio Grande do Sul, *Campus* Ibirubá [1]

O principal objetivo deste trabalho é a criação de uma abordagem genérica de

[1] Disponível em <https://ifrs.edu.br/ibiruba/documentosoficiais/boletim-de-servico/>. Último acesso em 21/11/2020.

descoberta, obtenção, conversão e estruturação de arquivos, extração de informação e resolução de entidades de Portarias institucionais que permita ao usuário realizar pesquisas em um banco de dados fazendo uso de informações extraídas dos documentos, como o nome dos servidores mencionados, suas matrículas identificadoras (SIAPE), a data de publicação das Portarias e seus números de identificação.

Isso é feito através da coleta dos documentos de seus repositórios, conversão e estruturação dos mesmos para um formato padrão XML, extração de informações relevantes de cada uma das Portarias e resolução das entidades encontradas nas Portarias. O resultado final é um banco de dados não relacional, flexível e com estruturas simplificadas para pesquisas de Portarias e servidores. Experimentos com dados reais da Universidade Federal do Rio Grande do Sul (UFRGS) e Instituto Federal do Rio Grande do Sul (IFRS), *Campus* Ibirubá, demonstram a eficácia da abordagem com precisão de 98% na coleta e estruturação dos arquivos da UFRGS, medida F1 de 89% no reconhecimento dos nomes dos servidores do IFRS, *Campus* Ibirubá e medida F1 de 90% na eficácia da técnica de resolução de entidades.

O restante do texto está organizado da seguine forma: o Capítulo 2, Tecnologias Utilizadas, aborda as técnicas e ferramentas já estabelecidas que são utilizadas na abordagem. O Capítulo 3, Trabalhos Relacionados, descreve de maneira breve outros projetos desenvolvidos com objetivos similares, mesmo que em diferentes domínios. O Capítulo 4, Proposta, aprofunda integralmente na extração e processamento dos dados conforme proposto pela abordagem ACERPI. O Capítulo 5, Avaliação Experimental, documenta os experimentos e casos de falha envolvidos na aplicação da abordagem em duas bases de dados de instituições diferentes, são elas a Universidade Federal do Rio Grande do Sul e o Instituto Federal do Rio Grande do Sul, *Campus* Ibirubá. Por fim, o Capítulo 6, Conclusão, revisa as contribuições do trabalho realizado e apresenta as possibilidades de trabalhos futuros.

2 TECNOLOGIAS UTILIZADAS

Neste capítulo são apresentadas as tecnologias (técnicas e ferramentas) utilizadas durante o desenvolvimento do trabalho. O uso das mesmas não só acelerou o processo de criação, como também deu mais confiança às decisões tomadas e caminhos seguidos. São mostradas, respectivamente, as técnicas envolvidas na abordagem proposta e as ferramentas que auxiliaram ou foram integradas à solução final.

2.1 *Web Scraping*

Web Scraping é o processo de recuperar dados de um *website* (ScrapingHub, 2020). Esta recuperação pode ser tanto manual quanto automática, e sua complexidade varia conforme a dificuldade em acessar, analisar e extrair os dados em questão.

Dentre as diversas técnicas de *Web Scraping*, destacam-se:

- Navegação e *download* manuais. Nessa abordagem, o próprio usuário realiza os acessos, identifica visualmente onde encontram-se os documentos de interesse e realiza, um a um ou em lotes, o *download* e gravação dos arquivos. Essa abordagem pode ser recomendada quando o volume de dados é pequeno e a forma de acesso é limitada.
- Caminhamento da estrutura de arquivos da página e *download* automatizado. Essa técnica é bastante comum na extração de dados de páginas web. São criados pequenos trechos de código (*scripts*) que analisam a estrutura da página em questão, identificando pontos comuns que caracterizam as páginas de interesse. Após a identificação das páginas de interesse é, então, realizada a coleta do conteúdo também de maneira automatizada.
- Inferência de um padrão de navegação (Palmieri Lage et al., 2004) de endereços relevantes e *download* automatizado do conteúdo completo das páginas. Esse procedimento pode ser utilizado quando, por exemplo, todo o conteúdo da página precisa ser armazenado.

Neste trabalho, são usadas as duas últimas abordagens devido ao volume de arquivos a serem recuperados.

2.2 *Transfer Learning*

Transfer Learning é a melhoria da aprendizagem em uma nova tarefa por meio da transferência de conhecimento de uma tarefa relacionada que já foi aprendida (OLIVAS et al., 2009). Um dos objetivos do uso de transferência de aprendizado entre modelos de aprendizado de máquina é a diminuição do tempo de anotação e treinamento necessários para o aprendizado de uma dada tarefa por um dado modelo. Além disso, em alguns casos, é possível notar uma melhora na eficácia do modelo final quando o treinamento acontece em um modelo pré-treinado (em detrimento de um modelo que nunca teve contato com qualquer forma de treinamento).

Na abordagem ACERPI, *transfer learning* é utilizado na etapa de reconhecimento de entidades nomeadas, fazendo uso de modelos pré-treinados para o reconhecimento de nomes de pessoas em português com dados de notícias para a extração do nome de pessoas de Portarias. Com isso, mostrou-se necessário um baixo volume de dados anotados para que os novos modelos conseguissem realizar a identificação das entidades no novo domínio endereçado (Portarias de Instituições).

2.3 *NoSQL: Bancos de Dados Orientados a Documentos*

Segundo AWS (2020), bancos de dados NoSQL (*Not only SQL* - Não apenas SQL) são bancos de dados não relacionais de alta performance com modelos de dados flexíveis. Existem 4 tipos principais de bancos de dados NoSQL (MONGODB, 2020b):

- Orientados a documentos.
- Chave-valor.
- Orientados a colunas.
- Orientados a grafos.

Para o contexto deste trabalho, o banco de dados utilizado segue o conceito de documentos. Documentos são estruturas livres de esquema pré-definido, armazenados em um formato padrão e compostos de pares chave e valor. Estes, por sua vez, podem ser compostos de *strings*, valores numéricos, arrays, ou mesmo outros documentos. A escolha de um banco de dados orientado a documentos deu-se em especial pela flexibilidade na evolução dos dados armazenados e na facilidade de uso que esse banco de dados NoSQL provê.

Dentro das características citadas acima, a sua popularidade (DB-ENGINES, 2020) e a vasta documentação, foi escolhido o banco de dados MongoDB (MONGODB, 2020a) para a realização deste trabalho. O MongoDB é um banco de dados orientado a documentos, que armazena os mesmos em formato JSON e possui uma interface simples para recuperação dos dados. Nas Listagens 2.1 e 2.2 tem-se, respectivamente, um exemplo de documento utilizado neste trabalho e uma consulta utilizada para acessá-lo.

Listagem 2.1 – Exemplo de documento do MongoDB

```
{
    "id": 4630,
    "name": "RENATA DE MATOS GALANTE",
    "siape": ["1488770"],
    "document": {
        "name": "50216"
    }
}
```

Listagem 2.2 – Exemplo de consulta no MongoDB

```
{
    "name": "RENATA DE MATOS GALANTE",
    "document.name": "50216"
}
```

2.4 *Spacy*

Spacy (Explosion.ai, 2020c) é uma biblioteca de código fonte aberto para a linguagem de programação Python focada em Processamento de Linguagem Natural (PLN) avançado. Com ela, pode-se fazer uso de modelos de Inteligência Artificial focados em PLN para diversas tarefas (Explosion.ai, 2020b). Dentre as principais, destacam-se Tokenização (Explosion.ai, 2020f), Reconhecimento de Entidades Nomeadas (Explosion.ai, 2020d) e Resolução de Entidades (Explosion.ai, 2020a).

Estão disponíveis modelos treinados de maneira genérica nos mais variados idiomas. Porém, para cada aplicação, sugere-se que seja utilizada transferência de aprendizado para adaptação dos modelos pré-existentes de acordo com o domínio específico da

tarefa em questão.

Essa biblioteca foi utilizada neste trabalho para o reconhecimento dos nomes dos servidores públicos mencionados nas Portarias, uma das etapas centrais da abordagem aqui apresentada. Quando dado como entrada à biblioteca a frase "Pedro conversou com Maria sobre o acontecido." e utilizado o reconhecimento de entidades nomeadas, o mesmo retorna os termos "Pedro"e "Maria".

2.5 *Prodigy*

Prodigy (Explosion.ai, 2020e) é uma ferramenta para acelerar e facilitar o processo de anotação e treinamento de modelos de aprendizado de máquina. A ferramenta é baseada na biblioteca Spacy (Explosion.ai, 2020c), que fornece a estrutura base para Processamento de Linguagem Natural, bem como modelos pré-treinados e com uso simplificado.

Através da interface gráfica da Prodigy, foi possível identificar o nome dos servidores de maneira facilitada, dada a pré-classificação do trecho a ser anotado utilizando um modelo genérico inicial treinado para a língua portuguesa. Um exemplo de anotação pode ser visto na Figura 2.1.

Figura 2.1 – Captura de tela de um exemplo sendo anotado utilizando a ferramenta Prodigy

A ferramenta é paga e foi utilizada no projeto a partir de uma versão disponibilizada gratuitamente para fins de pesquisa.

2.6 Apache PDFBox

Segundo The Apache Software Foundation (2020), Apache PDFBox é uma biblioteca de código fonte aberto para a linguagem de programação Java que permite ao desenvolvedor trabalhar com arquivos no formato PDF. Essa ferramenta pode ser utilizada para a criação de novos documentos, manipulação de documentos já existentes e extração de conteúdo a partir de arquivos PDF.

O uso desta biblioteca neste trabalho deu-se na conversão dos arquivos PDF coletados dos repositórios das Instituições para arquivos de texto, que depois são interpretados para a extração de informações-chave, como o número e a data das Portarias.

2.7 Considerações Finais

Nesse capítulo foram apresentadas as tecnologias que se mostraram úteis no desenvolvimento da abordagem descrita nesta monografia. Tais tecnologias foram imprescindíveis para que o objetivo fosse atingido com a rápida evolução e qualidade dispostas.

As técnicas mostradas foram utilizadas principalmente nas etapas de coleta dos arquivos (através de múltiplas formas de *Web Scraping* e conversão dos dados do formato original para um formato textual intermediário) e extração de informações (com o uso da *Prodigy* no treinamento dos modelos de reconhecimento de entidades nomeadas).

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os trabalhos relacionados à abordagem desenvolvida. Inicialmente, são descritos os trabalhos de (PIVETTA, 2013) e (MANICA; DORNELES; GALANTE, 2017), cujas abordagens incluem a coleta e estruturação de documentos de diferentes tipos. Depois, é descrito o trabalho de (DOZIER et al., 2010), que inclui a análise e implementação das etapas de reconhecimento de entidades nomeadas e resolução de entidades no contexto de documentos legais dos Estados Unidos da América. Por fim, é realizada uma comparação entre as abordagens e implementações mencionadas e a abordagem ACERPI, desenvolvida neste trabalho.

3.1 Coleta e Estruturação de Documentos

A abordagem Orion (MANICA; DORNELES; GALANTE, 2017) tem como objetivo descobrir e extrair entidades reais e valores de atributos de páginas entidade. Uma página-entidade é uma página *web* que publica dados que descrevem uma entidade de um determinado tipo (BLANCO et al., 2008).

A descoberta das entidades dá-se assumindo a similaridade entre o HTML e URL de páginas-entidade e realizando buscas de novas possíveis páginas-entidade a partir de um exemplo inicial. A busca ocorre percorrendo a estrutura hierárquica da árvore-entidade identificada a partir da URL da página-entidade exemplo e comparação das páginas encontradas com a página-entidade exemplo, até que não sejam identificadas novas páginas-entidade na árvore-entidade identificada.

Diferente da abordagem ACERPI, a extração dos atributos das entidades na abordagem Orion acontece a partir da estrutura das páginas-entidade encontradas, que correspondem a árvores DOM. Árvores DOM podem ser tratadas como grafos e por isso permitem o uso de linguagens de consultas de bancos de dados em grafos para acessar seus atributos. Dessa forma, não faz-se necessária a técnica de PNL de reconhecimento de entidades nomeadas proposta na abordagem ACERPI para a identificação do nome das entidades do mundo real mencionadas nos documentos. A abordagem utilizada em Orion só é possível dado o caráter estruturado das páginas-entidade, que em geral seguem um *template* e diferem-se através do conteúdo dos campos de dados.

No trabalho de Pivetta (2013), foi realizada a identificação de Boletins Internos do Exército Brasileiro que mencionam militares e a classificação de relevância de eventos

(sentenças) presentes nesses Boletins com o intuito de automatizar a confecção de relatórios (Folhas de Alterações) onde constam somente eventos relevantes de um dado militar durante um período de 6 meses.

O trabalho realiza a busca dos documentos relacionados aos militares através de consultas pelo nome completo do militar ou da sua "graduação mais o nome de guerra". Dos documentos identificados, são processados os respectivos conteúdos e então extraídas sentenças que mencionam o nome do militar em questão para posterior classificação. Durante o processamento dos textos dos documentos, são aplicadas transformações como o uso de *stemming* (redução das palavras ao seu radical), a remoção de *stopwords* e a remoção de palavras com menos de três caracteres. Após, é realizada a extração das sentenças, que acontece identificando a menção ao militar em questão e selecionando uma janela de k caracteres antes e depois da menção. É importante notar que, na abordagem de Pivetta (2013), o nome do militar é conhecido previamente, permitindo buscas tanto por documentos que o mencionam, quanto da localização exata das menções nos documentos recuperados. Na abordagem ACERPI, apesar da inexistência do conhecimento prévio do nome do servidor, também é utilizada uma janela de caracteres após a menção para identificação de informações relacionadas ao mesmo, como a matrícula SIAPE. Ainda, caso não sejam encontradas informações relacionadas ao servidor na janela, são associadas informações genéricas do contexto que possam vir a auxiliar na etapa de deduplicação das entidades.

A última etapa, de classificação, é realizada através do aprendizado supervisionado com o algoritmo de *Naive Bayes*, onde, para treinamento, foi possível extrair e utilizar-se de Folhas de Alteração de militares desenvolvidas de maneira não-automatizada no passado e os Boletins Internos citados e não citados nas respectivas Folhas de Alteração. A partir das anotações de treinamento e o pré-processamento dos dados, obteve-se uma métrica combinada de precisão e revocação de 78,5% na melhor combinação das técnicas exploradas na classificação da relevância das sentenças.

3.2 Reconhecimento de Entidades Nomeadas e Resolução de Entidades

No trabalho de Dozier et al. (2010), foram descritos métodos de reconhecimento de entidades nomeadas utilizando técnicas de *lookup*, regras de contexto e modelos estatísticos. Também foram descritas técnicas empregadas na resolução de entidades como blocagem, *features* para funções de *matching* e aprendizado supervisionado e semi-

supervisionado para a função de *matching*. Ainda, foram empregadas parte das técnicas na extração e resolução de entidades no contexto de documentos legais dos Estados Unidos da América, como casos de jurisprudência, depoimentos, defesas e outros documentos de julgamentos. O trabalho não descreve como foi realizada a coleta dos documentos legais analisados.

As entidades relacionadas no trabalho de (DOZIER et al., 2010) incluem juízes, advogados, empresas, jurisdições e tribunais. Foram criados *taggers*, isto é, sistemas de reconhecimento de uma classe de entidades, diferentes para cada classe. Cada *tagger* pode empregar uma metodologia diferenciada, como *lookup*, regras de contexto e modelos estatísticos no reconhecimento de uma entidade de sua classe.

As entidades encontradas passam, posteriormente, pela etapa de resolução. A resolução empregada, diferente da utilizada neste trabalho, ocorre com o objetivo de associar cada entidade encontrada a uma entrada em um arquivo de autoridade (*Authority File*). Os arquivos de autoridade possuem, em cada linha, uma entidade no mundo real da classe representada pelo respectivo arquivo (juízes, advogados, empresas, jurisdições e tribunais) e atributos conhecidos da entrada. A existência de um arquivo de autoridade também permitiu aos autores encontrarem a melhor combinação de atributos para os quais a ambiguidade das entradas é mínimo. Como exemplo, os autores mencionam que, caso o arquivo de autoridades de advogados possua muitos advogados com o mesmo nome e sobrenome, a ambiguidade do arquivo é alta quando utilizados esses atributos.

A resolução de entidades deu-se utilizando algoritmos de classificação do tipo Máquinas de Vetores de Suporte. Para isso, foram anotados casos positivos onde as entidades encontradas foram relacionadas a entradas nos arquivos de autoridade e casos negativos, que puderam ser inferidos dos casos positivos, onde a entidade que possui uma relação com uma entrada no arquivo de autoridade não possui relação com as outras entradas do mesmo arquivo.

3.3 Considerações Finais

Neste capítulo, foram apresentados três trabalhos relacionados à abordagem ACERPI, desenvolvida neste trabalho. Primeiramente, a abordagem *Orion*, que realiza a descoberta e extração de dados de páginas *web* baseadas em *template* que descrevem uma única instância de uma entidade no mundo real. Em seguida, foi descrita a abordagem desenvolvida por Pivetta (2013), onde é realizada a seleção de sentenças que mencionam

um dado militar em Boletins Internos do Exército Brasileiro para uma posterior classificação da relevância das sentenças na geração automática de Folhas de Alteração. Por fim, foi descrito o trabalho de Dozier et al. (2010), onde os autores descrevem abordagens de reconhecimento de entidades nomeadas e resolução de entidades, além de realizar a implementação de parte das técnicas em documentos legais dos Estados Unidos da América.

A Tabela 3.1 descreve, das etapas realizadas na abordagem ACERPI, quais são realizadas pelas abordagens propostas nos trabalhos mencionados. A abordagem ACERPI propõe um *pipeline* de processamento de dados genérico que cobre a união das etapas de coleta e estruturação, reconhecimento de entidades e extração de informações e resolução de entidades de Portarias institucionais.

Tabela 3.1 – Comparação entre abordagens propostas nos trabalhos relacionados

Abordagem	**Coleta e Estruturação**	**Reconhecimento de Entidades Nomeadas**	**Resolução de Entidades**
Orion	Sim	Não	Não
PIVETTA, 2013	Sim	Não	Não
DOZIER et al., 2010	Não	Sim	Sim
ACERPI	Sim	Sim	Sim

4 ACERPI

Este Capítulo apresenta a ACERPI, que tem como objetivo utilizar técnicas de descoberta, obtenção, conversão e estruturação de arquivos, extração de informação e resolução de entidades para criar uma abordagem de busca de informações profissionais de servidores públicos de maneira categorizada, filtrada e agrupada em bases de dados de portarias federais.

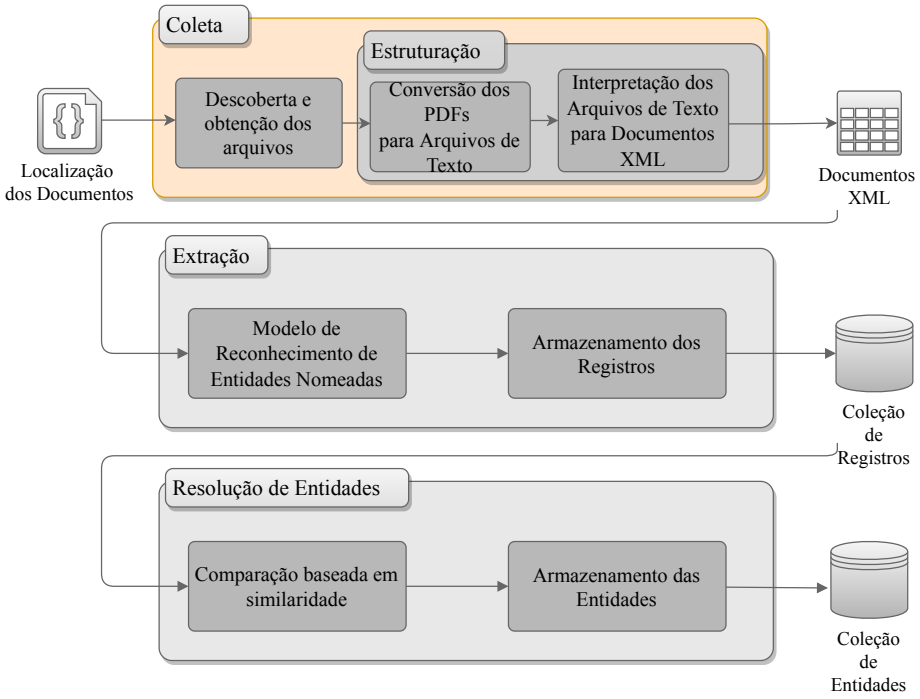
Inicialmente, uma visão geral da abordagem proposta é apresentada, seguida das três etapas principais, sendo elas: Coleta, Extração e Resolução de Entidades. As etapas são, então, divididas em subseções detalhando as metodologias e técnicas utilizadas, bem como onde se encaixam na abordagem como um todo.

4.1 Visão Geral

A ACERPI propõe a criação de um banco de dados não relacional para consultas avançadas a respeito dos documentos relacionados a um servidor de uma Instituição, bem como quais servidores são referenciados em um dado documento publicado. Com isso, a abordagem permite uma fácil pesquisa a respeito da jornada de servidores públicos através de portarias disponíveis abertamente em repositórios de órgãos federais. A Figura 4.1 ilustra o fluxo dos dados utilizados desde sua fonte até o armazenamento e pós-processamento. O ACERPI tem como entrada um conjunto repositórios de documentos de Portarias. Como saída, é gerado um banco de dados com as informações estruturadas dos servidores mencionados, bem como detalhes dos documentos e dos seus metadados.

A etapa de coleta inclui a descoberta e obtenção dos arquivos, a conversão para formato textual e a estruturação dos documentos de texto para arquivos XML identificando as portarias publicadas no documento em questão. A etapa de extração de informação utiliza de técnicas de *Named Entity Recognition* (SARKAR, 2018) e *Transfer Learning* (Seção 2.2) para identificar referências aos servidores e os metadados relacionados e armazená-los em um formato padrão. Por fim, são utilizadas técnicas de *Entity Resolution* (Data Community DC, 2013) para relacionar as referências encontradas com os servidores correspondentes e formar a base de dados final. A base de dados final pode ser utilizada para obter informações de um servidor, os documentos que o mencionam e os principais metadados extraídos.

Figura 4.1 – Visão Geral do funcionamento da ACERPI[1]



Um exemplo de entrada é o repositório UFRGS (2016) que contém, dentre os documentos disponíveis, a Portaria nº 10403 de 13/11/2017, ilustrada na Figura 4.2. Essa Portaria é do tipo substituição e refere-se às servidoras Renata de Matos Galante e Carla Maria dal Sasso Freitas. Essas informações são exemplos dos dados contidos no banco de dados de saída da execução do ACERPI.

4.2 Coleta

Nesta seção, é abordada a estratégia utilizada para descoberta, obtenção, conversão e *parsing* dos documentos para um formato intermediário, estruturado e utilizando a linguagem de marcação XML. A partir dessas marcações XML, são feitas análises subsequentes da etapa de extração, contempladas na Seção 4.3.

4.2.1 Descoberta e Obtenção dos Arquivos

Os dados iniciais, arquivos PDF de Portarias, são baixados dos repositórios das Instituições. Nesses repositórios, estão disponíveis de maneira heterogênea diversos do-

[1] A etapa de coleta, em destaque, foi realizada em parceria com Serigne Khassim Mbaye, estudante de graduação do Instituto Federal do Rio Grande do Sul (IFRS), *Campus* Ibirubá.

Figura 4.2 – Portaria número 10403 de 13/11/2017, emitida pela Administração Central Universidade Federal do Rio Grande do Sul

SERVIÇO PÚBLICO FEDERAL

PORTARIA Nº 10403 de 13/11/2017

A PRÓ-REITORA DE GESTÃO DE PESSOAS DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL EM EXERCÍCIO, no uso de suas atribuições que lhe foram conferidas pela Portaria nº. 8117, de 10 de outubro de 2016, e conforme a Solicitação de Afastamento n°32907,

RESOLVE

Designar, temporariamente, nos termos da Lei nº. 8.112, de 11 de dezembro de 1990, com redação dada pela Lei nº.9.527, de 10 de dezembro de 1997, a ocupante do cargo de PROFESSOR DO MAGISTÉRIO SUPERIOR, do Quadro de Pessoal desta Universidade, **RENATA DE MATOS GALANTE** (Siape: 1488770), para substituir CARLA MARIA DAL SASSO FREITAS (Siape: 0351477), Diretor do Instituto de Informática, Código CD-3, em seu afastamento no país, no período de 14/11/2017 a 15/11/2017, com o decorrente pagamento das vantagens por 2 dias.

VÂNIA CRISTINA SANTOS PEREIRA
Pró-Reitora

Documento gerado sob autenticação Nº LQH.287.570.56V, disponível no endereço http://www.ufrgs.br/autenticacao

Documento certificado eletronicamente, conforme Portaria nº 3362/2016, que institui o Sistema de Documentos Eletrônicos da UFRGS.

1/1

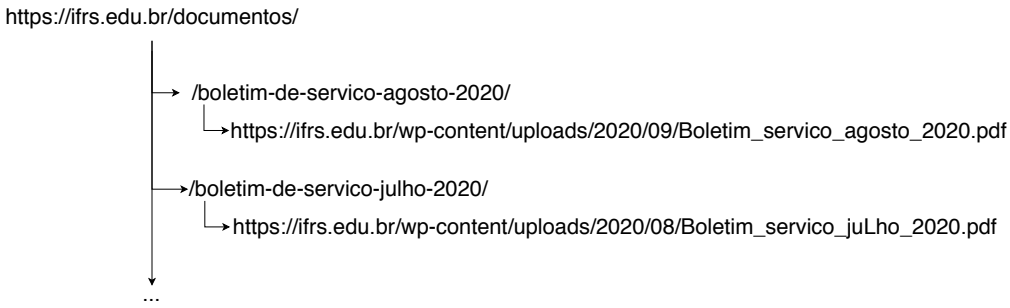
cumentos, incluindo as Portarias, conjunto de interesse para esse trabalho.

O método de descoberta e obtenção dos arquivos baseiam-se em técnicas de *Web Scraping* (Seção 2.1). Porém, para cada tipo de repositório, é utilizada uma ou mais técnicas conforme as restrições e possibilidades que cada repositório possui.

Para casos onde o repositório segue uma estrutura hierárquica, a descoberta e obtenção pode ser realizada caminhando pelas pastas e realizando a coleta dos arquivos conforme a estrutura disponível. Essa alternativa permite ainda a obtenção dos caminhos percorridos como metadados, caso seu conteúdo seja relevante para análise. A Figura

4.3 é um exemplo de repositório com uma estrutura hierárquica, e a obtenção dos dados segue a partir dos diretórios aos arquivos, como uma busca em profundidade realizada sobre uma árvore de pesquisa.

Figura 4.3 – Exemplo de estrutura do repositório de documentos do Instituto Federal do Rio Grande do Sul - *Campus* Ibirubá



Outra alternativa, também utilizada neste trabalho, é a inferência de um padrão de navegação (Palmieri Lage et al., 2004) que, através de uma expressão regular, generaliza as URLs relevantes do repositório para automatização da obtenção em uma etapa posterior. Essa técnica pode ser muito útil quando os arquivos tem um formato padrão que pode definir o sucesso de uma tentativa de *download*. O repositório de arquivos IEEE (2020) pode ser utilizado como exemplo de uso dessa abordagem, onde o padrão de navegação `https://ieeexplore.ieee.org/document/[0-9]{1-7}` possui uma expressão regular que generaliza os endereços de acesso aos arquivos do repositório.

Ainda, em alguns casos, a coleta pode acontecer através de APIs que disponibilizam a informação estruturada e de fácil acesso. Apesar disso, dificilmente os dados de Portarias disponíveis são encontrados dessa maneira.

4.2.2 Estruturação

Após a descoberta e obtenção dos arquivos PDF e antes de iniciar a etapa de extração, ocorre a subetapa de estruturação[2]. O objetivo da subetapa de estruturação é transformar os dados de seu formato original (PDF) para um formato intermediário com o conteúdo das portarias individuais. A estruturação é feita em duas fases: conversão do PDF para texto e interpretação para múltiplas portarias.

Primeiramente, é realizada a conversão do arquivo PDF para o formato de texto. Essa etapa é realizada utilizando a biblioteca de manipulação de arquivos PDF Apache PDFBox (Seção 2.6). Através de um programa desenvolvido na linguagem Java, são

[2]Esta subetapa foi realizada em parceria com Serigne Khassim Mbaye.

percorridos todos os arquivos PDF do diretório que contém os documentos coletados na subetapa anterior realizando a extração do respectivo conteúdo textual, que é salvo em um documento de mesmo nome, porém no formato de texto. O resultado da conversão, quando aplicada à Portaria ilustrada na Figura 4.2, pode ser visto na Listagem 4.1.

Listagem 4.1 – Exemplo de conversão da Portaria da Figura 4.2 para arquivo de texto

```
Documento gerado sob autenticação Nº LQH.287.570.56V, disponível no
endereço http://www.ufrgs.br/autenticacao
Documento certificado eletronicamente, conforme Portaria nº
3362/2016, que institui o Sistema de Documentos Eletrônicos da
    UFRGS.
1/1
PORTARIA Nº 10403 de 13/11/2017
A PRÓ–REITORA DE GESTÃO DE PESSOAS DA UNIVERSIDADE FEDERAL DO RIO
    GRANDE DO SUL EM
EXERCÍCIO, no uso de suas atribuições que lhe foram conferidas pela
    Portaria nº. 8117, de 10 de outubro de
2016, e conforme a Solicitação de Afastamento n°32907,
RESOLVE
Designar, temporariamente, nos termos da Lei nº. 8.112, de 11 de
    dezembro de 1990, com redação
dada pela Lei nº.9.527, de 10 de dezembro de 1997, a ocupante do
    cargo de PROFESSOR DO MAGISTÉRIO
SUPERIOR, do Quadro de Pessoal desta Universidade, RENATA DE MATOS
    GALANTE (Siape: 1488770), para
substituir CARLA MARIA DAL SASSO FREITAS (Siape: 0351477),
    Diretor do Instituto de Informática, Código
CD–3, em seu afastamento no país, no período de 14/11/2017 a
    15/11/2017, com o decorrente pagamento
das vantagens por 2 dias.
VÂNIA CRISTINA SANTOS PEREIRA
Pró–Reitora
```

Em seguida, na segunda fase, é realizada a interpretação dos arquivos de texto, extraindo as múltiplas Portarias que podem constar em cada arquivo e suas principais informações, para o formato intermediário XML. Os arquivos nesse formato seguem o padrão descrito pelo exemplo da Listagem 4.2

Listagem 4.2 – Padrão da estrutura intermediária dos dados armazenados em arquivos XML

```
<documento id="" nome_arquivo="" site="">
```

```
    <portaria nr="" data="">
        Conteúdo da Portaria
    </portaria>
</documento>
```

O formato intermediário XML possui um elemento raiz chamado “documento”, que tem como atributos o identificador único do documento, o nome do arquivo PDF original e o endereço do arquivo no repositório da Instituição. Um documento pode ter um número arbitrário de filhos, denominados “portarias”. As portarias correspondem a cada Portaria que o documento PDF original possui, tendo como atributos o número e a data da mesma. Dentro dos elementos portarias, são armazenados os conteúdos textuais puros referentes única e exclusivamente à Portaria identificada pelo número e data extraídos anteriormente. Um exemplo de arquivo XML estruturado pode ser visto na Listagem 4.3.

Listagem 4.3 – Exemplo de estrutura intermediária gerada a partir da Portaria da Figura 4.2

```
<documento id="47048" nome_arquivo="47048.pdf" site="https://www1.
    ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/
    ExibirPDF?documento=47048">
    <portaria nr="10403" data="13/11/2017">
        PORTARIA Nº 10403 de
            13/11/2017
        A PRÓ-REITORA DE GESTÃO DE PESSOAS DA UNIVERSIDADE FEDERAL
            DO RIO GRANDE DO SUL EM EXERCÍCIO, no uso de suas
            atribuições que lhe foram conferidas pela Portaria nº.
            8117, de 10 de outubro de 2016, e conforme a Solicitação
             de Afastamento n°32907,
        RESOLVE
        Designar, temporariamente, nos termos da Lei nº. 8.112, de
            11 de dezembro de 1990, com redação dada pela Lei nº
            .9.527, de 10 de dezembro de 1997, a ocupante do cargo
            de PROFESSOR DO MAGISTÉRIO SUPERIOR, do Quadro de
            Pessoal desta Universidade, RENATA DE MATOS GALANTE (
            Siape: 1488770 ), para substituir CARLA MARIA DAL
            SASSO FREITAS (Siape: 0351477 ), Diretor do Instituto de
             Informática, CódigoCD-3, em seu afastamento no país, no
             período de 14/11/2017 a 15/11/2017, com o decorrente
            pagamentodas vantagens por 2 dias.
        VÂNIA CRISTINA SANTOS PEREIRA
        Pró-Reitora
```

```
9       </portaria>
10 </documento>
```

A extração dos dados desta fase é realizada através de expressões regulares que capturam os intervalos de caracteres os quais compõem os padrões que descrevem o formato de uma portaria, seu número e data de publicação, respectivamente.

Os documentos analisados neste trabalho possuem estruturas bem definidas que permitem a identificação desses padrões facilmente. O número da Portaria, por exemplo, é sempre prefixado com o termo "PORTARIA Nº". A expressão regular `PORTARIA Nº *(/d+)` captura essa informação, que é então armazenada como atributo da Portaria em análise.

4.3 Extração

Nesta seção, são abordadas as técnicas utilizadas para o reconhecimento dos nomes dos servidores nas Portarias, bem como a estrutura escolhida para armazenar esse relacionamento. Com essas informações, em etapa posterior, é feita a resolução de entidades.

4.3.1 Reconhecimento de Entidades Nomeadas

Essa etapa consiste em extrair do conteúdo das Portarias os nomes dos servidores mencionados. Para isso, é utilizada a biblioteca de processamento de linguagem natural Spacy (Seção 2.4) e um modelo pré-treinado adaptado para o domínio de Portarias.

O modelo de partida é o *pt_core_news_sm*[3], treinado a partir de uma base de dados de notícias em português e redes neurais convolucionais. Tendo o modelo *pt_core_news_sm* como base, é feita a anotação de dados de Portarias das Instituições utilizando a ferramenta Prodigy (Seção 2.5) para permitir um ajuste mais fino do modelo. A anotação de dados consiste em identificar e armazenar os termos que caracterizam nomes de servidores em uma pequena parcela de documentos. Um exemplo de anotação usando a ferramenta Prodigy (Seção 2.5) pode ser visto na Figura 2.1 e o resultado da anotação é um banco de dados contendo informações sobre os dados anotados (nome do documento, caractere onde o termo inicia, caractere onde o termo termina e metadados).

[3]Disponívem em <https://spacy.io/models/pt#pt_core_news_sm>. Último acesso em 22/11/2020.

Após a anotação dos dados, faz-se necessário o re-treino do modelo para que ele melhor interprete os arquivos do domínio de Portarias. Esse estágio é fundamental para o aperfeiçoamento do reconhecimento das entidades nomeadas em dados com padrões antes desconhecidos pelo modelo genérico. Isso pode ser feito novamente pela ferramenta Prodigy, através do comando `prodigy train ner a b –output c –eval-id d –n-iter e`, onde `a` representa o(s) banco(s) de dados que contém as anotações, `b` consiste do modelo inicial (i.e. *pt_core_news_sm*), `c` é o caminho onde o modelo de saída será salvo, `d` é o banco de dados de anotações que serão utilizadas para avaliação dos treinamentos e `f` o número de iterações.

Além do nome dos servidores, são extraídos via expressões regulares as respectivas matrículas SIAPE (quando presentes). Esse dado corresponde a um identificador único do servidor, que é utilizado na etapa de resolução de entidades. Nas Portarias, esses números geralmente acompanham o nome do servidor, mencionam o termo SIAPE e podem ser extraídos a partir de expressões regulares como `[S|s][I|i][A|a][P|p][E|e][^0-9.]{1,3}([0-9]{6,8})`. Caso não seja identificada a matrícula SIAPE do servidor de maneira precisa nos 120 caracteres subsequentes ao último caractere do nome do servidor, é armazenado para o servidor uma lista contendo todas as matrículas encontradas na Portaria em análise. Essa alternativa prova-se útil quando são mencionados nomes dos servidores em uma lista, seguida por outra lista com as respectivas matrículas SIAPE.

Para o arquivo PDF da Figura 4.2, a saída da etapa de extração contempla o nome das servidoras, RENATA DE MATOS GALANTE, CARLA MARIA DAL SASSO FREITAS e VÂNIA CRISTINA SANTOS PEREIRA. As matrículas SIAPE associadas às servidoras, 1488770 para Renata de Matos Galante, 0351477 para Carla Maria Dal Sasso Freitas e ambas as matrículas 1488770 e 0351477 para a servidora Vânia Cristina Santos Pereira. E também é realizada a associação desses dados à portaria 10403 do dia 13 de Novembro de 2017.

4.3.2 Amazenamento dos Registros

Nesta seção, é detalhada a estrutura de armazenamento utilizada, bem como a motivação do uso de um banco de dados NoSQL (Seção 2.3). O armazenamento dos registros permite a estruturação parcial dos dados, dando acesso às informações extraídas não só para a próxima etapa, mas também para o usuário que tem interesse em já realizar

consultas.

O banco de dados escolhido para o desenvolvimento desse projeto foi o MongoDB (MONGODB, 2020a), em especial pela versatilidade da modelagem de dados em um banco de dados orientado a documentos, que permite a evolução do esquema conforme a aplicação evolui (AWS, 2020). Além disso, o MongoDB possui sua base de código aberta e uma abrangente documentação, facilitando o seu uso.

Na abordagem ACERPI, foi definida uma estrutura de documento principal denominada registro que concentra as informações de um servidor identificado em uma portaria. Esse documento possui, respectivamente:

1. Um identificador único do registro.
2. O nome do servidor identificado na etapa de reconhecimento de entidades nomeadas.
3. Uma lista de SIAPEs identificados em portarias relacionados ao servidor. Essa lista é preenchida quando não é possível identificar um número de matrícula específica para o servidor, sendo inseridas à lista, então, todas as matrículas identificadas no documento onde o respectivo nome foi encontrado.
4. O identificador da portaria de onde esse registro foi encontrado.

Três registros são gerados tendo em vista o documento da Listagem 4.3. O registro com identificador 131072 indicado na Listagem 4.4 da servidora Renata de Matos Galante, o registro com identificador 131073 indicado na Listagem 4.5 da servidora Carla Maria Dal Sasso Freitas e o registro com identificador 131074 indicado na Listagem 4.6 da Pró-Reitora Vânia Cristina Santos Pereira.

Listagem 4.4 – Registro gerado a partir da análise do arquivo XML da Listagem 4.3 para a servidora Renata de Matos Galante

```
{
    "id": 131072,
    "name": "RENATA DE MATOS GALANTE",
    "siape": ["1488770"],
    "document": {
        "name": "47048"
    }
}
```

Listagem 4.5 – Registro gerado a partir da análise do arquivo XML da Listagem 4.3 para a servidora Carla Maria Dal Sasso Freitas

```
{
    "id": 131073,
    "name": "MARIA DAL SASSO FREITAS",
    "siape": ["0351477"],
    "document": {
        "name": "47048"
    }
}
```

Listagem 4.6 – Registro gerado a partir da análise do arquivo XML da Listagem 4.3 para a servidora Vânia Cristina Santos Pereira

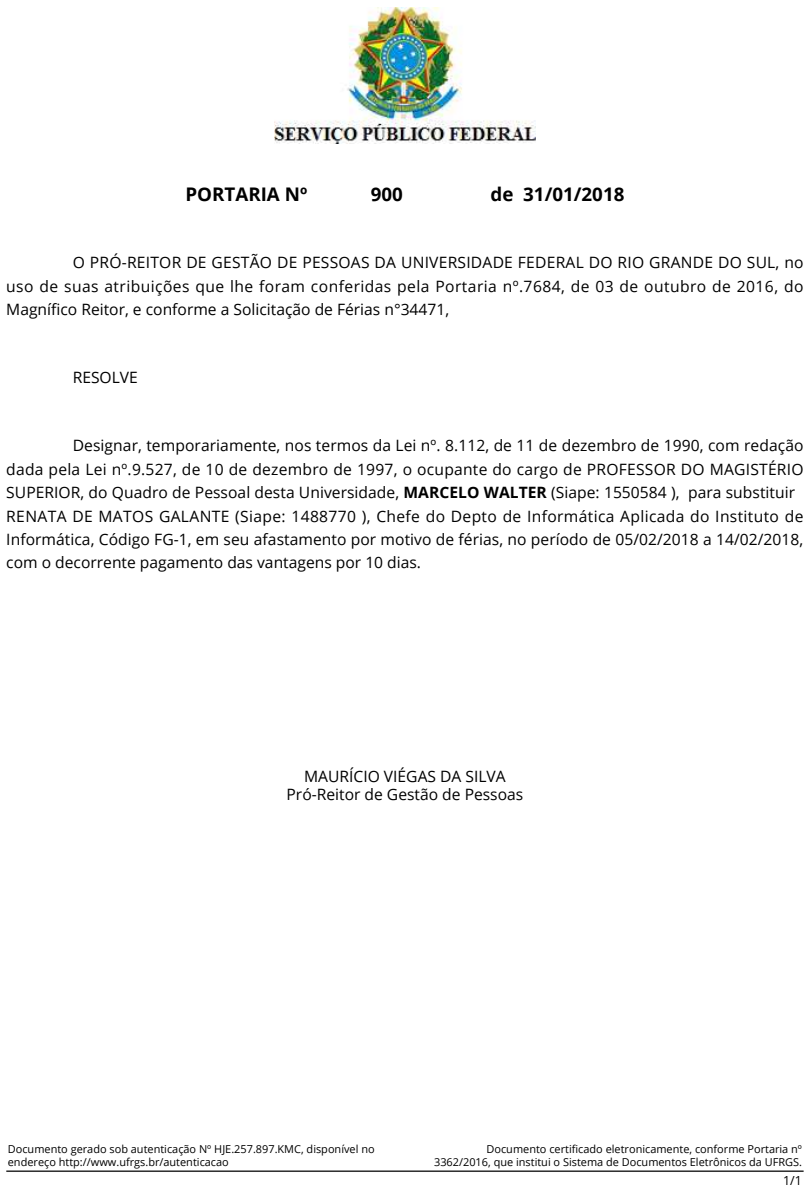
```
{
    "id": 131074,
    "name": "VÃNIA CRISTINA SANTOS",
    "siape": ["1488770", "0351477"],
    "document": {
        "name": "47048"
    }
}
```

4.4 Resolução de Entidades

A partir dos registros e seus dados associados, é feita a etapa de resolução de entidades. Esta etapa consiste em identificar, dadas as referências aos servidores nos documentos (registros) e as demais informações adquiridas, quais registros se referem às mesmas entidades no mundo real (i.e. quais documentos referenciam o mesmo servidor, refletindo portarias às quais ele esteve diretamente envolvido).

Como exemplo, as Portarias representadas nas Figuras 4.2 e 4.4 tem, dentre os servidores mencionados, a servidora Renata de Matos Galante. Ao final da etapa de resolução de entidades, é esperado que tenha sido identificado que a servidora foi mencionada em ambas as portarias e que as duas referências encontradas nas duas Portarias referem-se

Figura 4.4 – Portaria número 900 de 31/01/2018, emitida pela Administração Central Universidade Federal do Rio Grande do Sul



SERVIÇO PÚBLICO FEDERAL

PORTARIA Nº 900 de 31/01/2018

O PRÓ-REITOR DE GESTÃO DE PESSOAS DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, no uso de suas atribuições que lhe foram conferidas pela Portaria nº.7684, de 03 de outubro de 2016, do Magnífico Reitor, e conforme a Solicitação de Férias n°34471,

RESOLVE

Designar, temporariamente, nos termos da Lei nº. 8.112, de 11 de dezembro de 1990, com redação dada pela Lei nº.9.527, de 10 de dezembro de 1997, o ocupante do cargo de PROFESSOR DO MAGISTÉRIO SUPERIOR, do Quadro de Pessoal desta Universidade, **MARCELO WALTER** (Siape: 1550584), para substituir RENATA DE MATOS GALANTE (Siape: 1488770), Chefe do Depto de Informática Aplicada do Instituto de Informática, Código FG-1, em seu afastamento por motivo de férias, no período de 05/02/2018 a 14/02/2018, com o decorrente pagamento das vantagens por 10 dias.

MAURÍCIO VIÉGAS DA SILVA
Pró-Reitor de Gestão de Pessoas

Documento gerado sob autenticação Nº HJE.257.897.KMC, disponível no endereço http://www.ufrgs.br/autenticacao

Documento certificado eletronicamente, conforme Portaria nº 3362/2016, que institui o Sistema de Documentos Eletrônicos da UFRGS.

1/1

a mesma entidade do mundo real, a Professora Doutora Renata de Matos Galante, do Instituto de Informática da Universidade Federal do Rio Grande do Sul. Esse processo ocorre para todos os registros disponíveis de forma que, ao final, tem-se armazenado de maneira estruturada todas as associações entre portarias e servidores, bem como a identificação de quais menções a um servidor correspondem a mesma entidade do mundo real.

4.4.1 Comparação Baseada em Similaridade

A resolução de entidades na abordagem ACERPI é realizada conforme descrito no Algoritmo 1. O algoritmo recebe como entrada o conjunto de registros identificados a partir das Portarias e tem como saída o agrupamento dos registros que referem-se a mesma entidade do mundo real.

Algorithm 1: Algoritmo para a resolução de entidades.

```
   Entrada: conjunto de registros
   Saída  : agrupamentos de registros
 1 registros ← conjunto de registros;
 2 agrupamentos ← ∅;
 3 foreach registro r from registros do
 4 |   agrupado ← false;
 5 |   foreach agrupamento a from agrupamentos do
 6 |   |   if match(r, a) then
 7 |   |   |   a ← a ∪ r;
 8 |   |   |   agrupado ← true;
 9 |   |   |   break;
10 |   |   end
11 |   end
12 |   if not agrupado then
13 |   |   agrupamentos ← agrupamentos ∪ novo_agrupamento(r);
14 |   end
15 end
```

A função de *match* é parte principal do algoritmo, visto que define se o novo registro faz ou não faz parte do agrupamento (i.e. o quão similares são o novo registro e os registros já pertencentes ao agrupamento). A função de *match* pode ser simples, como uma comparação direta das entidades nomeadas dos registros, ou complexa, utilizando métodos de comparação de subtermos das entidades nomeadas e metadados dos registros. A abordagem ACERPI utiliza uma abordagem que analisa o contexto para a resolução das entidades, que no caso das portarias ocorre por meio do SIAPE. Quando únicas e idênticas, as matrículas SIAPE indicam a referência à mesma entidade do mundo real. Caso as matrículas SIAPE não sejam únicas e idênticas, é realizada a comparação das entidades nomeadas dos registros enriquecidas da remoção de espaços extras e descapitalização dos caracteres.

Assim, os registros das Listagens 4.4 e 4.5 não seriam agrupados por possuírem somente uma matrícula SIAPE associada a cada registro mas elas serem diferentes. Já os

registros das Listagens 4.4 e 4.7 seriam agrupados por possuírem um único número de matrícula SIAPE cada e eles serem idênticos.

Listagem 4.7 – Registro gerado a partir da análise da Portaria da Figura 4.4 para a servidora Renata de Matos Galante

```
{
    "id": 4630,
    "name": "RENATA DE MATOS GALANTE",
    "siape": ["1488770"],
    "document": {
        "name": "50216"
    }
}
```

4.4.2 Armazenamento das Entidades

Os agrupamentos resultantes da resolução de entidades são armazenados em uma estrutura denominada entidade. Uma entidade é gerada para cada agrupamento, representando uma entidade do mundo real. Cada entidade possui uma referência aos identificadores dos registros (*records*) que a compõem e um conjunto de nomes e matrículas SIAPE encontrados nos registros, de forma reduzir o custo computacional da resolução de entidades. Para os registros das Listagens 4.4 e 4.7, após a etapa de resolução de entidades, é gerada a entidade da Listagem 4.8.

Listagem 4.8 – Entidade gerada a partir da resolução de entidades dos registros das Listagens 4.4 e 4.7

```
{
    "records": [47048, 50216],
    "names": ["RENATA DE MATOS GALANTE"],
    "siapes": ["1488770"]
}
```

4.5 Considerações Finais

Neste capítulo, foi apresentada a abordagem ACERPI. A Abordagem ACERPI constitui de três etapas principais, são elas a coleta dos arquivos dos repositórios das Instituições (e estruturação dos documentos), a extração de informações e meta-informações a respeito das entidades nomeadas mencionadas nas Portarias, e a resolução das entidades identificadas em entidades do mundo real. Ao final, tem-se dados estruturados a respeito dos servidores mencionados nos documentos, bem como em quais documentos cada servidor foi referenciado.

5 AVALIAÇÃO EXPERIMENTAL

Este capítulo descreve os experimentos desenvolvidos para avaliar a Abordagem ACERPI. Primeiramente, são apresentados os experimentos para a coleta e estruturação das Portarias. Em seguida, são apresentados os experimentos que avaliam o reconhecimento de entidades nomeadas. Por fim, são apresentados os experimentos relacionados à resolução de entidades para servidores do mundo real.

5.1 Visão Geral

Nesta seção são detalhados os três principais experimentos realizados no trabalho. São eles:

1. Avaliar as técnicas de coleta dos arquivos das duas fontes de dados utilizadas e dos diversos repositórios disponíveis, a partir de técnicas de *Web Scraping* (Seção 2.1), automação de acesso a páginas web e estruturação das Portarias.
2. Avaliar estratégias de anotação de dados necessárias para extrair corretamente as entidades nomeadas das Portarias.
3. Experimentar estratégias de resolução de entidades para agrupar referências a um mesmo servidor e avaliar a(s) melhor(es) para uso na abordagem desenvolvida neste trabalho.

5.2 Configurações Gerais dos Experimentos

As partes em comum a todos os experimentos são explicadas nesta seção, enquanto as particularidades são abordadas individualmente no detalhamento de cada experimento.

5.2.1 Fontes de Dados

Para todos os experimentos, foram utilizadas duas fontes de dados, são elas:

- DOCS-UFRGS - Documentos públicos da Universidade Federal do Rio Grande do Sul. Desta fonte, foram extraídos documentos, em sua maioria Portarias, do repositório da Universidade. Os documentos podem ser acessados através

de endereços generalizados a partir do padrão de navegação (Palmieri Lage et al., 2004) `https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/ExibirPDF?documento=[0-9]{1,6}`, onde, para este trabalho, a expressão regular foi substituída pelos valores 18000 a 105995. Para essa fonte, cada arquivo contém no máximo uma portaria emitida pela Universidade.

- DOCS-IFRS - Documentos públicos do Instituto Federal do Rio Grande do Sul. Para esta fonte, foram extraídos documentos de três repositórios. São eles:

  1. Repositório antigo IFRS, *Campus* Ibirubá (IFRS, Campus Ibirubá, 2011). Este repositório é composto por arquivos apenas do *Campus* Ibirubá de 2013 a 2017 (inclusive).
  2. Repositório atual IFRS, *Campus* Ibirubá (IFRS, Campus Ibirubá, 2018). Este repositório complementa o anterior com os documentos a partir de 2017.
  3. Repositório geral do IFRS (IFRS, 2020). Este repositório é composto por documentos do IFRS, contemplando portarias relacionadas aos servidores dos 17 campi deste instituto desde 2017.

  Os dois últimos seguem sendo atualizados conforme novas publicações acontecem. Para esta fonte, cada arquivo pode conter uma quantidade arbitrária de portarias emitidas pela instituição.

As duas fontes de dados possuem repositórios que são continuamente atualizados. Estes experimentos foram realizados em dados coletados até 3 de Março de 2020 para a fonte DOCS-UFRGS e 17 de Fevereiro de 2020 para a fonte DOCS-IFRS. Informações adicionais a respeito do volume das fontes de dados encontram-se na Tabela 5.1.

Tabela 5.1 – Fontes de dados utilizadas nos Experimentos

Nome da Fonte de Dados	**Tamanho (Gb)**	**# Arquivos PDF**
DOCS-UFRGS	7.99	44865
DOCS-IFRS	0.152	313

5.3 Experimento 1 - Técnicas de Coleta

Este experimento tem como objetivo avaliar as técnicas de coleta dos arquivos. Foram utilizadas técnicas de *Web Scraping* (Seção 2.1), *scripts* para automação do acesso às páginas e *download* do seu conteúdo, ferramentas auxiliares de conversão de arquivos

e expressões regulares para estruturação das Portarias.

A realização do experimento 1 para a fonte de dados DOCS-IFRS foi realizada por Serigne Khassim Mbaye, estudante de graduação do Instituto Federal do Rio Grande do Sul, *Campus* Ibirubá, e será disponibilizada no respectivo trabalho.

5.3.1 Métricas

Para este experimento, a métrica utilizada foi a seguinte:

- Precisão - percentual dos documentos recuperados que são relevantes (BAEZA-YATES; RIBEIRO-NETO, 2011). A fórmula da precisão pode ser descrita por $(Relevantes \cap Recuperados)/Recuperados$.

Para a fonte de dados DOCS-UFRGS, entende-se como relevante os documentos que possuem termos que atendem a expressão regular `PORTARIA Nº[ ]*[0-9]{0,5}[ ]*de[ ]*[0-9]{2}[0-9]{2}[0-9]{4}`. A inferência desta expressão regular foi realizada através da análise visual de documentos desta fonte que foram classificados como Portarias.

A métrica de Revocação (BAEZA-YATES; RIBEIRO-NETO, 2011), de fórmula $(Relevantes \cap Recuperados)/Relevantes$, não foi utilizada pela falta da informação do total de documentos relevantes existentes para cada fonte de dados.

5.3.2 Ambiente de Configuração

Este experimento foi realizado em uma instância EC2 do tipo t2.micro [1]. Essa instância possui 1 CPU virtual, 1 GiB de memória e performance de rede baixa a moderada (não detalhado pela AWS).

5.3.3 Metodologia

Foram utilizadas técnicas diferentes para a descoberta e obtenção dos documentos de cada fonte de dados devido às particularidades dos respectivos repositórios.

[1] Disponível em <https://aws.amazon.com/pt/ec2/>. Último acesso em 21/11/2020.

5.3.3.1 DOCS-UFRGS

Essa fonte é composta por somente um repositório cujo acesso é restrito e controlado por uma ferramenta limitadora chamada CAPTCHA. Essa restrição impede o *download* dos arquivos de maneira simples, obrigando o uso de abordagens não convencionais para a coleta de todos os arquivos de sua base.

A abordagem utilizada foi a de *Web Scraping* (Seção 2.1). Foi identificado um padrão nos redirecionamentos das URLs fornecidos pelo repositório oficial (UFRGS, 2016) de forma que não fosse necessário o preenchimento do CAPTCHA para o acesso aos documentos.

Como exemplo, a portaria da Figura 4.2, está localizada no endereço <https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/ExibirPDF?documento=47048>. Foi inferido um padrão de navegação (Palmieri Lage et al., 2004) cuja expressão regular generaliza os endereços relevantes para este repositório, dado por `https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/ExibirPDF?documento=[0-9]{1,6}`. A partir dessa inferência, foi automatizado o *download* de todos os arquivos disponíveis no repositório, fazendo uso da variação de termos válidos pela expressão regular contida no padrão de navegação.

O uso dessa técnica gerou, também, a limitação no número de requisições ao repositório da Universidade. Isso foi percebido devido ao retorno da requisição HTTP, que possuía o `Status Code 429 - Too Many Requests`. Para contornar a limitação, foi inserido um *delay* de 60 segundos a cada 100 requisições ao servidor.

Após, foi realizada a conversão dos documentos PDF em arquivos de texto. Esta conversão obteve sucesso em todas as execuções para esta fonte de dados e o número de arquivos de saída foi exatamente igual ao número de arquivos recuperados.

A última etapa deste experimento, de extração de dados das Portarias de cada documento (agora textual), para esta fonte de dados deu-se através da criação de um arquivo XML no formato definido para cada um dos arquivos de texto e a extração dos dados de número da Portaria e data através das expressões regulares descritas na Seção 4.2.2. Como, para esta fonte de dados, cada arquivo recuperado possui exatamente uma portaria, não houve necessidade de realizar a identificação do número de Portarias por arquivo.

5.3.4 Resultados

Para a fonte de dados DOCS-UFRGS, foram recuperados 44865 arquivos ao total. Destes, 44094 atenderam a expressão regular `PORTARIA N°[ ]*[0-9]{0,5}[ ]*de[ ]*[0-9]{2}[0-9]{2}[0-9]{4}`, que foi utilizada para classificar o documento como relevante. A precisão obtida foi de 98.3%, indicando, de maneira aproximada, que para cada 100 documentos recuperados, 98 são relevantes.

Dos documentos recuperados relevantes, todos foram convertidos para o formato de texto e posteriormente para o formato XML corretamente.

5.3.5 Análise dos Casos de Falha

Durante o desenvolvimento da abordagem de coleta dos dados da fonte *DOCS-UFRGS*, foram encontrados alguns obstáculos que tiveram de ser ultrapassados. Foram eles:

- A existência de uma inteligência anti-automação na API oficial de recuperação dos documentos. Esse impedimento foi superado ao ser utilizado diretamente o endereço final dos arquivos, em detrimento dos filtros permitidos via o meio oficial.
- A heterogeneidade das respostas às requisições de recuperação dos arquivos. Isso deu-se de maneira que o servidor da Universidade respondia às requisições com arquivos PDF com diferentes *encodings* (mais especificamente `UTF-8` e `latin1`), o que não era previsto inicialmente. Esse impedimento foi superado através de múltiplas tentativas de armazenamento dos arquivos PDF recebidos utilizando-se de diferentes *encodings*, até que um obtivesse sucesso.
- O limite de requisições imposto pelo servidor da Universidade. Esse impedimento foi superado através do espaçamento das requisições, evitando a sobrecarga do servidor e, portanto, o bloqueio às requisições.

5.4 Experimento 2 - Reconhecimento de Entidades Nomeadas

Este experimento tem como objetivo avaliar estratégias de anotação de dados necessárias para extrair corretamente as entidades nomeadas das Portarias (i.e. os nomes dos servidores).

Para isso, foi utilizada a ferramenta de anotação e treinamento de modelos de Reconhecimento de Entidades Nomeadas intitulada Prodigy (Seção 2.5). Através dela foram realizadas as etapas de anotação, treinamento e extração de métricas relacionadas a eficácia dos modelos.

5.4.1 Métricas

Para este experimento, as métricas utilizadas foram as seguintes:

- Precisão - percentual de nomes do conjunto identificados corretamente pelo modelo como nomes de entidades. A fórmula da precisão pode ser descrita por $VerdadeirosPositivos/(VerdadeirosPositivos + FalsosPositivos)$.
- Revocação - o percentual de nomes existentes no conjunto e que o modelo conseguiu identificar corretamente. A fórmula da revocação pode ser descrita por $VerdadeirosPositivos/(VerdadeirosPositivos + FalsosNegativos)$.
- Medida F1, uma média ponderada que utiliza Precisão (P) e Revocação (R) para, através de uma só métrica, medir a eficácia do modelo. A fórmula da F1 pode ser descrita por $2 * P * R/(P + R)$.

As métricas anteriormente descritas são representadas em função de três valores, são eles:

- Verdadeiros Positivos - quantidade de termos identificados pelo modelo como nome de um servidor e que realmente representavam o nome do servidor.
- Falsos Positivos - quantidade de termos identificados pelo modelo como nome de um servidor incorretamente (ou mesmo parcialmente).
- Falsos Negativos - quantidade de termos não identificados pelo modelo como nome de um servidor porém que representavam o nome de um servidor.

Para todas as métricas, é importante notar que nomes encontrados parcialmente foram definidos como Falso Positivos, dado que era esperado que o modelo encontrasse o nome da entidade por completo. Por exemplo, um caso onde o nome da servidora é “Renata de Matos Galante” e o nome de entidade identificado na Portaria fosse apenas “Renata de Matos” seria considerado um falso positivo. A métrica de acurácia também é utilizada neste experimento. Esta métrica é calculada pela ferramenta *Prodigy* e não foi encontrada nenhuma forma de documentação a respeito de como o cálculo é realizado.

5.4.2 Ambiente de Configuração

Este experimento foi realizado em um Macbook Pro mid-2014 com processador 2,8 GHz Intel Core i5 Dual-Core, 8Gb de Memória, armazenamento SSD e sistema operacional macOS versão 10.15.5.

5.4.3 Metodologia

Foram realizadas sessões de anotação e treinamento de modelos especializados em reconhecimento de entidades nomeadas para que fossem adaptados a compreender as estruturas das Portarias. Todos os treinamentos foram baseados no modelo pré-existente *pt_core_news_sm*, treinado com notícias e textos em português, a partir do qual as nuances das portarias foram transferidas através dos treinamentos com os dados anotados.

O processo de anotação e treino seguiu o Algoritmo 2, onde dados são anotados a partir de sugestões de anotações geradas pelo modelo atual e incluídos no conjunto de treinamento até que seja identificado que o aumento do conjunto de treino não traz melhora na eficácia do modelo final. Isto é feito através de um método da ferramenta Prodigy chamado `train-curve`, que treina o modelo inicial com 25%, 50%, 75% e 100% dos dados anotados e calcula a acurácia dos modelos intermediários[2], indicando (de maneira simplificada) a melhora das predições dos modelos conforme o aumento do conjunto de treinamento. Quando não há uma melhora com o aumento do conjunto de treinamento, o processo de criação do modelo é encerrado e o modelo treinado com o conjunto de dados anotados até então é escolhido.

A partir deste ponto, melhorias ao modelo exigiriam o uso de outras técnicas de ajustes finos em modelos e treinamentos, que não foram abordadas por não se encaixarem no objetivo deste trabalho.

[2] A avaliação dos modelos, sempre que aplicada, foi realizada com um conjunto disjunto ao conjunto de treinamento.

Algorithm 2: Algoritmo seguido para implementação dos modelos.

```
dados ← conjunto de dados das Portarias;
dados_anotados ← ∅;
modelo_inicial ← pt_core_news_sm;
modelo_atual ← modelo_inicial;
repeat
    dados_anotados.add(Anotar(dados ∩ dados_anotados, modelo_atual));
    modelo_atual ← Treinar(modelo_inicial, dados_anotados);
    Avaliar(modelo_atual);
until acurácia diminuir ou estabilizar com mais dados anotados;
```

5.4.4 Resultados

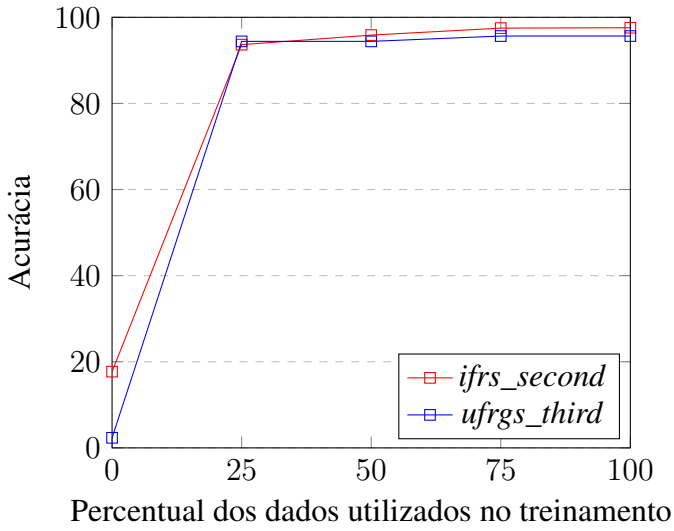
A Tabela 5.2 ilustra as métricas extraídas durante o processo de desenvolvimento do modelo *ifrs_second*, utilizado para a fonte de dados DOCS-IFRS. A primeira linha da tabela, referente ao modelo genérico *pt_core_news_sm*, indica a eficácia do modelo sem treinamento com dados específicos de Portarias. A segunda linha indica a eficácia do modelo *ifrs_first*, resultado do treinamento do modelo genérico com 204 exemplos de Portarias anotadas. A última linha da tabela indica, por fim, a eficácia do modelo *ifrs_second*, resultado do treinamento do modelo genérico com 418 exemplos de Portarias anotadas, os quais incluem as 204 Portarias anotadas utilizadas no treinamento do modelo anterior.

Tabela 5.2 – Avaliação dos modelos de Reconhecimento de Entidades Nomeadas para os documentos do IFRS

Modelo	Precisão	Revocação	Medida F1	# Exemplos
pt_core_news_sm	11,1%	41,1%	17,5%	0
ifrs_first	85,2%	87,8%	86,5%	204
ifrs_second	87,9%	92,2%	90,0%	418

Na Tabela 5.3, encontram-se as métricas extraídas durante o desenvolvimento do modelo *ufrgs_third*, utilizado para a fonte de dados DOCS-UFRGS. Da mesma forma como mencionado anteriormente, a primeira linha indica a eficácia do modelo *pt_core_news_sm* quando utilizado sem treinamento para o reconhecimento de entidades nomeadas nas Portarias da fonte DOCS-UFRGS, as linhas 2 e 3 indicam métricas de

Figura 5.1 – Curva de treinamento dos modelos escolhidos para cada fonte de dados



modelos intermediários *ufrgs_first* e *ufrgs_second* e a linha 4 indica métricas do modelo final utilizado, *ufrgs_third*.

Tabela 5.3 – Avaliação dos modelos de Reconhecimento de Entidades Nomeadas para os documentos da UFRGS

Modelo	Precisão	Revocação	Medida F-1	# Exemplos
pt_core_news_sm	1,0%	10,4%	1,9%	0
ufrgs_first	64,7%	71,4%	67,9%	100
ufrgs_second	65,1%	72,7%	68,7%	204
ufrgs_third	70,0%	72,7%	71,3%	534

O Gráfico da Figura 5.1 ilustra a estabilidade da acurácia no treinamento dos modelos definitivos (*ifrs_second* e *ufrgs_third*) com 75% e 100% dos dados de treinamento, evidenciando o momento de parada da anotação de dados. A diferença de acurácia entre os modelos sem treinamento e os modelos definitivos treinados com apenas 25% dos dados anotados foi de +428.7% para o modelo treinado para a fonte DOCS-IFRS e de +3934.6% para o modelo treinado para a fonte DOCS-UFRGS.

Além disso, a Medida F1 do modelo *ifrs_second* (Tabela 5.2) mostrou-se maior que a Medida F1 do modelo *ufrgs_third* (Tabela 5.3), apesar do menor número de exemplos. Uma justificativa é o baixo volume de entidades nomeadas da fonte de dados *DOCS-IFRS*, que permite ao modelo *ifrs_second* aprender os nomes dos servidores, e não as meta-informações sobre o reconhecimento em si das entidades, como capitalização, posição na frase ou identificação nos termos da oração (i.e. sujeito, objeto, entre outros).

5.4.5 Análise dos Casos de Falha

Após o reconhecimento de entidades nomeadas foi feita uma análise visual das entidades encontradas, evidenciando alguns casos de falha dos modelos. As principais falhas identificadas foram:

- Nomes incompletos. Apesar dos modelos serem treinados para identificar o nome completo dos servidores, ocorreram alguns casos onde o objetivo foi atingido apenas parcialmente. Como acontece na fonte de dados DOCS-UFRGS, onde o nome da servidora Carla Maria Dal Sasso Freitas foi identificado apenas parcialmente como MARIA DAL SASSO FREITAS em algumas Portarias.
- Nomes identificados como pertencentes a uma só entidade quando na verdade referem-se a duas ou mais. Este caso pode acontecer devido à estruturação das Portarias nas etapas anteriores, que podem acabar por remover quebras de linha utilizadas, por vezes, como delimitadores de término de um nome (e início de outro) em uma lista. Como acontece na fonte de dados DOCS-UFRGS, onde a entidade nomeada reconhecida foi ”NAIRA MARIA BALZARETTI RENATA JENISCH BARBOSA TANIRA“ porém ela se refere, na verdade, a três entidades: Naira Maria Balzaretti, Renata Jenisch Barbosa e Tanira Rodrigues Soares.
- Identificação de nomes de empresas prestadoras de serviço. Isto pode ocorrer devido ao modelo aprender a identificar os servidores como objetos nas sentenças, o que acaba por enganá-lo a induzir que o objeto de uma frase de uma Portaria é sempre um servidor, quando na verdade não é. Como acontece na fonte de dados DOCS-UFRGS, onde foi encontrado e identificado como entidade nomeada, em diversos documentos, o termo ”MEGATRON ENGENHARIA LTDA“.

Todos os casos listados anteriormente constituem Falsos Positivos neste experimento, dado que não correspondem ao nome completo de um servidor.

5.5 Experimento 3 - Resolução de Entidades

Este experimento tem como objetivo avaliar estratégias de resolução das entidades nomeadas para entidades do mundo real (servidores). Dessa forma, o foco do experimento é avaliar a eficácia da abordagem proposta para agrupar todas as menções a um servidor identificadas nos mais diversos documentos recuperados.

5.5.1 Métricas

As métricas escolhidas para avaliação da resolução de entidades tomam como base a comparação em pares, onde cada par indica um relacionamento entre entidades (de forma a referenciar uma mesma entidade no mundo real). Para um agrupamento de entidades $A = \{entidadeX, entidadeY, entidadeZ\}$ são gerados os pares $A_p = \{\{entidadeX, entidadeY\}, \{entidadeX, entidadeZ\}, \{entidadeY, entidadeZ\}\}$ e, a partir dos pares, são calculados os dados e métricas de avaliação. As métricas aqui descritas são representadas em função de três valores, sendo eles:

- Verdadeiros Positivos - quantidade de pares que realmente se referem a mesma entidade no mundo real.
- Falsos Positivos - quantidade de pares que foram identificados como referências a uma mesma entidade no mundo real, porém não são.
- Falsos Negativos - quantidade de pares que não foram identificados como referências a uma mesma entidade no mundo real, porém são.

Para este experimento, as métricas utilizadas foram as seguintes:

- Precisão - o percentual de pares identificados que realmente se referem à uma mesma entidade no mundo real. A fórmula da precisão pode ser descrita por $VerdadeirosPositivos/(VerdadeirosPositivos + FalsosPositivos)$.
- Revocação - o percentual de todos os pares que se referem à uma mesma entidade que foram identificados corretamente. A fórmula da revocação pode ser descrita por $VerdadeirosPositivos/(VerdadeirosPositivos + FalsosNegativos)$.
- Medida F1, uma média ponderada que utiliza Precisão (P) e Revocação (R) para, através de uma só métrica, medir a eficácia da abordagem. A fórmula da F1 pode ser descrita por $2 * P * R/(P + R)$.

5.5.2 Ambiente de Configuração

Este experimento foi realizado em um Macbook Pro mid-2014 com processador 2,8 GHz Intel Core i5 Dual-Core, 8Gb de Memória, armazenamento SSD e sistema operacional macOS versão 10.15.5.

5.5.3 Metodologia

Para este experimento, foram resolvidos 194 mil registros encontrados nas Portarias referentes a fonte de dados DOCS-UFRGS utilizando diferentes funções de *match* para que fosse possível identificar o impacto da variação da função na eficácia do processo de resolução de entidades.

As funções de *match* escolhidas foram:

1. A igualdade dos nomes presentes nos registros.
2. A igualdade dos nomes presentes nos registros com tratamento para redução de sequências de espaços em branco e descapitalização do nome.
3. A presença, em ambos os registros, de somente uma matrícula SIAPE e a igualdade das mesmas. Caso exista uma única matrícula SIAPE em cada registro mas elas sejam diferentes, a função de *match* retorna falso. Caso existam múltiplas matrículas SIAPE em pelo menos um registro, é ignorada a matrícula SIAPE e é realizada a comparação de nomes conforme o item anterior.

Para a avaliação, foram identificadas, manualmente, todas as referências aos 24 professores titulares do Instituto de Informática [3] da Universidade Federal do Rio Grande do Sul e gerados os respectivos pares indicando a relação entre os registros. A partir dos pares gerados para os registros que constituem a mesma entidade, foram extraídas as métricas de eficácia da resolução para a respectiva função de *match*.

5.5.4 Resultados

A Tabela 5.4 indica os resultados dos experimentos para os diferentes tipos de função de *match*. Pode-se observar que a precisão foi mantida próxima a 100% e diminuiu conforme a complexidade da função de *match*. Isso acontece porque as funções utilizadas foram conservadoras, de forma que somente agrupavam registros com mesmo nome ou nomes iguais porém com formatações diferentes (quantidade de espaços e / ou capitalização). Conforme a função de *match* torna-se mais complexa, a tendência é de que a precisão caia em detrimento da revocação, que aumenta dada a flexibilidade da condição de agrupamento dos registros.

A abordagem selecionada foi a número 3, que utiliza da matrícula SIAPE (quando

[3] Disponível em <https://www.inf.ufrgs.br/site/pessoas/corpo-docente/>. Último acesso em 10/07/2020.

Tabela 5.4 – Avaliação da eficácia das funções de *match* da resolução de entidades

Função de *match*	Precisão	Revocação	F1
1 - Nomes originais	100,0%	59,2%	74,4%
2 - Nomes processados	100,0%	75,5%	86,0%
3 - SIAPE + Nomes processados	99,5%	82,4%	90,1%

disponível) e os nomes pré-processados, com uma medida F1 de 90,1%. A função de *match* 3 demonstrou uma melhora de 9% na revocação, sem afetar drasticamente a precisão, que reduziu somente 0,5%.

5.5.5 Análise dos Casos de Falha

Apesar da alta precisão obtida (Tabela 5.4), a revocação da resolução de entidades não teve uma eficácia tão elevada. Isso acontece porque a definição de pertencimento de um registro a um agrupamento é muito restrita, impedindo que casos como os descritos a seguir sejam identificados e agrupados conforme o esperado. São os principais exemplos de casos não agrupados utilizando as funções descritas nesse experimento:

- Nomes identificados parcialmente em um registro e completamente em outros. Conforme descrito no experimento da Seção 5.4, podem ocorrer registros com identificações parciais dos nomes das entidades, que não seriam agrupadas pelas funções de *match* utilizadas na abordagem ACERPI.
- Nomes com prefixo ou sufixo identificados erroneamente. Casos com registro de nome “LUCIANE MACHADO CAETANO MOSSMANN-”, da fonte de dados DOCS-UFRGS, não seriam agrupados com os registros de nome “LUCIANE MACHADO CAETANO MOSSMANN”, o que é caracterizado como uma falha na resolução de entidades da abordagem descrita.
- Nomes escritos de maneira diferente mas que fazem referência a uma mesma entidade no mundo real. Por tratarem-se de documentos escritos por pessoas diferentes e serem suscetíveis a erros de escrita, pode acontecer de um mesmo nome ser escrito de formas diferentes, como é o caso do servidor Sérgio Bampi, que possui referências em documentos da fonte DOCS-UFRGS tanto como “SÉRGIO BAMPI” quanto como “SERGIO BAMPI”. Estes registros não seriam agrupados, o que é caracterizado, também, como uma falha na resolução de entidades da abordagem.

Os casos de falha descritos assumem que não foi possível resolver as entidades

dos registros baseado nos números de matrícula SIAPE.

5.6 Considerações Finais

Neste capítulo, foram apresentados os três experimentos realizados durante o desenvolvimento da abordagem ACERPI. O primeiro, destacando as técnicas de *Web Scraping* utilizadas para coleta das fontes de dados utilizadas. O segundo, avaliando métodos de reconhecimento de entidades nomeadas a partir de modelos de processamento de linguagem natural. E o terceiro, avaliando diferentes estratégias de identificação de duplicatas nas entidades identificadas nas Portarias.

6 CONCLUSÃO

Neste trabalho, foi apresentada uma abordagem para processamento de Portarias de Instituições que, ao final, inclui a criação de uma base de dados não relacional para pesquisa a respeito de servidores e Portarias. A abordagem foi avaliada em duas fontes de dados. A primeira foi a fonte de dados de Portarias da Universidade Federal do Rio Grande do Sul, com mais de 40 mil arquivos, cada um com exatamente uma Portaria, e milhares de servidores mencionados. A segunda foi a fonte de dados do Instituto Federal do Rio Grande do Sul, *Campus* Ibirubá, com aproximadamente 300 arquivos, cada um com uma ou mais Portarias, e centenas de servidores mencionados. Ainda, foram realizados experimentos durante os processos de avaliação da abordagem. Esses experimentos permitiram medir a eficácia dos métodos escolhidos para diferentes parâmetros e fontes de dados.

Foram identificados diversos pontos de melhoria e evolução que podem ser considerados possíveis trabalhos futuros. Desses, destacam-se:

- O aperfeiçoamento do modelo de reconhecimento de entidades nomeadas, permitindo uma maior eficácia no reconhecimento e redução do trabalho posterior em resolver as entidades reconhecidas.
- O aperfeiçoamento da função de *match* utilizada na resolução de entidades, de forma a resolver os casos de falha citados nos experimentos relacionados.
- O uso de uma técnica de blocagem (PAPADAKIS et al., 2020) para reduzir o número de comparações necessárias durante a resolução de entidades. O uso de uma técnica de blocagem seria muito relevante para a aplicação da abordagem em um grande volume de documentos, como ocorre na fonte de dados da Universidade Federal do Rio Grande do Sul, e para permitir o uso de funções de *match* com um custo mais elevado, porém de maior acurácia.
- A construção de uma interface gráfica para a pesquisa avançada a respeito dos servidores e Portarias. Isso facilitaria o uso e disponibilização das informações para pessoas não técnicas que tivessem interesse em realizar pesquisas.
- A classificação do conteúdo das Portarias em categorias conhecidas, como progressão funcional, afastamento, aposentadoria e substituição de função. Isso permitiria buscas ainda mais avançadas a respeito de informações dos servidores, bem como buscar por Portarias de um determinado tipo em um determinado período.

Por fim, a principal contribuição deste trabalho foi a criação de um *pipeline* de processamento dos dados de Portarias flexível, com etapas intermediárias independentes, com a definição de estruturas intermediárias extensíveis, explicação das escolhas e ferramentas utilizadas, bem como a avaliação da implementação em diferentes fontes de dados.

REFERÊNCIAS

AWS. O que é NoSQL? 2020. Disponível em: <https://aws.amazon.com/pt/nosql/>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval: The Concepts and Technology behind Search**. 2nd. ed. USA: Addison-Wesley Publishing Company, 2011. ISBN 9780321416919.

BLANCO, L. et al. Supporting the automatic construction of entity aware search engines. In: **Proceedings of the 10th ACM Workshop on Web Information and Data Management**. New York, NY, USA: Association for Computing Machinery, 2008. (WIDM '08), p. 149–156. ISBN 9781605582603. Disponível em: <https://doi.org/10.1145/1458502.1458526>.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. regula o acesso a informações previsto no inciso xxxiii do art. 5º , no inciso ii do § 3º do art. 37 e no § 2º do art. 216 da constituição federal; altera a lei nº 8.112, de 11 de dezembro de 1990; revoga a lei nº 11.111, de 5 de maio de 2005, e dispositivos da lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2011. ISSN 1677-7042. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>.

Data Community DC. Entity Resolution for Big Data. August 2013. Disponível em: <http://www.datacommunitydc.org/blog/2013/08/entity-resolution-for-big-data>.

DB-ENGINES. DB-Engines Ranking. September 2020. Disponível em: <https://db-engines.com/en/ranking>.

DOZIER, C. et al. Named entity recognition and resolution in legal text. In: ____. **Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language**. Berlin, Heidelberg: Springer-Verlag, 2010. p. 27–43. ISBN 364212836X.

Explosion.ai. Entity Linking. 2020. Disponível em: <https://spacy.io/usage/linguistic-features#entity-linking>.

Explosion.ai. Features. 2020. Disponível em: <https://spacy.io/usage/spacy-101#features>.

Explosion.ai. Industrial-Strength Natural Language Processing. 2020. Disponível em: <https://spacy.io/>.

Explosion.ai. Named Entity Recognition. 2020. Disponível em: <https://spacy.io/usage/linguistic-features#named-entities>.

Explosion.ai. Prodigy · An annotation tool for AI, Machine Learning NLP. 2020. Disponível em: <https://prodi.gy/>.

Explosion.ai. Tokenization. 2020. Disponível em: <https://spacy.io/usage/linguistic-features#tokenization>.

IEEE. IEEE Xplore. 2020. Disponível em: <https://ieeexplore.ieee.org/Xplore/home.jsp>.

IFRS. Documentos. 2020. Disponível em: <https://ifrs.edu.br/documentos/>.

IFRS, Campus Ibirubá. Boletins de Serviço. 2011. Disponível em: <https://ibiruba.ifrs.edu.br/site/conteudo.php?cat=50>.

IFRS, Campus Ibirubá. Boletim de Serviço. 2018. Disponível em: <https://ifrs.edu.br/ibiruba/documentosoficiais/boletim-de-servico/>.

MANICA, E.; DORNELES, C. F.; GALANTE, R. Orion: A cypher-based web data extractor. In: BENSLIMANE, D. et al. (Ed.). **Database and Expert Systems Applications**. Cham: Springer International Publishing, 2017. p. 275–289. ISBN 978-3-319-64468-4.

MONGODB. The database for modern applications. 2020. Disponível em: <https://www.mongodb.com/>.

MONGODB. What is NoSQL. 2020. Disponível em: <https://www.mongodb.com/nosql-explained>.

OLIVAS, E. S. et al. **Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes**. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2009. ISBN 1605667668.

Palmieri Lage, J. et al. Automatic generation of agents for collecting hidden web pages for data extraction. **Data Knowledge Engineering**, v. 49, n. 2, p. 177 – 196, 2004. ISSN 0169-023X. Web Information and Data Management. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169023X03001769>.

PAPADAKIS, G. et al. Blocking and filtering techniques for entity resolution: A survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 2, mar. 2020. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3377455>.

PIVETTA, S. P. Classificação de documentos do exército brasileiro utilizando o classificador naive bayes e técnicas de seleção de sentenças. Alegrete, Rio Grande do Sul, Brasil, 2013. Disponível em: <http://oasisbr.ibict.br/vufind/Record/UNIP_fdeeac79eb001f280074dc2a5a5a64b9>.

SARKAR, D. Named Entity Recognition: A Practitioner's Guide to NLP. August 2018. Disponível em: <https://www.kdnuggets.com/2018/08/named-entity-recognition-practitioners-guide-nlp-4.html>.

ScrapingHub. What is web scraping? 2020. Disponível em: <https://www.scrapinghub.com/what-is-web-scraping/>.

The Apache Software Foundation. Apache PDFBox® - A Java PDF Library. 2020. Disponível em: <https://pdfbox.apache.org/>.

UFRGS. Consulta a Portarias geradas pela Reitoria da UFRGS. 2016. Disponível em: <https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/consultar/>.