



Universidade de São Paulo
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Faculdade de Medicina de Ribeirão Preto



**"Mineração de Textos para Auxílio ao Processo de
Estruturação da Informação contida em Laudos
Radiológicos"**

Luana Peixoto Annibal

Ribeirão Preto

2008



Universidade de São Paulo

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Faculdade de Medicina de Ribeirão Preto



"Mineração de Textos para Auxílio ao Processo de Estruturação da Informação contida em Laudos Radiológicos"

Monografia apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto e à Faculdade de Medicina de Ribeirão Preto, ambas da Universidade de São Paulo, como requisito da disciplina de Desenvolvimento de Projetos II.

Aluna: Luana Peixoto Annibal, luapatela@gmail.com

Orientador: Prof. Dr. Joaquim Cezar Felipe, jfelipe@ffclrp.usp.br

Co-Orientador: Prof. Dr. José Augusto Baranauskas, augusto@usp.br

Ribeirão Preto

2008

A melhor lição que pude aprender nestes quatro anos ...



*Don't worry about a thing,
Cause every little thing is gonna be alright*



if you work

Agradecimentos

Agradeço, inicialmente, a Deus por todas as conquistas de minha vida e tudo que tenho hoje.

Aos meus pais, Italo e Celia, por todo amor, carinho e apoio que estes me ofereceram nesta jornada. Por me ensinarem a acreditar em meu potencial e ensinarem que as dificuldades da vida são para nos tornar mais fortes.

Ao meu namorado, Gustavo, pelo apoio e conselhos, que fizeram com que esta etapa fosse vencida com mais leveza e tranquilidade.

Aos demais familiares e ao saudoso Rui, pela alegria e auxílio.

Aos meus inesquecíveis e eternos amigos, Gisele, Daniane, Diego, Carla, Juliana e Flávia, por toda a amizade, apoio e ajuda doada para que eu alcançasse mais esta vitória em minha vida. Também agradeço a estes e aos amigos: Caio, Cristina, Ian, Yuri, Lariza, Hugo, Maycon e Fábio, pela companhia e por todos os momentos de alegria e divertimento, que tornaram estes quatro anos inesquecíveis e incomparáveis.

Às amigas, Fernanda, Érika, Juliana e Angelina, por me acolherem em Ribeirão Preto e me ajudarem em todos estes quatro anos.

Às amigas eternas, Mariane e Lucimara, por, mesmo distantes, estarem sempre ao meu lado.

Ao meu exemplar orientador Joaquim, co-orientador Augusto e professor Renato, por tudo que me foi ensinado e toda a ajuda prestada para a realização deste trabalho e a conclusão de minha graduação.

À Universidade de São Paulo, por todo o conhecimento oferecido e ensinado. Assim como seus contribuintes e professores, por tornar isto acessível.

A toda turma III do curso de Informática Biomédica, pelos ótimos momentos de convivência e pelas oportunidades de aprendizado.

E à Fundação de Amparo a Pesquisa do estado de São Paulo pelo auxílio financeiro.

RESUMO

Um dos grandes desafios enfrentados pelos desenvolvedores de sistemas computacionais consiste em possibilitar uma maior adaptação e satisfação de seus usuários às inovadoras tecnologias voltadas à geração, recuperação, análise e gerenciamento de informação e conhecimento. Nota-se que, no contexto da área médica, os conflitos e dificuldades neste tipo de implementação são ainda mais elevados, devido ao perfil dos profissionais da saúde que se caracteriza por apresentar uma maior dificuldade no processo de migração tecnológica; e devido ao fato de anotações textuais simples, ou seja, textos abertos em linguagem natural, serem a prática médica mais freqüente e, geralmente, a única utilizada por parte destes profissionais no processo de descrição e diagnóstico do laudo do paciente. Conseqüentemente, seu armazenamento é feito de forma não estruturada, limitando, assim, a consulta à base de dados, minimizando seus resultados e impossibilitando a implantação de aplicativos baseados em inteligência artificial. O presente trabalho consistiu em investigar a aplicação de conceitos e técnicas de ontologia e mineração de texto (tais como algoritmos de *stopwords* e de *stemming*) sobre um conjunto de textos de descrição e diagnóstico presentes em laudos radiológicos com o objetivo de possibilitar a identificação de termos presentes nesse tipo de documento. Em seguida, foram desenvolvidas duas ferramentas, uma para o auxílio da estruturação de tais textos, e outra para a geração de um sistema de busca de laudos. Torna-se possível, dessa forma, que trabalhos futuros para o auxílio a diagnóstico sejam implementados, utilizando os termos estruturados nesta ferramenta como atributos em algoritmos de Aprendizado de Máquinas e/ou Inteligência Artificial.

SUMÁRIO

LISTA DE FIGURAS	I
LISTA DE TABELAS.....	II
LISTA DE ABREVIACÃO	III
CAPÍTULO 1. INTRODUÇÃO	1
1.1. CONTEXTUALIZAÇÃO	1
1.2. MOTIVAÇÃO.....	2
1.3. OBJETIVOS	3
1.4. ORGANIZAÇÃO DA MONOGRAFIA.....	3
CAPÍTULO 2. FUNDAMENTOS TEÓRICOS.....	4
2.1. CONSIDERAÇÕES INICIAIS	4
2.2. MINERAÇÃO DE TEXTO	4
2.3. PROCESSAMENTO DE LÍNGUA NATURAL - PLN.....	7
2.3.1. REMOÇÃO DE <i>STOPWORDS</i>	7
2.3.2. RADICALIZAÇÃO – <i>STEMMING</i>	9
2.4. TABELA ATRIBUTO-VALOR (<i>BAG-OF-WORDS</i>).....	13
2.5. <i>PRETEXT</i>	13
2.5.1. <i>PRETEXT</i> : UMA FERRAMENTA PARA PRÉ-PROCESSAMENTO DE TEXTOS UTILIZANDO A ABORDAGEM <i>BAG-OF-WORDS</i> (VERSÃO 1).....	13
2.5.2. <i>PRETEXT</i> : A REESTRUTURAÇÃO DA FERRAMENTA DE PRÉ-PROCESSAMENTO DE TEXTOS (VERSÃO 2)	16
2.6. ONTOLOGIA.....	17
CAPÍTULO 3: METODOLOGIA	21
3.1. CONSIDERAÇÕES INICIAIS	21
3.2. DOMÍNIO DE CONHECIMENTO DOS LAUDOS	21
3.3. ESTUDO DAS CARACTERÍSTICAS DE PREENCHIMENTO DOS LAUDOS	22
3.4. PRÉ-PROCESSAMENTO DOS LAUDOS.....	22
3.5. MODELAGEM DE UMA ONTOLOGIA	23
3.6. MODELAGEM E ALIMENTAÇÃO DA BASE DE DADOS	23
3.7. ESTRUTURAÇÃO DOS LAUDOS.....	24
CAPÍTULO 4. FERRAMENTA E-RAD.....	25
4.1. CONSIDERAÇÕES INICIAIS	25
4.2. MÓDULO <i>TAKEREPORT.PM</i>	27
4.3. MÓDULO <i>REMOVEACCENT.PM</i>	27
4.4. MÓDULO <i>START.PM</i>	28
4.5. MÓDULO <i>TAKETERMSENTENCE.PM</i>	29
4.6. MÓDULO <i>MAKESENTENCES.PM</i>	30
4.7. MÓDULO <i>E-RAD.PL</i>	30
4.8. INTERFACE E SISTEMA DE BUSCA.....	31
CAPÍTULO 5. RESULTADOS.....	33
5.1. CONSIDERAÇÕES INICIAIS	33
5.2. BREVE DEFINIÇÃO DO DOMÍNIO DE CONHECIMENTO DOS LAUDOS.....	34
5.2.1. IDENTIFICAÇÃO DAS CARACTERÍSTICAS DE PREENCHIMENTO.....	35
5.3. RESULTADOS DO PRÉ-PROCESSAMENTO DOS LAUDOS	36
5.4. RESULTADOS DA MODELAGEM DE UMA ONTOLOGIA	37
5.5. RESULTADO DA MODELAGEM E ALIMENTAÇÃO DA BASE DE DADOS	39
5.6. RESULTADOS DA FERRAMENTA <i>E-RAD</i>	42
5.6.1. RESULTADOS DO MÓDULO <i>TAKEREPORT.PM</i>	43
5.6.2. RESULTADOS DO MÓDULO <i>REMOVEACCENT.PM</i>	44
5.6.3. RESULTADOS DO MÓDULO <i>START.PM</i>	45
5.6.4. RESULTADOS DO MÓDULO <i>TAKETERMSENTENCE.PM</i>	45

5.6.5.	RESULTADOS DO MÓDULO <i>MAKESENTENCES.PM</i>	46
5.7.	RESULTADO DE UTILIZAÇÃO DA INTERFACE DE BUSCA DO <i>E-RAD</i>	48
5.7.1.	BUSCA UTILIZANDO UMA PALAVRA	49
5.7.2.	BUSCA COM UM CONJUNTO DE PALAVRAS.....	50
5.7.3.	BUSCA BASEADA EM UM CONCEITO.....	51
5.7.4.	BUSCA BASEADA EM DOIS CONCEITOS	52
CAPÍTULO 6. DISCUSSÃO E CONCLUSÕES		54
6.1.	CONSIDERAÇÕES FINAIS.....	54
6.2.	DISCUSSÃO.....	55
6.3.	CONCLUSÃO.....	56
REFERÊNCIAS BIBLIOGRÁFICAS.....		58
APÊNDICES.....		61
	APÊNDICE 1 – ARQUIVO DE CONFIGURAÇÃO DA FERRAMENTA <i>PRETEXT</i>	62
ANEXOS		63
	ANEXO A – CONJUNTO DE PALAVRAS QUE FORAM REMOVIDAS DO ARQUIVO <i>STOPLIST</i>	64

LISTA DE FIGURAS

Figura 1: Principais etapas da Mineração de Texto.....	5
Figura 2: Variáveis de comprimento mínimo da palavra	10
Figura 3: Fluxograma de execução da ferramenta <i>PreText</i> versão 1.....	16
Figura 4: Fluxograma de execução da ferramenta <i>PreText</i> versão 2.....	17
Figura 5: Comparação entre categorização em níveis <i>versus</i> listagem de todos os vinhos e tipos.....	19
Figura 6: Fluxograma de execução da ferramenta <i>E-Rad</i>	26
Figura 7: Diagrama de classe do módulo <i>Start.pm</i> e os módulos que este se relaciona	29
Figura 8: Diagrama de classe da ferramenta <i>E-Rad</i>	31
Figura 9: Interface para busca por laudos estruturados	32
Figura 10: Visão parcial da ontologia proposta.....	39
Figura 11: Modelagem da base de dados <i>RadOn</i> , proposta para armazenar de forma estruturada os laudos e diagnósticos médicos.....	42
Figura 12: Resultado do módulo <i>TakeReport.pm</i>	44
Figura 13: Resultado do módulo <i>RemoveAccent.pm</i>	44
Figura 14: Resultado do módulo <i>Start.pm</i>	45
Figura 15: Resultados do módulo <i>TakeTermSentence.pm</i>	46
Figura 16: Resultado do módulo <i>MakeSentences.pm</i>	48
Figura 17: Resultado da busca por laudo que relatam a ocorrência de cisto	49
Figura 18: Busca para teste do módulo <i>MakeTermSentence.pm</i>	50
Figura 19: Resultado de busca pelo conceito “hipointensidade”.....	52
Figura 20: Resultado da busca por “LCA com forma preservada”	53

LISTA DE TABELAS

Tabela 1: Conjunto de stopwords da ferramenta PreTextT	8
Tabela 2: Regras de remoção de sufixos da língua portuguesa	11
Tabela 3: Terminações verbais do Português	12
Tabela 4: Tabela atributo-valor	13
Tabela 5: Exemplo de laudo radiológico	33
Tabela 6: Laudo Modelo para o exame de Ressonância Magnética de Joelho	35
Tabela 7: Exemplo de laudo radiológico utilizado para demonstrar os resultados da ferramenta E-Rad	43
Tabela 9: Arquivo de configuração da ferramenta <i>PreTextT</i>	62
Tabela 10: Palavras desconsideradas <i>stopwords</i>	64

LISTA DE ABREVIACÃO

CCIFM - Centro de Ciências das Imagens e Física Médica

EI - Extração de Informação

HCFMRP – Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto

KDD – *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bases de Dados)

KDT - *Knowledge Discovery in Text* (Descoberta de Conhecimento em Texto)

LCA - Ligamento Cruzado Anterior

PEP - Prontuário Eletrônico da Paciente

PLN - Processamento de Língua Natural, ou também denominado por Processamento de Linguagem Natural

RI - Recuperação de Informação

SIH - Sistemas de Informação Hospitalar

SIR - Sistema de Informação Radiológica

CAPÍTULO 1. Introdução

1.1. Contextualização

Nas últimas décadas, observa-se um crescente processo de informatização dos procedimentos hospitalares baseados nos chamados Sistemas de Informação Hospitalar (SIH). Esse sistema consiste em um conjunto de aplicativos inter-relacionados para realizar coletas, armazenamentos, possíveis processamentos e distribuição de dados e informações [1] com o objetivo de fornecer suporte às necessidades de uma organização hospitalar.

Dentro deste contexto, o Prontuário Eletrônico do Paciente (PEP) é um dos principais temas de pesquisa e desenvolvimento no âmbito da Informática Médica. Esta, enquanto campo da ciência, é voltada à resolução de problemas e auxílio na tomada de decisão pelo processamento e gerenciamento de informações [2].

O desenvolvimento de um PEP visa melhorar a eficiência e a organização do armazenamento das informações de saúde e não, tão somente, substituir o prontuário em papel, mas, principalmente, prover novos recursos e aplicações do conhecimento para elevar a qualidade da assistência à saúde do paciente [3].

Atualmente, os Sistemas de Informação Hospitalares consistem-se, em sua maioria, de aplicativos para o armazenamento de dados, abrindo-se mão de ferramentas de recuperação da informação ou descoberta do conhecimento baseadas em Inteligência Artificial ou Aprendizado de Máquina. Tais dados são considerados elementos puros e quantificáveis em determinado evento. Podem ser fatos, números, textos ou qualquer elemento que possa ser processado pelo computador. No entanto, os dados por si só não oferecem informação e, muito menos, conhecimento para o entendimento da situação [4].

Em contrapartida aos sistemas atuais, idealiza-se que SIH possua informações, em sua maioria, dos elementos armazenados, pois estas consistem em um conjunto de dados organizados de forma compreensível, registrado em papel ou em outro meio, e suscetível de ser comunicado e conservado para a transmissão do conhecimento obtido a partir de sua análise [5] e [6].

Ou seja, a informação é a interpretação de um conjunto de dados, padrões ou associações, passível de ser transmitida. É desejado possuir um SIH repleto de informações, pois informações podem gerar conhecimento, auxiliando na análise de padrões históricos e realizando previsões dos fatos com maiores probabilidades de certeza.

O Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP) possui, em seu centro de radiologia, um sistema informatizado - Sistema de Informação Radiológica (SIR), responsável pela informatização dos processos que envolvem a aquisição e o controle das imagens e exames, além de ser utilizado por médicos docentes e residentes para visualizar imagens digitais e gerar laudos.

O HCFMRP conserva, de forma eletrônica, cerca de 660.000 exames radiológicos. Contudo, os laudos lá gerados não estão atingindo a total funcionalidade de um SIH, uma vez que consiste em simples campos textuais nos quais o radiologista registra suas observações em linguagem natural, apresentando uma série de limitações para os processos computacionais de auxílio à recuperação, processamento da informação e auxílio ao diagnóstico.

1.2. Motivação

Como dito na seção anterior, o HCFMRP possui em seu centro de radiologia um sistema informatizado utilizado por seus profissionais da saúde para visualização de imagens obtidas a partir de exames radiológicos e para geração de seus respectivos laudos médicos. Todavia, esses laudos informatizados consistem em simples campos textuais em que o radiologista registra suas observações em língua natural - também denominada por linguagem natural -, esse tipo de dado dificulta, relevantemente, a implantação de processos computacionais de auxílio à recuperação, análise da informação e obtenção de padrão de achados radiológicos para futuros sistemas de auxílio ao diagnóstico.

Textos abertos, tais como as descrições e diagnósticos de laudos de pacientes, são classificados como tipos de dados não estruturados. Estrutura é a organização de um conjunto de elementos, propriedades e relações, podendo ser representado como um conjunto de elementos em interação, que realizam determinadas funções para determinados propósitos [7].

Dados não estruturados são pouco adequados aos algoritmos de Aprendizado de Máquina e Inteligência Artificial, pois, além de não viabilizarem o resgate de informação, concomitantemente, esses dados são apresentados de forma desorganizada em relação à exposição de seus componentes (ou seja, não há extração ou distinção computacional entre as partes) e, principalmente, não há extração ou distinção entre os elementos chaves/diferenciais.

Sendo assim, o pré-processamento de textos e sua representação em uma forma inteligível aos sistemas, tal como em uma tabela atributo-valor, é tida por Sebastiani [8] como uma tarefa de fundamental influência no desempenho destes algoritmos.

Contudo, a intenção deste trabalho não é estruturar os textos empregando, somente, a listagem de seus termos em tabela (tabela atributo-valor), pois a estruturação de um texto deve manter os relacionamentos entre os termos a fim de preservar a informação existente nas sentenças. Ou seja, ao armazenar a presença dos termos: ligamento cruzado anterior, patela, normal, osteofítos, orientação, não é determinada em que região há a ocorrência de osteofítos, a que estrutura a característica orientação está relacionada ou mesmo que estrutura está normal.

A estruturação elaborada de textos, mantendo o relacionamento entre os termos existentes em uma mesma sentença, também pode servir como base de dados em sistemas de busca para laudos, uma vez que a semântica das sentenças de busca e a semântica das sentenças dos laudos são mantidas e comparadas, em vez de, tão somente, verificar a ocorrência de termos em um texto. Dessa forma, aumentar-se-ia a compatibilidade do resultado em relação a sua respectiva sentença de busca.

1.3. Objetivos

Este trabalho pretende permitir que os profissionais da saúde possam continuar descrevendo e concluindo os laudos com texto aberto e armazenando a informação destes de forma estruturada e imperceptível ao usuário. Isso é possível a partir de uma metodologia de filtragem da informação relevante, aplicando conceitos e técnicas de mineração de texto – algoritmos de *stopwords* e de *stemming* –, além do desenvolvimento de uma ontologia buscando identificar termos e suas possíveis semânticas.

O desenvolvimento deste trabalho possibilitará que futuras pesquisas relacionadas a sistemas de buscas complexas e sistemas de auxílio ao diagnóstico sejam realizadas.

1.4. Organização da Monografia

O presente trabalho está organizado da seguinte forma: o Capítulo 2 descreve os fundamentos teóricos; o Capítulo 3 relata a metodologia deste trabalho; o Capítulo 4 descreve a ferramenta *E-Rad*; Capítulo 5 apresenta os resultados obtidos e o Capítulo 6 apresenta as conclusões e discussões.

CAPÍTULO 2. Fundamentos Teóricos

2.1. Considerações Iniciais

Neste capítulo são abordados os principais conceitos que permeiam a base teórica deste trabalho. É descrito a técnica de Mineração de Texto, bem como os princípios de Ontologia e dos algoritmos de mineração, aqui denominados *PreText* versão 1 e *PreText* versão 2.

Ambos os algoritmos *PreText*, foram desenvolvidos no Instituto de Ciências Matemáticas e de Computação, na linguagem *Perl*. A versão 1, por Edson Takashi Matsubara, Claudia Aparecida Martins, Maria Carolina Monard em 2003 e titulada por *PreText*: uma ferramenta para pré-processamento de textos utilizando a abordagem *bag-of-words*. E a versão 2, por Matheus Victor Brum Soares, Ronaldo C. Prati, Maria Carolina Monard, em 2008 e titulada por *PreText*: A Reestruturação da Ferramenta de Pré-Processamento de Textos.

Essas ferramentas foram escolhidas, pois suas características pesaram de forma favorável aos objetivos referidos no trabalho, destacando o fato de essas ferramentas possuírem a capacidade de pré-processar textos em três línguas distintas: Espanhol, Inglês e Português; e por realizarem o processo de *Stemming* (a ser explicado na seção 2.3.2).

2.2. Mineração de Texto

A informatização de documentos de diversas áreas nem sempre garante que estes possam ser recuperados de forma mais rápida e mais eficiente em comparação à busca manual. Conseqüentemente, teorias e ferramentas computacionais têm sido desenvolvidas para realizar análises automáticas em texto, visto que a sobrecarga de informação disponível dificultava sua análise manual, localização e acesso.

Inicialmente foram desenvolvidas técnicas de extração de informações úteis presentes em Banco de Dados, originando assim a área de Descoberta de Conhecimento de Banco de Dados (*Knowledge Discovery in Data – KDD*), ou como também pode ser denominado, Mineração de Dados. Aplicando-as em um grande número de dados, observou-se que eram gerados padrões e tendências na busca. Esse fato acentua a diferença entre a busca manual e a computadorizada.

Em seguida, as técnicas de Descoberta do Conhecimento foram aplicadas em dados unicamente textuais. Contudo, devido ao fato desses dados serem geralmente armazenados em registros não-estruturados, era impossibilitada a aplicação dos métodos existentes no banco de dados, visto que tais métodos necessitavam de registros estruturados para serem executados. Sendo assim, novas técnicas de busca voltadas a registros não estruturados foram desenvolvidas, originando a área de Descoberta de Conhecimento em Texto (*Knowledge Discovery in Text – KDT*), ou também denominado por Mineração de Texto (*Text Mining*).

Essas tarefas de KDT são realizadas utilizando abordagens das áreas: Recuperação de Informação, Processamento de Língua Natural e Descoberta de Conhecimento em Banco de Dados. Mesmo que essas tradicionais técnicas aplicadas em texto gerem informações sem utilidade para o usuário, seus conceitos foram reaproveitados para a manipulação de textos [9].

A Recuperação de Informação tem como intuito identificar os melhores meios de armazenamento e localização das informações [10]. A área de Processamento de Língua Natural foca a análise de texto, identificando as classes morfosintáticas existentes dentro dele. A Descoberta de Conhecimento em Banco de Dados é composto pela técnicas de Extração de Informação e Mineração de Dados, a primeira, com a pré-definição de *slots* (itens de dados formados por pares: atributo-valor), permite com que informações sejam obtidas. A segunda é definida como o campo de pesquisa que soluciona os problemas relacionados à descoberta de informação implícita em banco de dados [11].

A aplicação de todas essas áreas resulta nas quatro principais etapas de uma Mineração de Texto: Coleta de Documentos, Pré-processamento, Mineração de dados e por fim, Avaliação e Interpretação dos Resultados.



Figura 1: Principais etapas da Mineração de Texto

A Mineração de Texto é iniciada com a obtenção de documentos relevantes ao domínio de aplicação do conhecimento extraído, com a etapa de Coleta de Documentos. O sucesso dessa etapa é alcançado com a contribuição de um especialista, pois além de viabilizar o conhecimento sobre o domínio, ele também auxilia a tarefa de encontrar os objetivos almejados [12].

O Pré-processamento ocorre com o intuito de converter os documentos desestruturados em uma forma estruturada, na maioria das vezes, em uma tabela atributo-valor. Contudo, ao término de sua criação, observa-se que essa tabela possui uma alta dimensionalidade, uma vez que cada termo, existente em algum documento, pode ser indexado como um elemento do conjunto de atributos da tabela. Uma alternativa para a redução da dimensionalidade consiste na aplicação de técnicas de Recuperação de Informação (RI) ou Extração de Informação (EI), as quais identificando um conjunto de palavras-chaves, juntamente com seus pesos e suas frequências associadas [13].

De acordo com Dixon [14], a Recuperação de Informação é a etapa responsável, na Mineração de Dados, por localizar e recuperar documentos relativamente importantes, utilizando um filtro para selecionar os documentos especificados pelo usuário, além de indexar os documentos filtrados com palavras-chave que os caracterizam.

Outra etapa existente no processo de Mineração de Dados é a Mineração de Informação. Ela se resume em encontrar padrões de agrupamento entre documentos com os mesmos conceitos e ainda na possibilidade de extrair relacionamentos entre os documentos a partir da execução de algoritmos específicos [15].

Finalizando as etapas existentes no processo de Mineração, há a Interpretação, a qual possui o objetivo de interpretar os padrões resultantes da etapa de mineração, desenvolvendo, como consequência, a descoberta do conhecimento.

As técnicas existentes para se obter a Descoberta do Conhecimento são diversas. A Sumarização visa encontrar palavras ou frases para sumarizar o conceito dos documentos ou de um conjunto de documentos. Outra técnica é a Regra de Associação, ou seja, se há a existência de um determinado atributo X no banco, então Y tende a existir. A Clusterização que coloca em grupos os documentos similares. Há, também, a Classificação, a qual utiliza um treinador que aprende as regras e as aplica em classes pré-definidas [15].

Todas essas etapas e processos são realizados com algoritmos baseados em princípios de três áreas: Aprendizado de Máquina: algoritmos desenvolvidos com o intuito de fazer com que o computador “aprenda” a tomar decisões diferentes baseado em exemplos estudados anteriormente; Estatística: “Ramo da matemática que estuda as técnicas de inferência dos

fenômenos previsíveis, através de amostragem do todo” e; Mineração de Dados: objetiva-se a descobrir estrutura de dados [14].

2.3. Processamento de Língua Natural - PLN

Segundo Allen J. [16], o Processamento de Língua Natural caracteriza-se pela análise e manipulação ou codificação de informações expressas em língua natural. Com o objetivo de alcançar um melhor entendimento sobre a língua através do uso de computadores. Por exemplo, uma grande quantidade de textos em língua natural pode ser mais compreensível e útil quando sumarizada. Técnicas lingüísticas, estatísticas e manipulação de cadeias de caracteres (*strings*) podem ser empregadas utilizando este processamento automático de textos.

Esta técnica de processamento textual também é compreendida como uma tarefa complexa que possui a finalidade de determinar a relação entre as palavras das sentenças para extrair o significado destas.

O entendimento da linguagem natural pode se dar em vários níveis [17].

- Nível Morfológico: estudo da constituição das palavras em elementos básicos;
- Nível Sintático: determinação da relação (papéis) de um conjunto de palavras em uma sentença;
- Nível Semântico: determinação do significado e inter-relacionamento semântico das palavras;
- Nível Discursivo: objetiva-se em determinar o significado de um conjunto de sentenças;
- Nível Pragmático: Visa determinar o objetivo do uso da língua.

Quanto maior o nível de entendimento, mais conhecimento sobre a linguagem se faz necessário. Ferramentas PLN são muito utilizadas nas etapas de pré-processamento e modelagem dos textos, dentre os tipos de ferramentas aplicadas, as mais comuns são os sentenciadores, os tokenisadores, a remoção de *stopwords*, a lematização e o *stemming* [18].

2.3.1. Remoção de *Stopwords*

As *stopwords* são termos freqüentes em textos e não dotados de informação de maior relevância, ou seja, não representativa no documento. Sua remoção tem como finalidade a compressão de textos, pois resulta na redução de termos analisados no documento e na

diminuição do número de palavras armazenadas na base de dados, além de ser considerada como um meio de diminuição da dimensionalidade da tabela atributo-valor [19].

As ferramentas *PreText*, versão 1 e versão 2, realizam a remoção das *stopwords* comparando a palavra dos textos a ser processada e as palavras existentes no arquivo *stoplist*. Este arquivo é composto por um conjunto de conectivos tais como palavras adverbiais, possessivas, quantitativas, adjuntos adnominais, adjuntos adverbiais, numerais e outras. Caso a palavra comparada estiver presente na lista de *stopwords* a mesma é desconsiderada em processos futuros. Além das *stopwords*, verbos irregulares dar, dizer, estar, fazer, haver, ir, poder, saber, ser, ter, ver e vir, também são ignorados no processo de seleção de *stems*. Na tabela 1 estão representados alguns *stopwords* presentes no arquivo *stoplist*.

Tabela 1: Conjunto de stopwords da ferramenta PreText

a	basicamente	e	mesmas mesmo	pelo	voce
abaixo	bastante	eis	mesmos	pelos	voces
acaso	bastantes	ela	meu	per	vos
acerca	bem	elas	meus	perante	vossa
acima	bom	ele	mim	pero	vossos
acola	ca	eles	minha	pois	vulgo
ademais	cada	em	minhas	por	cadern
adentro	cade	embaixo	mui	porem	um
adiante	caso	embora	muita	porquanto	dois
afinal	certa	enfim	muitas	porque	duas
afora	certamente	enquanto	multissimo	portanto	tres
agora	certas	entanto	muito	porventura	quatro
agorinha	certo	entao	muitos	possivelmente	cinco
ai	certos	entre	mutuamente	posteriormente	seis
ainda	chez	entretanto	na	posto	sete
alem	com	exceto	nada	pouca	oito
algo	comigo	essa	nadinha	poucas	nove
alguem	como	essas	nalgum	pouco	dez
algum	comumente	esse	nalguma	poucos	onze
algumas	conforme	esses	nalgumas	pra	doze
alguns	confronte	esta	nalguns	praquela	treze
ali	conosco	estas	naquela	praquelas	quatorze
alias	conquanto	este	naquelas	praquele	quinze
amiude	consigo	estes	naquele	praqueles	dezesseis
ante	consoante	eu	naqueles	praquilo	dezessete
antes	contanto	exatamente	naquilo	pras	dezoito
ao	contigo	exceto	nao	praticamente	dezenove
aonde	contra	felizmente		prela	vinte

2.3.2. Radicalização – *Stemming*

Em Processamento de Linguagem Natural (PLN), algoritmos de Radicalização, ou também denominados de *Stemming*, consistem em uma normalização lingüística, em que as palavras e suas variantes são simplificadas a uma forma comum, em um “quase radical” ou *stem* [20], resultando na diminuição do número de palavras armazenadas na base de dados. É válido salientar que o radical resultante da normalização de uma palavra não é necessariamente igual a sua raiz lingüística.

Assim como a maioria dos programas de *Stemming*, ambos os *PreText* são baseados no algoritmo de Porter [21]. Seu funcionamento consiste na remoção de sufixos e/ou prefixos das palavras, o que o torna altamente dependente da linguagem de escrita dos textos utilizados. Fato esse, facilmente observado na diferença quantitativa de regras dos algoritmos direcionados à língua inglesa em comparação a algoritmos desenvolvidos para línguas originadas do Latim, como o Português e o Espanhol.

Tanto para o Inglês, quanto para as línguas originadas do Latim, o algoritmo remove os sufixos possuidores de um tamanho mínimo estabelecido de acordo com um conjunto de regras pré-estabelecidas. Contudo, para o Português e o Espanhol, caso não seja possível eliminar algum sufixo após a verificação destas primeiras regras, é analisada a existência de terminações verbais que, em caso positivo, são eliminadas. Essa análise é necessária, pois estas línguas possuem formas verbais amplamente conjugadas em sete tempos verbais e cada uma com seis terminações diferentes.

As regras que determinam quando uma palavra é suficientemente grande para eliminar seu sufixo é estabelecida com o auxílio de duas variáveis, **posV**, utilizada para eliminar terminações verbais, e **pos2**, para eliminações de outras formas de sufixo. Ambas são complementares e usam a medida *m* de Porter para eliminar o sufixo e são posicionadas como um intervalo para preservar uma região crítica da palavra. Esta medida *m* é imprescindível, pois, em sua ausência, *stems* não interessantes podem ser gerados de palavras muito pequenas. A figura a seguir exemplifica a utilização das variáveis no processo de *Stemming* [22].

abandoná-lo	augusto
↑ ↑	↑ ↑
V 2	V 2
baseando-se	proposição
↑ ↑	↑ ↑
V 2	V 2

Figura 2: Variáveis de comprimento mínimo da palavra¹

A variável **pos2** se posiciona no fim de uma seqüência de consoantes que segue a segunda seqüência de vogais, como mostrado pela seta referente ao número “2” na figura 2. Já o posicionamento da variável **posV** depende de como é o início da palavra, ou seja, se a palavra começa com duas vogais ou mais, como em “augusto” (figura 2), **posV** é posicionada na primeira consoante; já se a palavra começa com vogal-consoante, como em “abandoná-lo” (figura 2), **posV** posiciona-se na segunda vogal; e por fim, se a palavra começa com consoante, como em “baseando-se” (figura 2), **posV** é posicionada na terceira letra mais a direita.

Após o posicionamento das variáveis, a redução das palavras pode ser realizada conforme as regras presentes na tabela 2. No entanto, em casos nos quais as regras da tabela 2 não se aplicam, a palavra é submetida à remoção de terminações verbais, como exposto na tabela 3. Vale ressaltar que toda a análise aqui descrita é feita para a língua portuguesa. Outro cuidado que o algoritmo apresenta refere-se à remoção de terminações referenciais como –lo, –la e –se, antes da submissão das palavras às regras de remoção descritas anteriormente.

¹Fonte: Tutorial *PreText* [22]

Tabela 2: Regras de remoção de sufixos da língua portuguesa²

N _o	Condição	Ação	Exemplo
1		[ae]is → [ae]l ns → m res → r s →	casais → casal álbuns → álbum fatores → fator casas → casa
2	depois(pos2)	idade →	adversidade → advers
3	se(2) e depois(pos2)	abil → iv → ic →	amigabil(idade) → amig exclusiv(idade) → exclus infelic(idade) → infel
4	depois(pos2)	ic[ao] →	irônic[ao] → irôn
5	depois(pos2)	ável → ível →	lamentável → lament acessível → acess
6	depois(pos2)	ismo →	alcooolismo → alcool
7	se(1) ou depois(pos2)	çõe → ção →	acomodaçõe(s) → acomoda resolução → resolu
8	se(1) ou depois(pos2)	ador → adora →	colonizador(es) → coloniz colonizador(as) → coloniz
9	depois(pos2)	os[ao] →	corajos[ao] → coraj
10	depois(pos2)	ista →	dermatologista → dermatolog
11	depois(pos2)	amento → imento → amente →	pensamento → pens discernimento → discern discretamente → discret
12	se(11) ou depois(pos2)	os → ativ → ad → iv → ic →	generos(amente) → gener comparativ(amente) → compar deliberad(amente) → deliber compulsiv(amente) → compuls demagogic(amente) → demagog
13	depois(pos2)	mente →	difícilmamente → difícil
14	se(13) ou depois(pos2)	avel → ível →	favoravel(mente) → favor possivel(mente) → poss
15	depois(pos2)	iv[ao] →	primitiv[ao] → primit
16	se(15) ou depois(pos2)	at →	recreat(ivo) → recre
17	depois(pos2)	eza →	sutileza → sutil

²Fonte: Tutorial *PreTeX*T [22]

Tabela 3: Terminações verbais do Português³

1ª Conjugação						
Infinitivo	-ar					
	1ª PS	2ª PS	3ª PS	1ª PP	2ª PP	3ª PP
Presente	-o	-as	-a	-amos	-ais	-am
Subjuntivo	-e	-es	-e	-emos	-eis	-em
Futuro	-arei	-arás	-ará	-aremos	-areis	-arão
Condicional	-aria	-arias	-aria	-aríamos	-aríeis	-ariam
Imperfeito	-ava	-avas	-ava	-ávamos	-áveis	-avam
Passado	-ei	-aste	-ou	-ávamos	-astes	-aram
Imp. Subjuntivo	-asse	-asses	-asse	-ássemos	-ásseis	-assem
Mais Perfeito	-ara	-aras	-ara	-áramos	-áreis	-aram
Presente Partic.	-ando					
Passado Partic.	-ada -ado -adas -ados					
2ª Conjugação						
Infinitivo	-er					
	1ª PS	2ª PS	3ª PS	1ª PP	2ª PP	3ª PP
Presente	-o	-es	-e	-emos	-eis	-em
Subjuntivo	-a	-as	-a	-amos	-ais	-am
Futuro	-erei	-erás	-erá	-eremos	-ereis	-erão
Condicional	-eria	-erias	-eria	-eríamos	-eríeis	-eriam
Imperfeito	-ia	-ias	-ia	-íamos	-íeis	-iam
Passado	-i	-este	-eu	-emos	-estes	-eram
Imp. Subjuntivo	-esse	-esses	-esse	-éssemos	-ésseis	-essem
Mais Perfeito	-era	-eras	-era	-éramos	-éreis	-eram
Presente Partic.	-endo					
Passado Partic.	-ida -ido -idas -idos					
3ª Conjugação						
Infinitivo	-ir					
	1ª PS	2ª PS	3ª PS	1ª PP	2ª PP	3ª PP
Presente	-o	-es	-e	-imos	-is	-em
Subjuntivo	-a	-as	-a	-amos	-ais	-am
Futuro	-irei	-irás	-irá	-iremos	-ireis	-irão
Condicional	-iria	-iriam	-iria	-iríamos	-iríeis	-iriam
Imperfeito	-ia	-ias	-ia	-íamos	-íeis	-iam
Passado	-i	-iste	-iu	-imos	-istes	-eram
Imp. Subjuntivo	-isse	-isses	-isse	-íssemos	-ísseis	-issem
Mais Perfeito	-ira	-iras	-ira	-íramos	-íreis	-iram
Presente Partic.	-indo					
Passado Partic.	-ida -ido -idas -idos					

³Fonte: Tutorial *PreTeX* [22]

2.4. Tabela Atributo-Valor (*Bag-of-words*)

A tabela atributo-valor, também denominada por *bag-of-words*, consiste em uma abordagem de representação de documentos a partir de um vetor de termos presentes nesses documentos. Os textos são representados com exemplos (linhas da tabela) e os *stem*, ou termos, identificados são os atributos (colunas da tabela), tal qual na tabela 4. A forma com que os atributos são mensurados é a partir de valores tanto booleanos, os quais relatam se um termo foi encontrado ou não em um texto, quanto esparsos, os quais contabilizam a frequência de aparecimento de um termo, como exemplificado na tabela 4, em que os números nas células centrais da tabela referem-se à frequência dos termos (colunas) em cada texto (linha) [22].

Como cada palavra pode ser um possível elemento do conjunto de atributos da tabela atributo-valor, esta estrutura geralmente possui alta dimensionalidade [23].

Tabela 4: Tabela atributo-valor⁴

documento	amig	ciudad	trabalh	filh	cas
texto1.txt	0	0	0	0	1
texto2.txt	0	0	0	1	1
texto3.txt	0	0	1	1	1
texto4.txt	0	1	1	1	1
texto5.txt	5	5	5	5	5
texto6.txt	5	8	13	28	28
texto7.txt	5	6	10	8	8

2.5. *PreText*

A ferramenta *PreText* consiste de uma ferramenta de mineração de texto, baseada na abordagem *bag-of-words*, utilizando técnicas de remoção de *stopwords* e *stemming*, desenvolvida no Instituto de Ciências Matemáticas e de Computação (USP-SC).

2.5.1. *PreText*: uma Ferramenta para Pré-processamento de Textos Utilizando a Abordagem *Bag-of-words* (versão 1)

A ferramenta *PreText* (versão 1) foi computacionalmente implementada em linguagem *Perl* para pré-processamento de textos e utilizando a abordagem *bag-of-words* na sintaxe padrão do *DISCOVER*. Esse formato estruturado é utilizado pela maioria dos algoritmos de

⁴ Fonte: Tutorial do *PreText* [22]

Aprendizado de Máquina (AM). Uma das principais funcionalidades da ferramenta é transformar palavras dos textos processados em *stem* (“quase radicais”) [22].

A interação com este *PreText* é iniciada com a definição dos parâmetros de processamento dos textos durante a elaboração do arquivo *parameters.cfg* em que se deve detalhar: em que linguagem os textos foram escritos; especificar os diretórios que estarão armazenando os arquivos intermediários; assim como especificar os arquivos de listas de *stopwords*, de gráficos do tipo *.data* e *.names* e; os arquivos a serem processados. Também pode ser definido o nome do arquivo com o *log* de execução e a quantidade de palavras que gerarão um *stem* (termo), nesse programa é possível escolher em um, dois ou três *gramas*, ou combinações de divisões, além de escolher seus respectivos tipos de estatísticas de processamento, *measure* (*boolean*, *freq*, *tfidf* e *tflinear*), *normalize* (*linear*, *quadratic* e *disable*), *smooth* (habilitados ou não), *min*, *max* e *std_dev*.

A variável *measure* detém-se em mensurar a presença dos termos nos textos, por valores booleanos, de frequência (*freq*), de *tfidf* ou de *tflinear*. A medida booleana representa a presença de um termo de forma binária, ou seja, caso este esteja presente, seu campo na tabela atributo-valor assume o valor 1 e, em caso contrário, admite o valor 0. Já a medida *freq* ou *tf* contabiliza o número de ocorrência dos termos em um dado documento.

Contudo, o fato de um termo ocorrer em quase todos os documentos, torna-o pouco relevante e assim pouco utilizável em tarefas de mineração de texto. Nesses casos, a medida *tfidf* pode trazer bons resultados, uma vez que seu cálculo de presença de termos possui um fator de ponderação multiplicado a frequência da palavra em um documento. O fator de ponderação age de forma a “forçar” que os resultados dos termos muito frequentes em todos os documentos tenham uma importância menor, e pode ser obtido a partir do *log* da razão entre o número de documentos do conjunto pelo número de documentos em que o termo aparece. E por fim, a medida *tflinear* foi desenvolvida pelos autores do *PreText* e propõe que o fator de ponderação seja linear variando entre 0 e 1.

Em casos nos quais os documentos possuem assuntos muito semelhantes ou mesmo de um domínio do conhecimento muito restrito, a aplicação destes fatores de ponderação não age de forma positiva, pois seus valores serão tão baixo que zerarão os valores de ocorrência desses termos. Assim, outra solução proposta pelos autores do *PreText* foi não permitir que esses fatores de ponderação assumam o valor zero, a partir da ativação de uma variável denominada *smooth* que, temporariamente, aumenta em 10% o número de documentos da coleção.

A variável *normalize* é utilizada para tornar o algoritmo insensível à diferença do tamanho dos arquivos processados, normalizando de forma quadrática ou linear os valores da tabela atributo-valor.

E por fim, as variáveis *max*, *min* e *std_dev* são utilizadas para a redução de dimensionalidade da tabela atributo-valor pela definição manual dos pontos de corte de Luhn (*max* e *min*), uma vez que, conforme Luhn [24], os termos mais significativos estão posicionados no meio destes dois pontos, logo os com frequência fora deste intervalo são ignorados na construção da tabela atributo-valor. E caso a *std_dev* seja definida no arquivo *parameters*, a tabela atributo-valor será reduzida de acordo com este desvio padrão sobre o *rank* de *stems*.

A obtenção de *stems* a partir das palavras é feita pelo módulo *stem.pl*. Em sua execução são utilizados os diretórios de listas de *stopwords*, diretórios da base de textos e o arquivo de parâmetros. Finalizada as tarefas do módulo *stem.pl*, são gerados oito arquivos intermediários.

A partir desses arquivos o módulo *report.pl* do *PreText* cria a tabela atributo-valor em que os textos são representados como exemplos e os *stems* encontrados na fase anterior são os atributos, tal qual na tabela 4. A forma com que os atributos são contados é feita a partir do parâmetro *measure* escolhido. Outra saída do módulo *report.pl* é um arquivo com dados para a geração de gráfico (caso o usuário deseje, este módulo permite a aplicação dos cortes de Luhn).

A linha de execução explicada anteriormente pode ser visualizada na figura 3. Esta demonstra as entradas e saídas dos módulos *Stem.pl* e *Report.pl*.

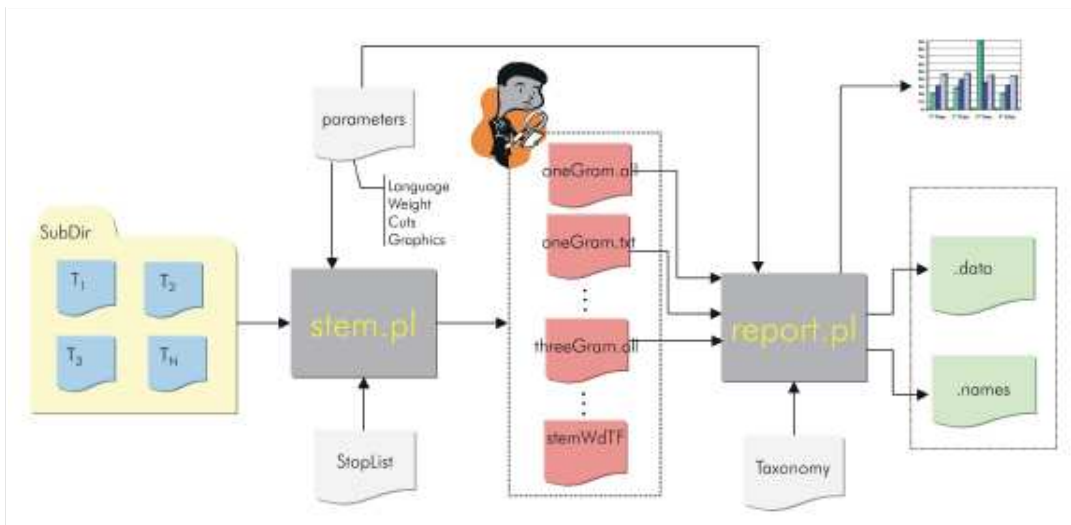


Figura 3: Fluxograma de execução da ferramenta *PreText* versão 1⁵

Contudo percebeu-se que esta versão do programa produz alguns resultados inapropriados na aquisição de *tokens* (termos compostos por um ou mais *stems*) gerados por mais de um grama. Por exemplo, no texto “Estruturas ósseas regionais com morfologia e intensidade de sinal preservado. Cartilagem articular com espessura e sinal conservados.”, o algoritmo gera o *stem* preserv_cartilag_articul, mesmo que suas palavras de origem estejam separadas por pontuação como o ponto final. Por essa liberdade na obtenção dos *tokens*, um número significativo de *tokens* é desprovido de semântica e, conseqüentemente, inapropriados para a estruturação dos laudos radiológicos.

2.5.2. *PreText*: A Reestruturação da Ferramenta de Pré-processamento de Textos (versão 2)

Esta ferramenta consiste na reestruturação e reimplementação da ferramenta *PreText* (versão 1) de forma a corrigir a geração de *tokens*, disponibilizar um número maior de funções, além de realizar o pré-processamento de forma ágil.

Nesta versão da ferramenta foi mantida a abordagem *bag-of-words* para o pré-processamento de textos, assim como as técnicas de tokenização, *stemming* e *stopwords*. Também foram mantidas as técnicas de redução da tabela atributo-valor a partir do valor do atributo (frequência) dos termos em relação ao conjunto de textos processados, assim como o

⁵ Fonte: Tutorial do *PreText* [22]

conceito de n-grama (*n-gram*) em que a associação de uma ou mais palavras posicionadas consecutivamente ou não, resultam em um termo (*token*) [25].

Com a remodelagem e reimplementação do *PreText*, a nova versão da ferramenta possui as seguintes características ilustradas na figura 4.

O módulo *Start.pl* lê o arquivo de configuração gerado pelo usuário e gerencia os demais módulos. Em seguida, o módulo *Maid.pm* realiza a limpeza dos textos com a remoção de *stopwords*, símbolos, *tags HTML* e faz o tratamento de símbolos e gera os *stems* a partir das palavras resultantes pela classe *Stemmer.pm*. Em seguida, o módulo *NGram.pm* gera os n-gramas, documentando-os nos arquivos *NGram.txt* e *NGram.all*. E por fim, o módulo *Report.pm*, a partir dos resultados anteriores, faz o processamento da taxonomia e calcula as medidas e normalizações solicitadas, resultando em uma tabela atributo-valor no formato *DSX* do *Discover* e em alguns gráficos.

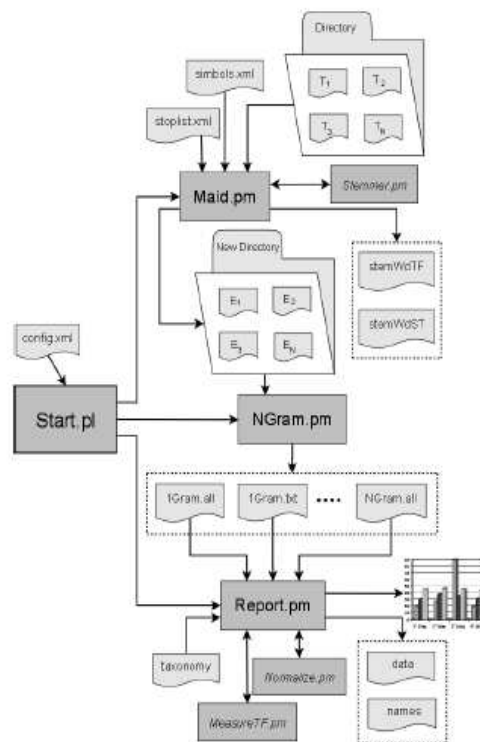


Figura 4: Fluxograma de execução da ferramenta *PreText* versão 2⁶

2.6. Ontologia

O termo ontologia teve sua primeira definição nas ciências filosóficas e consistia na sistemática explicação da existência. Em meados da década de 90, houve uma nova utilização

⁶ Fonte: Tutorial do *PreText* [25]

desse termo na área da inteligência artificial e, segundo Gruber [26], a especificação explícita de uma conceitualização é a definição de ontologia para o campo da inteligência artificial. A partir desta definição, muitas outras foram propostas na literatura. Para Guarino e Giarretta, ontologia é uma descrição de um domínio a partir de uma hierarquia estruturada de um conjunto de termos, podendo ser usada como estrutura de uma base de dados e fornecendo meios para expor explicitamente a conceitualização do conhecimento neste domínio [26].

Uma ontologia é, portanto, um meio de demonstração explícita e estrutural de termos e seus relacionamentos, livre de ambigüidade e pertencentes a um mesmo domínio, com a finalidade de permitir a elaboração e compartilhamento de vocabulário comum entre seus distintos usuários.

Na figura 5, é demonstrada uma clássica comparação entre uma categorização em níveis de vinhos (primeiro esquema da figura 5) e a simples listagem dos vinhos e tipos de vinhos (segundo esquema da figura 5). Ambas possuem o objetivo de representar a realidade, contudo o primeiro esquema apresenta uma quantidade maior de informação, assim como induz o aprendizado deste domínio.

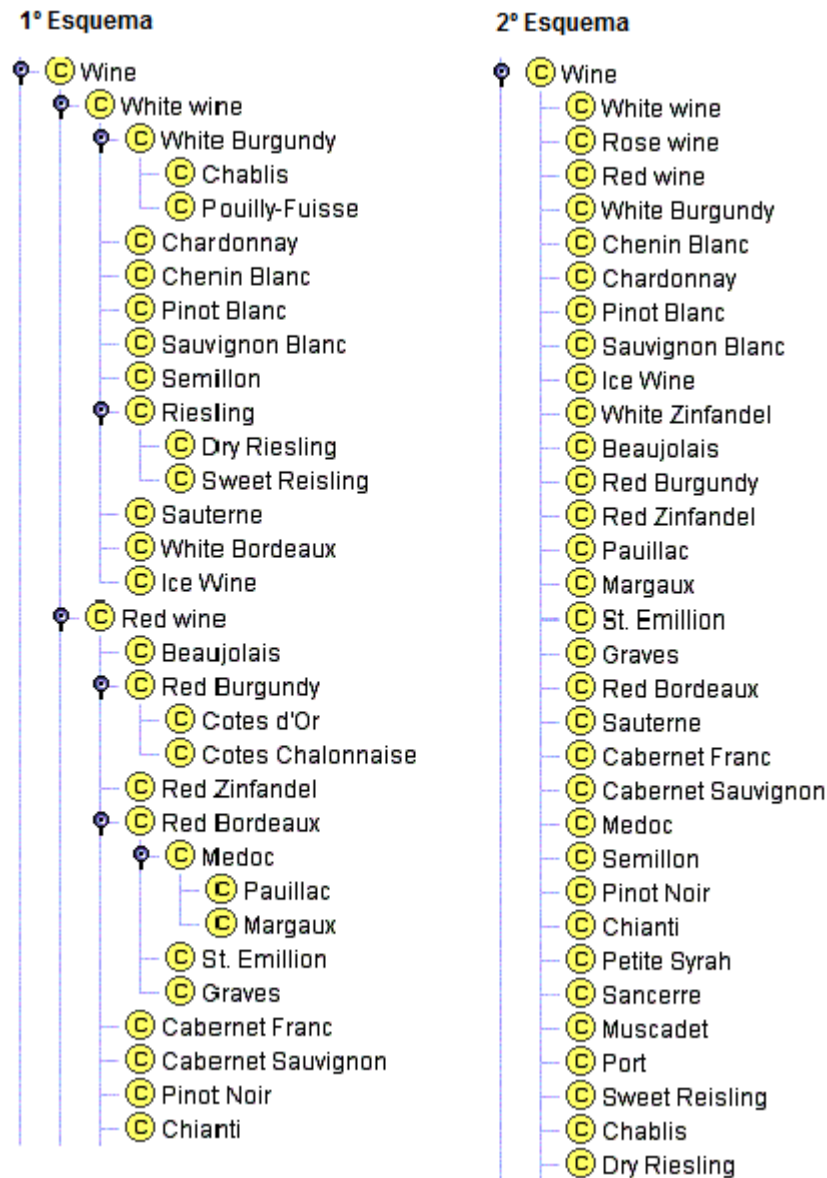


Figura 5: Comparação entre categorização em níveis *versus* listagem de todos os vinhos e tipos⁷.

Um modelo ontológico é constituído de cinco componentes básicos: as classes, que representam todo e qualquer elemento do domínio de conhecimento da ontologia; os relacionamentos, que expõem a interação entre essas classes; as funções, que determinam relações de causa e consequência entre os elementos da relação; os axiomas, determinando a verdade absoluta deste domínio e; as instâncias representando os elementos da ontologia.

O conhecimento pode ser representado por quatro diferentes modos: pela Taxonomia, a qual determina o como se organiza as classes e subclasses de uma ontologia pelas relações de generalização/especialização em hierarquia simples ou múltipla; pela Partonomia e Mereologia, que possuem o objetivo de especificar relações semânticas relativas aos objetos

⁷ Site para download <http://protege.stanford.edu/>, acesso no dia 21/06/2008

tal como sinônimos e antônimos; pela Cronológica, determinando uma relação temporal entre as classes relacionadas; e por Topologia, definindo as conexões existentes entre as classes, podendo ser irreflexibilidade, simetria e não transitividade [27].

Uma ontologia tem o intuito de compartilhar de um conhecimento comum entre as pessoas e os agentes de *software* [26], provendo consistência de representação, livre de ambigüidade. Por exemplo, imagine uma grande quantidade de diferentes *web sites* que contêm informações médicas ou de comércio eletrônico. Se esses *web sites* dividem e apresentam a mesma ontologia, os agentes de *software* podem utilizar-se destas para responder às pesquisas de usuários ou como alimentação para outras aplicações.

Durante o processo de desenvolvimento de uma ontologia, todos os termos devem ser acompanhados de documentação em linguagem natural e definições que expressam as condições necessárias e suficientes para representá-los, devendo manter a coerência dessas definições para que suas inferências sejam consistentes, além de permitir que novos termos sejam incluídos sem alterações das definições existentes. As classes da ontologia não precisam ser dependentes e sem superposição de conceitos. Também devem, sempre que possível, utilizar nomes padronizados e explorar ao máximo os mecanismos de herança múltipla, assim como o de modularidade [27].

CAPÍTULO 3: Metodologia

3.1. Considerações Iniciais

Este trabalho foi desenvolvido a partir das atividades de estudo do domínio de conhecimento de laudos radiológicos assim como da implementação de conceitos e técnicas de mineração de texto para a estruturação desses laudos.

Para efetivar a proposta desta pesquisa adotaram-se, com o auxílio de um radiologista experiente, textos de laudos referentes a exames de Ressonância Nuclear Magnética do Joelho como estudo de caso.

Os dados utilizados foram obtidos da base de laudos radiológicos do HCFMRP, atualizada até a data de 06/08/2007. Os dados consistem em exames de Ressonância Magnética do Joelho inseridos no sistema desde 1999 até agosto de 2007, totalizando 2051 exames.

Vale ressaltar que este trabalho foi realizado preservando os conceitos de bioética, mantendo em sigilo toda e qualquer informação de pacientes do HCFMRP. Concomitantemente, valores de identificação dos pacientes foram ficticiamente inseridos na base de dados desenvolvida neste trabalho.

Para a execução deste trabalho foram realizadas algumas etapas que caracterizaram metodologia do trabalho.

3.2. Domínio de Conhecimento dos Laudos

Com o intuito de garantir qualidade à proposta do presente projeto, torna-se indispensável uma boa compreensão dos elementos descritos durante a geração dos laudos utilizados. Visto que o mesmo está contido em um domínio de conhecimento médico, foi necessária à graduanda a revisão dos conceitos sobre anatomia humana, anatomia clínica e aquisição de exames radiológicos.

Aditivamente, a graduanda realizou atividade de visita ao centro radiológico do hospital HCFMRP e clínica Documenta do Hospital São Francisco a fim de conhecer o cotidiano e o processo de geração dos laudos.

3.3. Estudo das Características de Preenchimento dos Laudos

Uma parcela dos textos utilizados neste trabalho foi estudada pela graduanda a partir de análise manual, tendo por finalidade a observação das características de preenchimento como os tipos conceituais existentes, a frequência e disposição destes, e outras características textuais presentes nesses tipos de dados, assim como a tendência descritiva nestes textos como, por exemplo, a importância em relatar a normalidade de determinada estrutura ou priorizar apenas as anormalidades.

Esta atividade foi realizada com o intuito de prever e avaliar os resultados da execução dos conceitos de mineração de textos, além de observar como as informações são mantidas conforme as características de preenchimento.

3.4. Pré-processamento dos Laudos

Para estudar os resultados gerados pelas técnicas de *stemming* e remoção de *stopwords*, a representação do texto com a abordagem *bag-of-words* e suas estatísticas, foram pré-processados um conjunto inicial de 1659 laudos utilizando a ferramenta *PreText* versão 1⁸. Durante essa etapa do trabalho, foram analisados os gráficos gerados com diferentes estatísticas, além de seus arquivos textuais, e principalmente os arquivos resultantes do processo de *NGram*, contendo os *n_gramas* identificados no texto.

Esse estudo foi necessário para caracterizar o perfil textual dos laudos, notando necessário o tratamento de símbolos, para a geração de *tokens* adequados. Em seguida, o mesmo estudo foi realizado para a versão 2 dessa ferramenta⁹, que apresentou um desempenho notável e de fato tratou a geração inadequada de *tokens*.

Percebeu-se, nessa fase, a necessidade da remoção de um conjunto de palavras da lista de *stopwords* por serem relevantes no contexto desses textos ou necessárias na estratégia de representação semântica das sentenças.

Durante essa atividade, foi possível concluir que a forma com que ambas as versões do *PreText* expõem seus resultados, abordagem *bag-of-words*, não agrega informação e semântica, por relatarem apenas a presença ou frequência de termos em determinado documento. Logo, listar os termos presentes não permite afirmar que em uma região um achado (anormalidade) foi identificado pelo radiologista. Além disso, esse tipo de abordagem

⁸ Esta tarefa foi realizada no sistema operacional *Windows*.

⁹ Esta tarefa foi realizada no sistema operacional *Linux*.

não trata eventos de negação que é muito empregado em textos radiológicos como, por exemplo, “Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações”.

Com o intuito de acrescentar alguma semântica na estruturação dos textos, decidiu-se por estudar e aplicar conceitos de ontologia neste trabalho.

3.5. Modelagem de uma Ontologia

A partir dos estudos anteriores e, principalmente, com os resultados da atividade de análise das características de preenchimento dos laudos, o domínio da ontologia proposta foi definido como a representação dos elementos descritivos do exame de Ressonância Nuclear Magnética em Joelho, composto por seus respectivos achados, diagnósticos, regiões anatômicas e outras definições clínicas e anatômicas.

Para o desenvolvimento da ontologia, foi pressuposto de que a ferramenta desenvolvida e utilizada para a estruturação dos laudos realizaria um pré-processamento desses textos a partir de uma ferramenta de mineração de texto. E cada *stem* gerado seria categorizado dentre as entidades (classes) da ontologia.

A ontologia foi desenvolvida utilizando a ferramenta *Protégé 3.3.1*¹⁰ com a linguagem *OWL*. Essa linguagem pode ser utilizada por aplicações que precisam disponibilizar e processar o conteúdo da informação, e usa a lógica descritiva para representação do conhecimento, podendo ser lida em conteúdo *Web* suportado por *XML*, *RDF* e *RDF-Schema*. [2].

3.6. Modelagem e Alimentação da Base de Dados

A partir da criação das entidades e relacionamentos existentes na ontologia e considerando a saída gerada pela ferramenta *PreText*, foi possível elaborar a modelagem da base de dados necessária para o armazenamento estruturado das descrições de laudos e diagnósticos radiológicos.

Após avaliação e validação desta modelagem, a base de dados foi construída utilizando a plataforma de bando de dados *Postgresql 8.3*, com a ferramenta *pgAdmin III*, iniciando-se, finalmente, o processo de alimentação da mesma.

¹⁰ Site para download <http://protege.stanford.edu/>, acesso no dia 21/06/2008

As tabelas da base de dados proposta: *Laudo*, *Descrição_Textual*, *Um_Grama* e *Item_Term* foram alimentadas automaticamente, as duas primeiras a partir dos dados fornecidos pelo HCFMRP e a terceira foi alimentada com o resultado do pré-processamento do léxico e de um conjunto de laudos pela ferramenta *PreText* versão 2. A quarta tabela foi alimentada a partir do cruzamento de dados entre as tabelas alimentadas, *Termo* e *Um_Grama*, e o léxico.

Como, para a alimentação das tabelas *Termo* e *Conceito*, é necessário um conhecimento prévio de cada elemento a ser inserido na base, seus dados foram armazenados de forma manual.

3.7. Estruturação dos Laudos

Observou-se a necessidade de desenvolvimento de uma ferramenta para a estruturação de laudos. Essa ferramenta necessitaria identificar cada palavra ou conjunto de palavras existentes no texto em forma de termos e, para que a semântica fosse mantida, notou-se necessário o tratamento de elementos textuais de conjunção como, por exemplo, “e”, “mas” e vírgulas.

Toda essa ferramenta, apelidada por *E-Rad*, foi desenvolvida durante este trabalho, utilizando paradigma orientado a objetos, na linguagem **Perl**, e reutilizando alguns objetos do *PreText* versão 2 para o processo de *Stemming* e remoção de *Stopwords*, e acessando a base de dados *RadOn*, desenvolvida na base de dados *Postgresql 8.3*, para a categorização dos termos entre as classes da ontologia. Também foram desenvolvidos módulos para a adaptação e ajuste dos resultados gerados pelo *PreText* e para o armazenamento na base da estrutura gerada.

E para permitir que pesquisas por laudos fossem realizadas, um sistema de busca foi desenvolvido. Sua interface foi elaborada em **HTML** e **PHP**, seus sistemas de identificação de termos é baseado na mesma técnica de estruturação da ferramenta *E-Rad*, também programada em **PERL**, utilizando paradigma orientado a objetos e com interface para a base *RadOn*. Esse sistema de busca foi desenvolvido para “comparar” a sentença de busca e sua semântica com as sentenças existentes na base e assim retornar os laudos possuidores da mesma semântica.

Capítulo 4. Ferramenta E-Rad

4.1. Considerações Iniciais

A tentativa de estruturar laudos radiológicos utilizando abordagem *bag-of-words* não é apropriada por resultar na perda de relacionamento entre os termos existentes, isto é, não demonstra a qual termo um segundo se refere, além de não tratar a presença e o relacionamento entre elementos negativos e os outros componentes. Assim, com o objetivo de reduzir esses resultados inapropriados para a estruturação dos laudos, a ferramenta computacional *E-Rad* (**E**struturação **R**adiológica) foi desenvolvida utilizando paradigma orientado a objetos, na linguagem **Perl** e reutilizando alguns objetos da ferramenta *PreText* versão 2 com a chamada ao objeto *Start.pm* e com interface a base de dados *RadOn*.

A atual composição do *E-Rad* está ilustrada na figura 6, esta é formada por seis módulos principais: *E-Rad.pl*, *TakeReport.pm*, *RemoveAccent.pm*, *Start.pm*, *TakeTermSentence.pm* e *MakeSentences.pm*. A seguir uma breve descrição de cada módulo.

- O módulo *E-Rad.pl* é responsável pelo gerenciamento das atividades, ou seja, é este que requisita aos outros módulos a realização de determinada tarefa. Sua primeira requisição é feita ao módulo *TakeReport.pm*, solicitando as descrições textuais (descrição e diagnóstico dos laudos) que ainda não passaram pelo processo de estruturação;
- O módulo *TakeReport.pm* acessa a base de dados *RadOn* para obter os textos que ainda não foram processados e retorna ao *E-rad.pl* estes textos;
- Ao receber esses documentos, o *E-Rad.pl* repassa-os para o módulo *RemoveAccent.pm*, que os adapta para o pré-processamento do *PreText*, removendo acentos, letras maiúsculas e trocando a letra “e” por “and”;
- Em seguida, esses textos de saída são enviados ao módulo *Start.pm*, que consiste em uma reformulação do *Start.pl* da ferramenta *PreText* versão 2. Esse módulo é responsável pela leitura do arquivo de configuração¹¹ do processamento a ser realizado, pela remoção de *stopwords*, de símbolos *tags html*, e pela geração de *stemming* por uma das classes que herdam a classe *Stemmer.pm*, resultando no envio de um conjunto de gramas posicionados como as suas palavras de origem, com as

¹¹ O arquivo referido é denominado *config.xml*

vírgulas mantidas em suas devidas posições e as sentenças separadas por *breaks* (separadores);

- Ao receber o resultado do processamento do *Start.pm*, o *E-Rad.pl* envia cada sentença gerada para o módulo *TakeTermSentence.pm*, responsável por identificar os termos que compõem a sentença a partir do acesso e comparação dos dados da base de dados *RadOn*;
- Finalmente, o módulo *MakeSentences.pm* é solicitado a cada resultado diferente de nulo gerado pelo módulo *TakeTermSentence.pm*. A tarefa do *MakeSentences.pm* consiste em verificar se há a necessidade de gerar novas sentenças ao comparar o que lhe foi enviado a um conjunto de condições, a fim de manter sua semântica. Em caso positivo, a sentença é submetida a um conjunto de regras e novas sentenças são inseridas na base de dados. Caso contrário, a sentença de entrada é inserida na base.

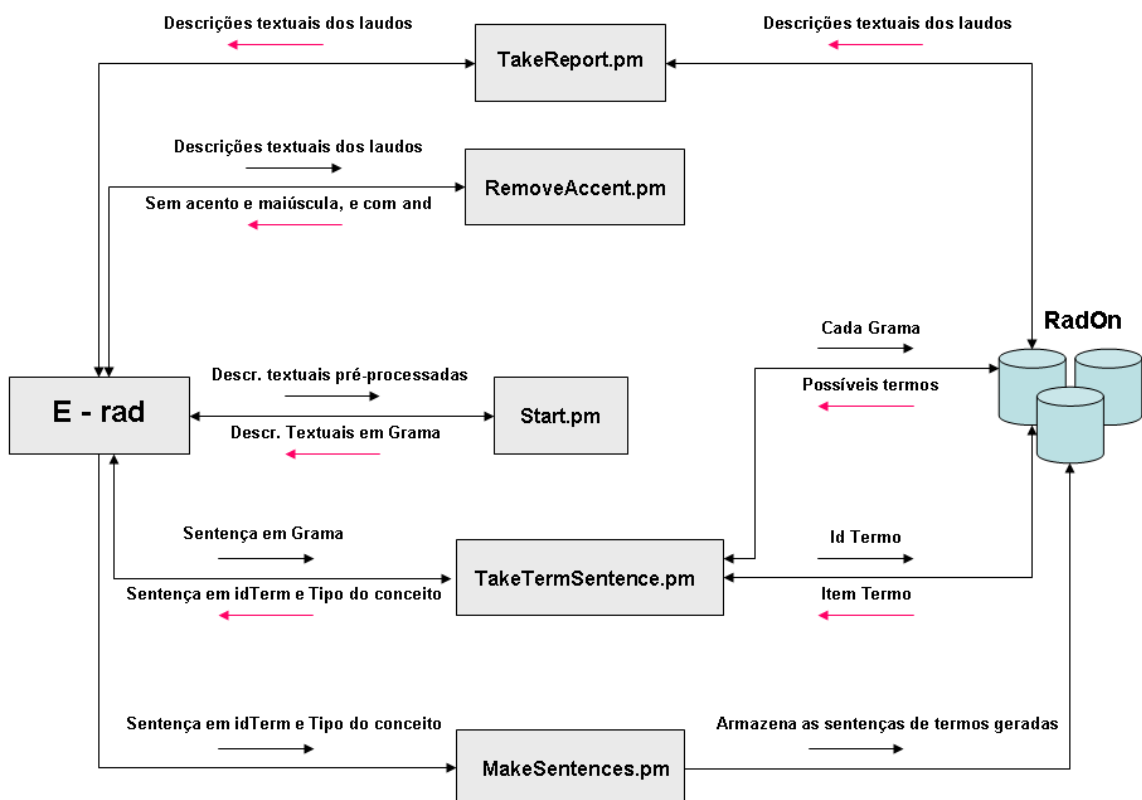


Figura 6: Fluxograma de execução da ferramenta *E-Rad*

As sub-seções seguintes descrevem detalhadamente os módulos da ferramenta, bem como seus resultados.

4.2. Módulo TakeReport.pm

O módulo *TakeReport.pm*, nesta versão, é iniciado manualmente com a execução do módulo *E-Rad.pl*. Contudo, pequenas manipulações nesse módulo habilitam sua execução de tempos em tempos¹², tornando-se um artifício para a verificação e obtenção de laudos novos (não estruturados).

O módulo *TakeReport.pm* acessa a base de dados *RadOn* para obter os textos que ainda não foram processados. Esse acesso é feito a partir de uma busca com junção entre as tabelas *Laudo* e *Descrição_Textual*, e condicionada em relação ao tipo da modalidade, da região e do identificador do laudo, quando este é maior do que o identificador do último laudo estruturado.

Este módulo retorna ao *E-Rad.pl* o identificador da descrição textual (*id_descr_text*), uma variável que especifica o tipo textual (descrição do laudo ou diagnóstico) e o texto.

4.3. Módulo RemoveAccent.pm

O módulo *RemoveAccent.pm* foi implementado nesta ferramenta por conta de resultados inconsistentes gerados pela ferramenta *PreText*, devido à existência de palavras acentuadas no texto. Por exemplo, se a palavra “matemática” estiver presente no texto, o resultado a ser obtido após o processo de *Stemming* corresponderá a “matem” e “tic”, mesmo que haja o tratamento de acentos por essa ferramenta. Outro tratamento da *PreText* que não é efetuado refere-se à substituição de letras maiúsculas por letras minúsculas, essa troca faz-se necessária durante a comparação do grama existente no texto em relação ao existente na base de dados, pois a linguagem **Perl** é sensível à diferença de tamanho de letra (maiúscula e minúscula). Assim, este módulo realiza a substituição das letras quando necessário.

O módulo *RemoveAccent.pm* também trata, de forma especial, a ocorrência da palavra “e”, pois como dito na seção 3.7., na tentativa de manter a semântica da sentença, as conjunções, elementos aditivos, adversativos e consecutivos, devem ser mantidos. Na execução normal do *PreText*, essa palavra é removida do texto por ser considerada uma *stopword*, porém apenas removê-la do *stoplist* não garante que a mesma seja mantida no texto, pois durante o processo de *Stemming* esta será confundida com o verbo “é”¹³ após a

¹² Esta abordagem não foi implementada nesta ferramenta, pois o presente trabalho está simulando o cotidiano hospitalar

¹³ Conjugação na terceira pessoa do singular do verbo irregular ser

remoção dos acentos e assim também será removida do texto. O tratamento realizado por este módulo é verificar a ocorrência desta palavra e, em caso positivo, substituí-la por “*and*”.

Assim, a saída deste módulo consiste no resultado de todas essas substituições aplicadas ao texto enviado pelo *E-Rad.pl*.

4.4. Módulo *Start.pm*

O texto pré-processado por *RemoveAccent.pm* é enviado pelo *E-Rad.pl* para o módulo *Start.pm*, que consiste na reformulação do objeto *Start.pl* do *PreText* versão 2. Este módulo utiliza apenas os módulos *Maid.pm*, *StopList.pm*, *Simbols.pm*, *Stemmer.pm*, *ProgressBar.pm* e *IO::File*. Os módulos *NGram.pm*, *Message.pm* e *Report.pm* não serão utilizados, pois a ferramenta *PreText*, neste trabalho, tem a finalidade de obter apenas os *stems* (*Gramas*) presentes nos textos a serem estruturados dispostos na mesma ordem de suas palavras de origem.

O arquivo de configuração deste módulo define que:

- A língua dos textos a serem processados é o português;
- O diretório¹⁴ de textos é denominado “joelho”;
- Sua execução será silenciosa;
- O diretório contendo os *stoplist* é denominado “*stoplist*” e o arquivo contendo os *stopwords* é enunciado por “*port.xml*”;
- Será realizada a remoção de *tags html*, de símbolos, *stopwords*;
- A técnica *stemming* será aplicada.

Em Apêndice 1 é apresentada a composição deste arquivo de configuração.

Na figura 7 é apresentada a arquitetura deste módulo e as marcações em formato de “x” em vermelho demonstram os módulos que não serão usados nesta ferramenta por não ser necessários para a estruturação.

Para a estruturação dos textos, os arquivos de *stoplist* e *simbols* sofreram pequenas alterações. O primeiro teve seu conjunto de *stopwords* reduzido, pois algumas palavras são consideradas relevantes no processo de estruturação (estas podem ser visualizadas em anexo A). E o arquivo *simbols* foi alterado para que o símbolo vírgula não fosse removido. A vírgula será muito importante durante a execução do módulo *MakeSentences.pm*, por se tratar de um parâmetro das regras de geração de sentenças.

¹⁴ Mesmo não sendo utilizados diretórios nesse módulo, achou-se conveniente definir este no arquivo de configuração

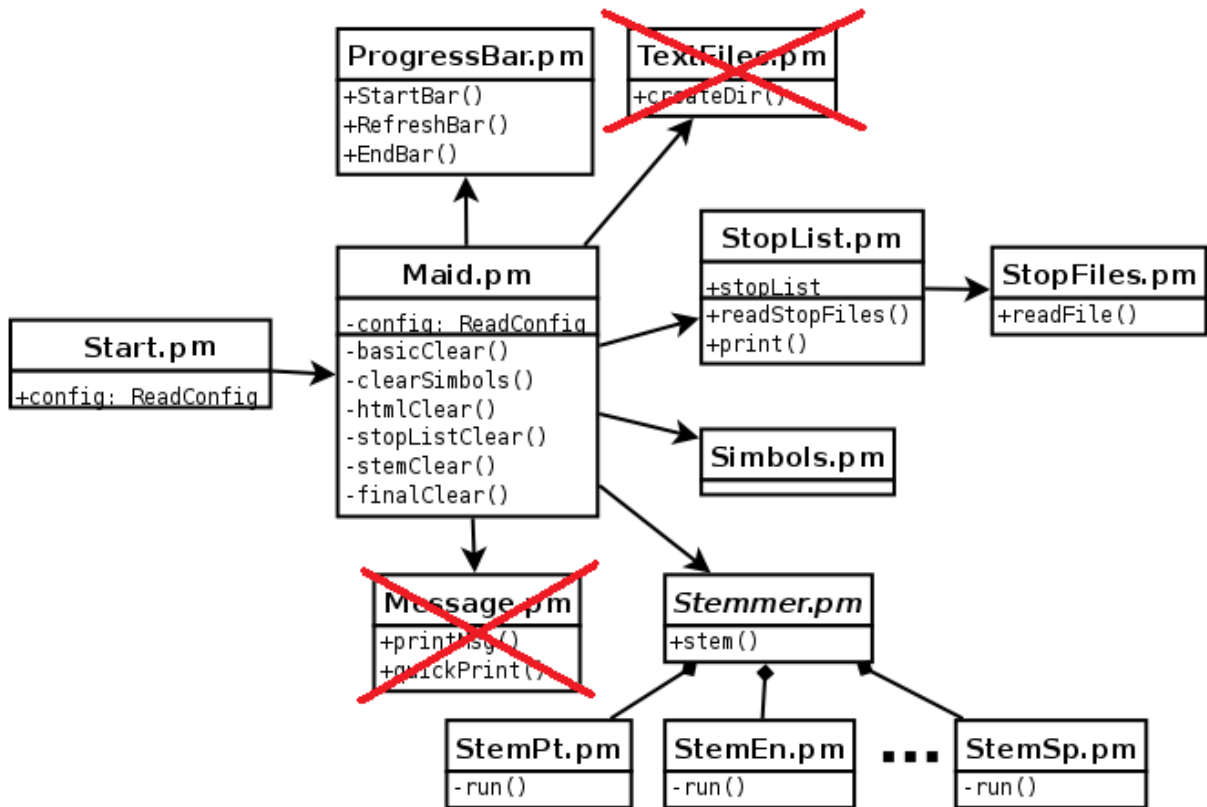


Figura 7: Diagrama de classe do módulo Start.pm e os módulos que este se relaciona

O resultado deste módulo consiste no envio de um conjunto de gramas ordenados tais quais as suas respectivas palavras de origem, mantendo as vírgulas e as sentenças separadas por *breaks* (separadores).

4.5. Módulo TakeTermSentence.pm

Após o pré-processamento do texto, realizado pela ferramenta *PreText* com o módulo *Start.pm*, o gerenciador *E-Rad.pm* divide esse resultado em sentenças conforme a aparição de *breaks* (separadores) e repassa cada sentença para o módulo *TakeTermSentence.pm*.

Este é responsável por gerar uma sentença composta por termos e outra pelos respectivos conceitos a partir de uma sentença composta por gramas. Esse módulo é iniciado com a chamada da sub-rotina *getIdGram()* para obter o identificador de cada grama existente na sentença. Nessa etapa, alguns gramas podem ser descartados por não possuírem referência na base de dados.

Após a obtenção dos identificadores de grama, a sub-rotina *findTerm()* é chamada para cada grama identificado com a finalidade de buscar os termos que iniciem com este grama.

Em seguida, para cada termo encontrado é verificado sua existência na sentença. Para tanto, a sub-rotina *findElem()* faz-se necessária ao retornar a composição deste termo, ou seja, os gramas que compõem este termo e suas respectivas posições (seqüência). Assim, com estes dados, cada identificador de grama que compõe o termo é comparado com os identificadores de gramas presentes na sentença, sendo importante salientar que a ordem ou seqüência de ocorrência dos gramas deve ser mantida conforme o resultado da sub-rotina *findElem()*.

Ao término das comparações, a saída deste módulo é composta por um conjunto de termos ordenados em relação ao seu grama final (o último elemento que compõe este termo) e seu respectivo tipo conceitual.

4.6. Módulo MakeSentences.pm

Para cada sentença diferente de nula, resultante de *TakeTermSentence.pm*, o módulo *MakeSentences.pm* é solicitado. Com o intuito de detalhar o relacionamento entre os termos de uma mesma sentença semântica, este módulo é responsável por verificar se há a necessidade de gerar novas sentenças ao comparar o que lhe foi enviado a um conjunto de condições. Em caso positivo, a sentença de termo enviado como parâmetro é submetida a um conjunto de regras e as novas sentenças são inseridas na base de dados. Caso contrário, a sentença de entrada é inserida diferentemente na base.

O conjunto de condições está presente na sub-rotina *getSentences()*, esta é composta por uma seqüência lógica de condições (“if”) que comparam a igualdade dos termos separados por vírgulas e/ou conjunções (aditivo, adversativo e consecutivo), ressaltando que para cada situação há uma determinada regra de geração de sentença. Em seguida, há uma chamada recursiva com a chamada da sub-rotina *callAnalyse()* enviando essa nova sentença. A recursão terminará quando a sentença verificada não possuir mais vírgula, elementos aditivos, adversativos e consecutivos. Neste momento a sub-rotina *insertSentence()* é chamada e esta sentença é inserida na base. Se a sentença enviada pelo *E-Rad.pl* não possuir vírgula, elementos aditivos, adversativos e consecutivos, esta será inserida diretamente na base de dados chamando a sub-rotina *insertSentence()*.

4.7. Módulo E-Rad.pl

O módulo *E-Rad.pl* tem a responsabilidade de gerenciar o envio e a saída dos módulos da ferramenta, assim como a chamada destes. É composto pela instanciação de cada módulo,

takeReport objeto de *TakeReport.pm*, *removeAccent* objeto de *RemoveAccent*, *start* objeto de *Start.pm*, *takeSentence* objeto de *TakeTermSentence* e *makeSentences* objeto de *MakeSentences*. Para este trabalho sua execução foi realizada manualmente. Na figura 8 é apresentado o diagrama de classe da ferramenta *E-Rad*.

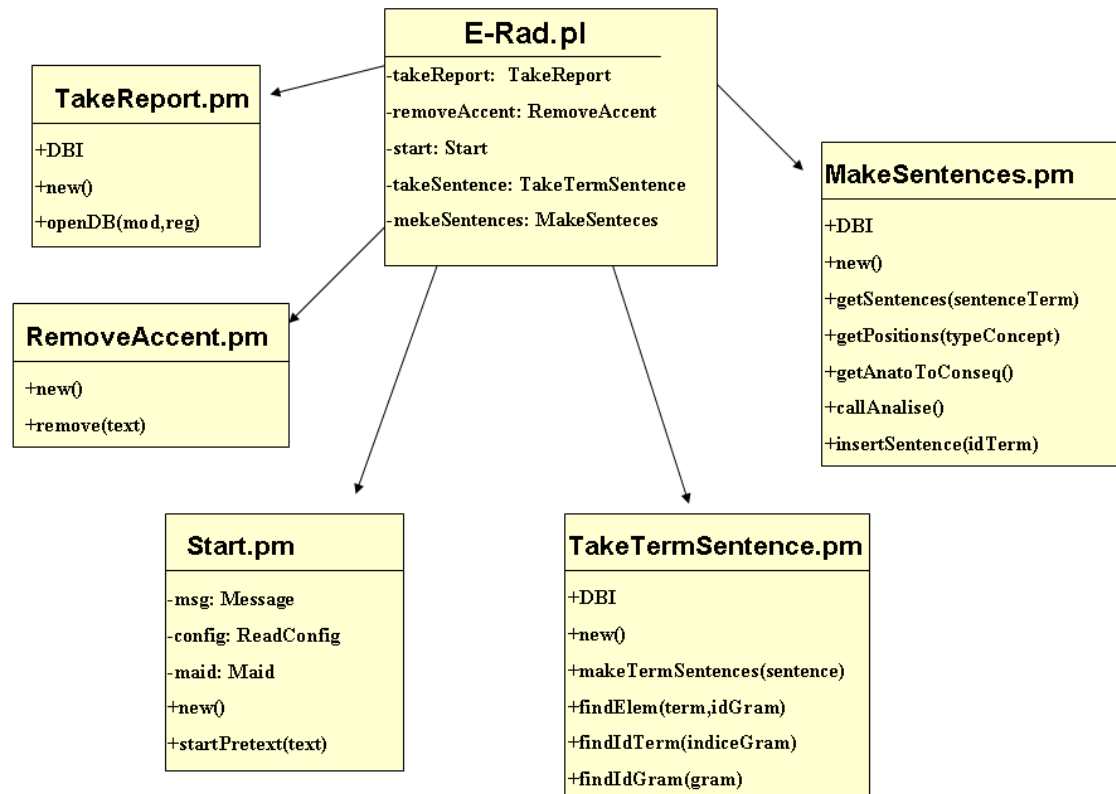


Figura 8: Diagrama de classe da ferramenta E-Rad

Notou-se necessário o desenvolvimento de um sistema de busca para permitir que os usuários pesquisem por laudos utilizando a estruturação proposta neste trabalho. Na sub-seção a seguir esta interface e seu sistema serão apresentados mais detalhadamente.

4.8. Interface e Sistema de Busca

Para que os usuários pudessem utilizar-se da estruturação dos textos para procurar por laudos radiológicos, notou-se a necessidade de desenvolver uma interface, apresentada na figura 9, e

um sistema com esse objetivo. A interface foi desenvolvida utilizando as linguagens **PHP e HTML**, utilizando um servidor *Apache3*.

Esta interface, após obter a sentença de busca do usuário, a envia para o sistema *findReports* que é composto por quase todos os módulos iguais ao *E-Rad*. O primeiro módulo, que é muito parecido com o *E-Rad.pl*, denominado *findReports.pl*, obtém a sentença de busca inserida na interface pelo usuário e, em seguida, envia essa sentença para o módulo *RemoveAccent.pm* cujo resultado é repassado para o módulo *Start.pm*. Assim como no *E-Rad*, cada sentença do resultado de *Start.pm* é enviada para *TakeTermSentence.pm* e sua respectiva sentença de termos é gerada. Por fim, a sentença de termos é analisada no módulo *TakeResultReport.pm* para que, caso seja necessário, este gere novas sentenças. Após este processamento verifica-se na base de dados a existência de algum laudo que possua alguma sentença com um conjunto de termos parecido ao conjunto de termos da sentença de busca ou por conceitos semelhantes.

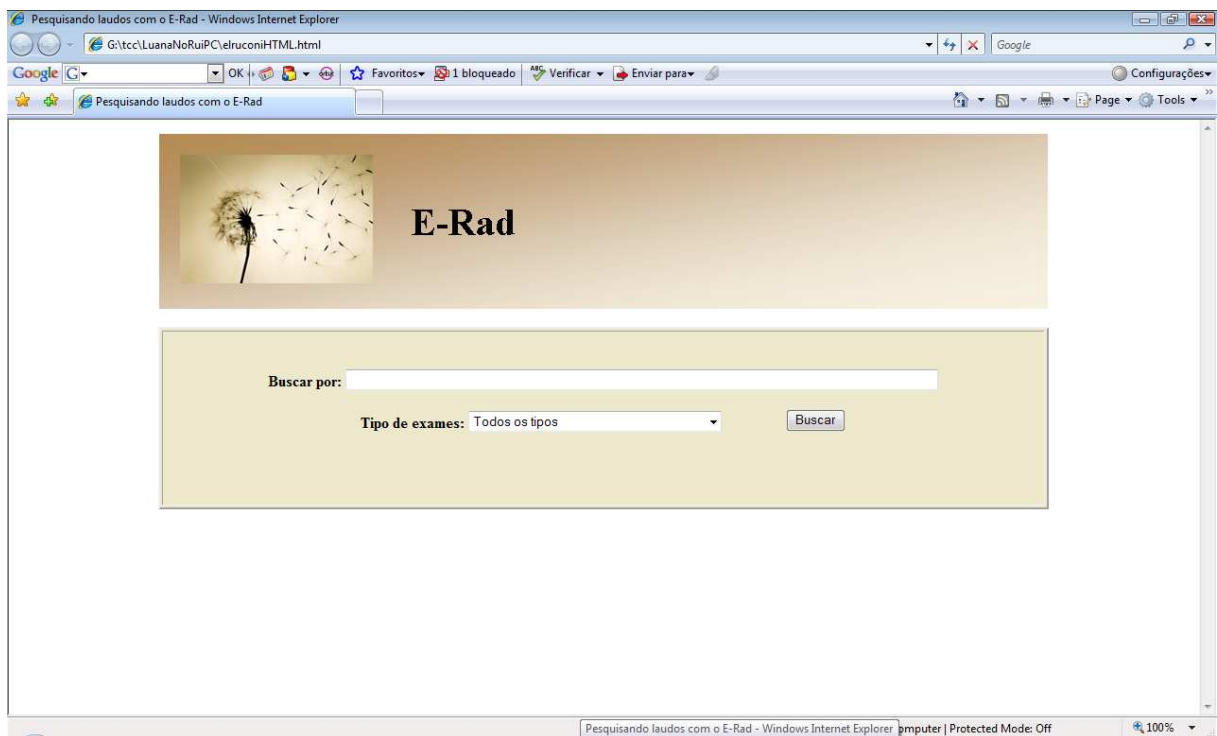


Figura 9: Interface para busca por laudos estruturados

Capítulo 5. Resultados

5.1. Considerações Iniciais

Este trabalho foi desenvolvido a partir das atividades de estudo do domínio de conhecimento de laudos radiológicos, assim como da implementação de conceitos e técnicas de mineração de texto para a estruturação de laudos radiológicos. Para tanto, foram realizadas algumas etapas que caracterizaram a metodologia do trabalho.

Para efetivar a proposta deste trabalho, foram utilizados dados da base de laudos radiológicos fornecida pelo HCFMRP, totalizando 2051 exames. Na tabela 5, são apresentados alguns dados relatados nesses laudos. Observa-se que cada laudo carrega informações relativas a: data do exame, região anatômica examinada, nome do exame, descrição (laudo) e conclusão (diagnóstico).

Como estudo de caso, os laudos de exames de Ressonância Nuclear Magnética tendo o Joelho como a região anatômica examinada foram escolhidos com o auxílio de um radiologista experiente.

Tabela 5: Exemplo de laudo radiológico

Data do exame	Nome da região	Nome do exame	Descrição do exame	Conclusão do laudo do exame
29/11/00	Joelho	Ressonância Magnética	<p>Estruturas ósseas regionais com morfologia e intensidade de sinal preservados.</p> <p>Cartilagem articular com espessura e sinal conservados.</p> <p>Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações.</p> <p>Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais.</p> <p>Meniscos lateral, com forma, contornos e intensidade de sinal normais.</p> <p>Laceração na transição corpo-corno posterior do menisco medial, com deslocamento da porção central para incisura intercondilar, porção periferia de aspecto irregular.</p> <p>Pequena quantidade de líquido livre intrarticular.</p> <p>Observa-se distensão da bursa gastrocnêmio-semimembranosa medial por coleção líquida.</p>	<p>1. Ruptura do menisco medial (alça de balde)</p> <p>2. Cisto de baker.</p> <p>3. Pequeno derrame articular</p>

Aditivamente, como estudo de caso, a graduanda realizou atividade de visita ao centro radiológico do hospital a fim de conhecer o cotidiano e o processo de geração dos laudos e, em fases seguintes (como, por exemplo, durante a análise manual dos laudos), seu regresso ao HCFMRP foi necessário para que um especialista confirmasse as características de preenchimento obtidas no trabalho.

5.2. Breve Definição do Domínio de Conhecimento dos Laudos

A radiologia é uma área da medicina que utiliza técnicas de raios X, isótopos radioativos e radiações não-ionizantes, como os ultra-sons para a geração de imagens de órgãos, tecidos e organismos internos; tanto para fins diagnósticos como para fins terapêuticos.

A ressonância magnética nuclear foi escolhida como técnica de estudo neste projeto e consiste na reconstituição das imagens anatômicas, utilizando a propriedade de certos núcleos atômicos por se comportarem simultaneamente como pequenos ímãs devido à absorção ressonante de energia eletromagnética na faixa das ondas de rádio. Sendo assim, as imagens anatômicas são geradas, pois como o campo magnético gerador da energia é levemente afetado pelos débeis campos eletromagnéticos gerados a partir dos elétrons envolvidos nas ligações químicas no ambiente químico da vizinhança do núcleo, cada núcleo responde diferentemente de acordo com sua localização no objeto em estudo, atuando como uma sonda sensível a estrutura onde se situa.

Como citado em seções anteriores, o joelho foi a estrutura anatômica escolhida e indicada por um especialista para o presente estudo de caso. Esta região do corpo humano, primeiramente, é uma articulação sinovial do tipo gínglimo (em dobradiça) que permite flexão e extensão; entretanto, os movimentos da articulação são combinados com deslizamento e rolamento, e com rotação sobre um eixo vertical. Embora a articulação do joelho seja bem construída sua função é comumente prejudicada em hiperextensão.

Esta região anatômica é detentora de uma estrutura óssea composta pela patela, tíbia, fíbula e fêmur, na qual somente a fíbula não está fisicamente envolvida na articulação. Além da estrutura óssea, o joelho possui um conjunto de ligamentos, que impedem os deslocamentos inapropriados dos ossos e hiperextensões das articulações. O ligamento colateral tibial, mais fraco do que o ligamento colateral fibular, é mais frequentemente danificado juntamente com o menisco medial, essas dilacerações ocorrem, principalmente, durante a prática de esportes de contato.

Já se referindo aos ligamentos cruzados, o ligamento cruzado anterior (LCA) é tido como o mais fraco, este é responsável por impedir o deslocamento posterior do fêmur sobre a tíbia e a hiperextensão da articulação do joelho, ou seja, quando a articulação é fletida formando um ângulo reto, a tíbia não pode ser tracionada anteriormente porque é contida pelo LCA. Aditivamente, o ligamento cruzado posterior impede o deslocamento anterior do fêmur sobre a tíbia ou o deslocamento posterior da tíbia sob o fêmur.

Outros elementos constituintes do joelho são os músculos, tendões, cartilagem, meniscos e cornos. Os músculos e tendões estão diretamente relacionados ao movimento, uma vez que são eles os geradores dessa ação, enquanto as cartilagens, meniscos e cornos são responsáveis por amortecer os impactos sofridos por fricção e sustentação desta região anatômica [28].

5.2.1. Identificação das Características de Preenchimento

Os laudos no HCFMRP são gerados de forma eletrônica e, na maioria das vezes, seguindo um padrão de preenchimento apelidado por laudo modelo. A utilização desse padrão é necessária como sugestão pelo departamento de imagens do hospital para que os laudos sejam gerados apresentando alto nível descritivo de informação. Logo, esses documentos devem expor achados de anormalidades além de relatar as normalidades analisadas. Concomitantemente, o laudo modelo especifica as regiões, estruturas e características que devem ser observadas para a obtenção do diagnóstico.

Outro fato que deve ser especificado é que para cada tipo de exame e sua respectiva região analisada é relacionado um modelo de laudo. Na tabela 6, a seguir, é apresentado o laudo modelo para o exame de Ressonância Nuclear Magnética de Joelho.

Tabela 6: Laudo Modelo para o exame de Ressonância Magnética de Joelho

Modelo de descrição de laudo
<i>Estruturas ósseas regionais com morfologia e intensidade de sinal preservados.</i>
<i>Cartilagem articular com espessura e sinal conservados.</i>
<i>Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações.</i>
<i>Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais.</i>
<i>Meniscos medial e lateral, com forma, contornos e intensidade de sinal normais.</i>
<i>Fossa poplíteia sem alterações.</i>

Após a análise dos laudos obtidos foi possível gerar estruturas hierárquicas das entidades anatômicas examinadas, das modalidades de exame, das observações e das regiões anatômicas. Também foi possível obter outros tipos hierárquicos que podem ser utilizados em diferentes domínios de informação como, por exemplo, léxicos temporais, quantitativos, qualitativos, posicionais, causalidade, equivalência, normalidade, negação, incerteza, desconhecimento e presença. E, por fim, foi organizada a maior quantidade possível de palavras referentes a achados e anormalidades relatadas nos laudos.

Essas hierarquias são fundamentais para a base de construção da estrutura de armazenamento proposta neste projeto, pois quanto mais granulosidade houver na hierarquia do domínio lexical estudado, mais facilmente essas informações serão registradas de forma estruturada.

Foram realizadas visitas ao CCIFM e à clínica Documenta do Hospital São Francisco para que os especialistas avaliassem e validassem o léxico desenvolvido a partir das características de preenchimento obtidas no trabalho.

5.3. Resultados do Pré-processamento dos Laudos

Para dar continuidade ao estudo de termos e semânticas dos elementos textuais de laudos radiológicos, a etapa seguinte desta metodologia consistiu no pré-processamento dos textos de ressonância magnética em joelho e do conjunto léxico desenvolvido com as ferramentas *PreText*.

Esta etapa iniciou-se com o pré-processamento de 1670 arquivos de extensão *.txt* utilizando a versão 1 do *PreText*., tornando-se possível a verificação da presença de uma coleção relativamente pequena de palavras na maioria dos laudos, devido ao estilo descritivo aconselhado pelo hospital.

Como dito em seções anteriores, esta versão do *PreText* gera alguns resultados indesejados para o presente trabalho, tanto na geração de *tokens*, como na eficiência de sua execução. Assim, a versão 2 desta ferramenta foi estudada e utilizada. Seus resultados foram mais adequados, contudo a forma com que ambas as versões os expõem (tabela atributo-valor) não agrega informação e semântica, por relatarem apenas a presença ou frequência de termos em determinado documento. Ou seja, listar os termos presentes não permite afirmar onde um achado (anormalidade) foi identificado pelo radiologista. Para tanto, decidiu-se por estudar e adotar conceitos de ontologia.

Também se observou, nesta etapa da metodologia, que seria adequado remover algumas palavras do conjunto de *stopwords*, uma vez que certas palavras como as de negação e quantidade podem influenciar na semântica das sentenças. As palavras desconsideradas *stopwords* podem ser observadas no Anexo A.

Ademais, notou-se necessário desenvolver algum método auxiliar que pré-processe os textos antes de serem repassados para a ferramenta *PreTexT* versão 2. Esse pré-processamento consiste na remoção dos acentos nas palavras e na substituição de letras maiúsculas por letras minúsculas¹⁵.

5.4. Resultados da Modelagem de uma Ontologia

A ontologia proposta possui nove classes (entidades) relacionadas entre si: *Característica da Modalidade*, *Diagnóstico*, *Entidade Anatômica*, *Estado*, *Modalidade*, *Laudo*, *Observação*, *Região Anatômica* e *Sentença*.

Cada uma das nove classes representa um dos principais termos que compõem os laudos médicos, e seu conceito é único. A classe *Característica de Modalidade* é composta pelo conjunto de palavras que especifica os procedimentos de um determinado exame, por exemplo, o exame de tomografia pode ser realizado com a administração de contraste ou com um corte axial. A classe *Diagnóstico* representa todo o campo textual que relata a conclusão do laudo. A *Entidade Anatômica* é constituída por todas as estruturas anatômicas humanas estudadas neste trabalho e organizada de forma hierárquica. Já o *Estado* determina em que condições o elemento examinado está: Normal, Anormal ou Não Visualizado. A classe *Modalidade* especifica qual modalidade de exame foi realizada. *Laudo* representa todo o campo textual que descreve o laudo. *Observação* é a classe que contém os elementos visualizados em cada entidade anatômica, essa classe pode relatar características como intensidade de sinal e contorno; ou pode ser um achado, ou seja, uma observação patológica. A *Região Anatômica* é uma classe composta, hierarquicamente, por um conjunto de termos que divide em regiões determinada entidade anatômica. E por fim, a entidade *Sentença* representa uma frase existente em uma descrição de laudo ou em um diagnóstico.

Na Figura 10 é apresentada parte da ontologia. Apesar de ser possível observar apenas as especializações cabeça e pescoço e membro inferior da classe Entidade Anatômica na caixa especificada pelo número 1, essa classe subdivide-se em 4 subclasses, cabeça e pescoço,

¹⁵ Estudando o algoritmo da ferramenta *PreTexT* versão 2, observou que esta situação é tratada, contudo não ocorre o resultado esperado.

membro inferior, membro superior e tronco. Nessa figura, também é exposta a descrição da classe Entidade Anatômica na caixa 2 e, ainda, os relacionamentos dessa classe com as demais (caixa 3).

A granulosidade da especialização foi realizada conforme o recomendado pelos especialistas. Os relacionamentos de alto nível entre as classes são:

- 1- *Característica da Modalidade eh característica de Modalidade*
- 2 *Diagnóstico refere-se a um Laudo*
- 3 *Diagnóstico possui algumas Sentenças*
- 4 *Entidade Anatômica eh examinada por uma Modalidade*
- 5 *Entidade Anatômica esta presente em algumas Sentenças*
- 6 *Entidade Anatômica possui um Estado*
- 7 *Entidade Anatômica possui algumas Regiões Anatômicas*
- 8 *Entidade Anatômica possui uma Observação*
- 9 *Estado esta presente em algumas Sentenças*
- 10 *Estado refere-se a algumas Entidades Anatômicas*
- 11 *Modalidade eh caracterizada por algumas Características de Modalidade*
- 12 *Modalidade examina algumas Entidades Anatômicas*
- 13 *Modalidade gera um Laudo*
- 14 *Laudo possui algumas Sentenças*
- 15 *Laudo possui um Diagnóstico*
- 16 *Observação esta presente em algumas Sentenças*
- 17 *Observação refere-se a algumas Entidades Anatômicas*
- 18 *Região Anatômica refere-se a algumas Entidades Anatômicas*
- 19 *Região Anatômica esta presente em algumas Sentenças*
- 20 *Sentença refere-se a um Diagnóstico*
- 21 *Sentença refere-se a um Laudo*
- 22 *Sentença possui um Estado*
- 23 *Sentença possui uma Observação*
- 24 *Sentença possui uma Entidade Anatômica*
- 25 *Sentença possui uma Região Anatômica*

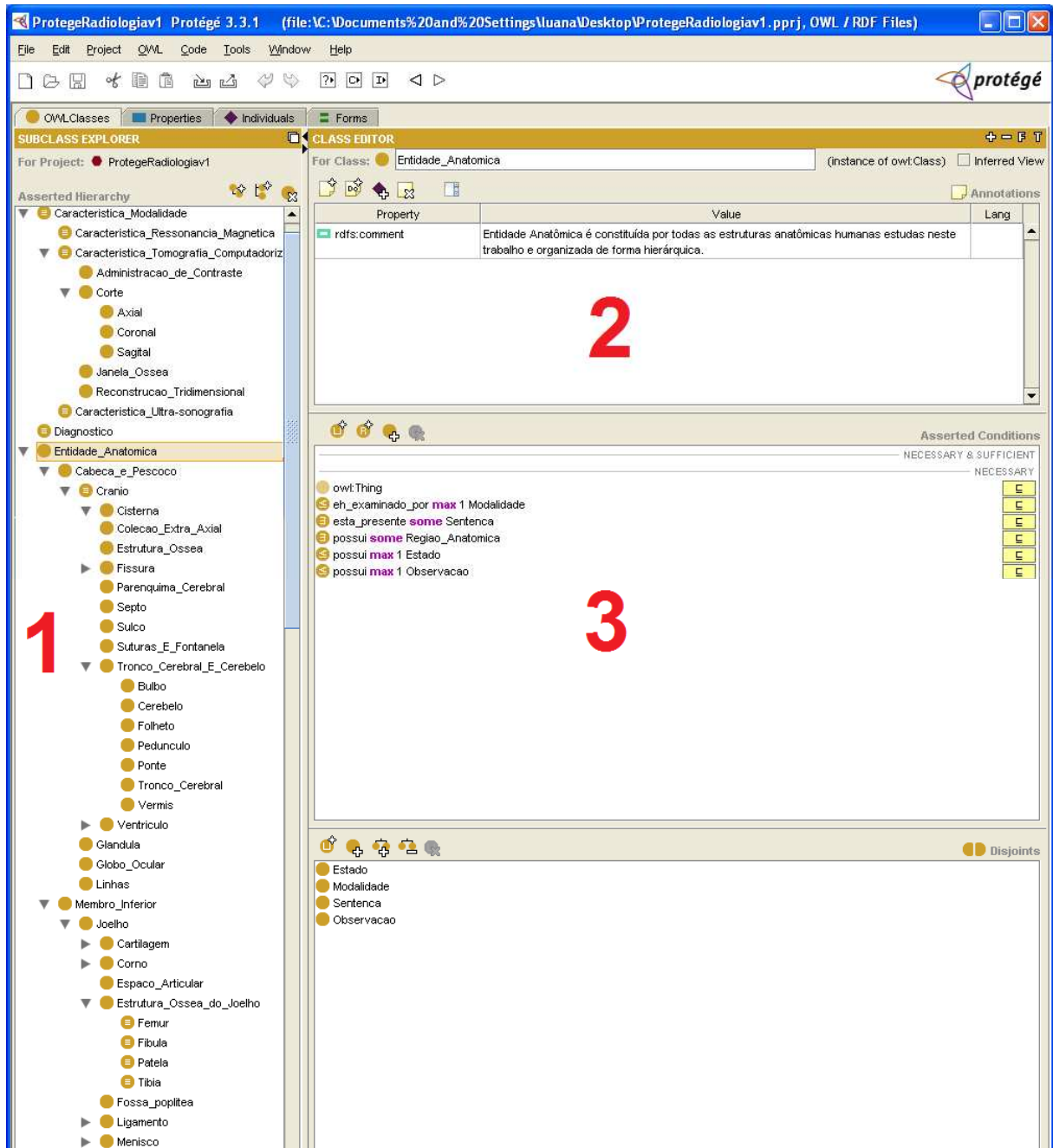


Figura 10: Visão parcial da ontologia proposta

5.5. Resultado da Modelagem e Alimentação da Base de Dados

Esta base foi apelidada de *RadOn*, por representar uma ontologia radiológica - “*Radiology Ontology*”. Como observado na figura 11, o *RadOn* possui oito tabelas relacionadas entre si para representação das classes abstraídas da ontologia: *Laudo*, *Descrição_Textual*, *Sentença*, *Sentença_has_Termo*, *Termo*, *Conceito*, *Item_Termo*, *Um_Grama*.

A tabela *Laudo* é uma representação dos laudos existentes nos hospitais como no HCFMRP, essa tabela contém dados sobre o identificador do laudo (*idLaudo*), a data em que esse exame foi realizado, o médico que o realizou, o paciente examinado, a região anatômica que se deseja examinar e a modalidade de exame realizada.

O texto de diagnóstico ou de descrição do laudo é armazenado na tabela *Descrição_Textual*, que se relaciona com a tabela *Laudo*, pois uma descrição textual pertence a um laudo e um laudo possui duas descrições textuais – descrição do laudo e o diagnóstico. Nessa tabela estão armazenados os dados sobre o identificador do texto (*idDescri_Textual*), o identificador do laudo que esta pertence (*Laudo_idLaudo*), um atributo que especifica se este texto é um diagnóstico ou não e, por fim, o texto deste relato.

Foi tomada a estratégia de fragmentar o texto em sentenças, uma vez que estas detêm as informações de um texto. Portanto necessitou-se da criação da tabela *Sentenca*, que armazena os dados de identificador do texto (*Descr_laudo_idDescr_Laudo*) e o identificador da sentença (*idSentenca*).

Esta informação contida em uma sentença é oriunda de um conjunto de termos e não somente de palavras, pois, exemplificando, a presença da palavra “cruzado” em uma sentença designa um conceito de algo em forma de cruz, intersecção; porém esta palavra no conjunto “ligamento cruzado anterior” deve ser interpretada como parte de um termo cuja semântica refere-se à denominação de um importante ligamento do joelho.

Logo, é imprescindível o relacionamento entre a tabela *Sentenca* e *Termo*, que é representado por um relacionamento N:M gerando a tabela *Sentenca_has_Termo*, a qual relata que um termo pode estar presente em uma ou mais sentenças e, uma sentença pode possuir um ou mais termos. Cada termo desse domínio de conhecimento é armazenado na tabela *Termo* que armazena os dados de identificador do termo, identificador do conceito e um elemento discriminante¹⁶.

Como evidenciado na modelagem do *RadOn*, um conceito pode estar presente em mais de um termo, por exemplo, a sigla JE e joelho esquerdo possuem o mesmo conceito, porém são escritos com caracteres diferentes; sendo assim necessária a criação da tabela *Conceito* que registra os dados sobre o identificador do conceito (*idConceito*), a definição deste, o nome e o tipo.

¹⁶ Este discriminante foi evidenciado, pois um termo pode possuir mais de um conceito dependendo da modalidade de exame realizada, como foi pouco relatado pelos especialistas durante idas ao HCFMRP este elemento deverá ser trabalhado em projetos futuros

Os conceitos presentes no *RadOn* retratam tanto os elementos representados na ontologia anteriormente relatada, quanto os termos de léxicos temporais, quantitativos e outros listados durante o estudo das características de preenchimento. Esses conceitos podem assumir os seguintes valores considerados relevantes neste trabalho: entidade_anatomica, regio_anatomica, modalidade, observacao, caract_anatomica, caract_modalidade, quantitativo, negacao, presenca, consequencia, adversativo, aditivo, caract_imagem, incerteza, qualificador, temporal e unidade.

Como exemplificado anteriormente, um termo pode ser formado de uma ou mais palavras e como, pressuposto na modelagem da ontologia, os textos dos laudos seriam pré-processados pelo *PreText*, logo, o *RadOn* retrata que um termo pode ser formado de um ou mais gramas a partir da tabela *Item_Termo*, que especifica o identificador do termo (*Termo_idTermo*), o identificador do grama (*Um_grama_id_Um_grama*) e a seqüência ou posição em que esse grama está presente no termo referenciado.

E por fim, a tabela *Um_Grama* armazena os gramas (grams) de tamanho um, resultantes do processamento das palavras presentes no léxico desenvolvido. Essa tabela contém o identificador do grama (*idUm_Grama*) e o grama que este representa (*palavra_grama*).

Toda a modelagem desta base pode ser visualizada na figura 11.

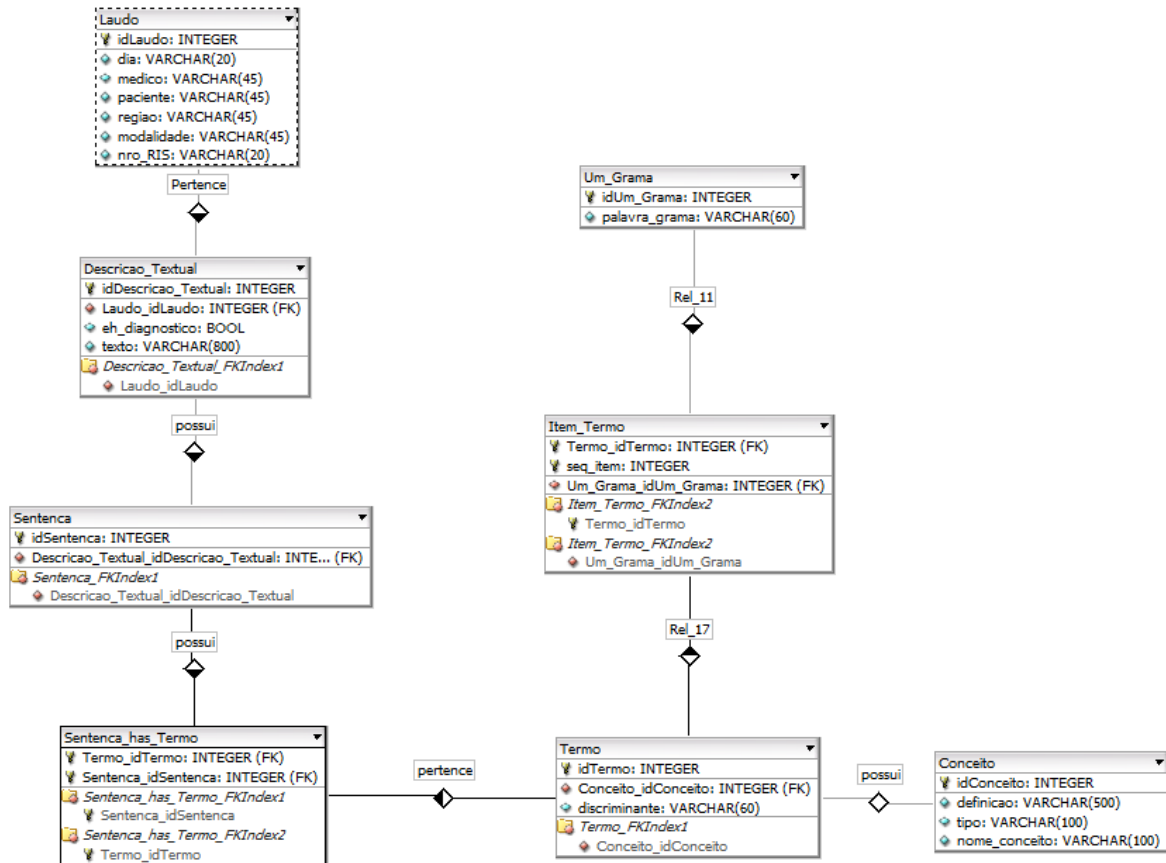


Figura 11: Modelagem da base de dados RadOn, proposta para armazenar de forma estruturada os laudos e diagnósticos médicos

5.6. Resultados da Ferramenta E-Rad

Para exemplificar a saída dos módulos da ferramenta *E-Rad*, será utilizado o laudo ilustrado na tabela 7, que apresenta suas descrições textuais não estruturadas e outros dados de identificação. Este laudo foi escolhido por ser de grande ocorrência dentre os dados fornecidos, além de sua complexidade semântica que pode ser compreendida na descrição de laudo, pois, quando citado “Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial”, o que se deseja relatar é a ocorrência dos Ligamentos Cruzados Anterior, Ligamento Cruzado Posterior, Ligamento Colateral Lateral e Ligamento Colateral Medial.

Tabela 7: Exemplo de laudo radiológico utilizado para demonstrar os resultados da ferramenta E-Rad

Identificador	Nro_RIS	Descrição do Laudo	Diagnóstico
2	"119-1"	<p>Estruturas ósseas regionais com morfologia e intensidade de sinal preservados.</p> <p>Cartilagem articular com espessura e sinal conservados.</p> <p>Tendões patelar e do quadriceps e demais estruturas do aparelho extensor sem alterações.</p> <p>Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais.</p> <p>Meniscos medial e lateral, com forma, contornos e intensidade de sinal normais Fossa poplíteia sem alterações.</p>	ID: Dentro dos limites da normalidade

5.6.1. Resultados do Módulo *TakeReport.pm*

A saída do módulo *TakeReport.pm* contém o identificador da descrição textual (*id_descr_text*), o texto e um elemento booleano (*eh_diagnóstico*) que determina se esta descrição é um diagnóstico ou uma descrição do laudo. Na figura 12 é exemplificada a saída deste módulo.

Como é possível observar na figura, o elemento booleano determina se a descrição textual refere-se a um diagnóstico, assumindo o valor “1”, ou a uma descrição de laudo, apresentando o valor “0”.

Comparando o texto apresentado na tabela 7 em relação à saída deste módulo, nota-se que o conjunto de caracteres, símbolos, assim como a disposição dos elementos são idênticos.

```

luana@luana-laptop: ~/Desktop/Elruconi
Arquivo Editar Ver Terminal Abas Ajuda

luana@luana-laptop:~/Desktop/Elruconi$ perl Elruconi.pl
Descricao textual: ID: Dentro dos limites da normalidade

Identificador: 3

Eh diagnostico: 1

-----

Descricao textual: Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilage
m articular com espessura e sinal conservados. Tendões patelar e do quadriceps e demais estruturas do aparelho e
xtensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com fo
rma, orientação e intensidade de sinal habituais. Meniscos medial e lateral, com forma, contornos e intensidade
de sinal normais Fossa poplítea sem alterações.

Identificador: 4

Eh diagnostico: 0

-----

```

Figura 12: Resultado do módulo *TakeReport.pm*

5.6.2. Resultados do Módulo *RemoveAccent.pm*

O resultado do módulo *RemoveAccent.pm* consiste no resultado de todas as substituições aplicadas ao texto enviado pelo *E-Rad.pl*, ou seja, remoção de acentos, substituição das letras maiúsculas por minúsculas e o tratamento da ocorrência da palavra "e". A saída *RemoveAccent.pm* é demonstrada na figura 13.

Foi desenhado nesta figura um retângulo azul para ilustrar a ocorrência da substituição da palavra “e” por “and”. E a palavra “alteracoes”, foi sublinhada em laranja para demonstrar o tratamento de acentos.

```

luana@luana-laptop: ~/Desktop/Elruconi
Arquivo Editar Ver Terminal Abas Ajuda

luana@luana-laptop:~/Desktop/Elruconi$ perl Elruconi.pl
Resultado do módulo RemoveAccent.pm

id: dentro dos limites da normalidade

-----

Resultado do módulo RemoveAccent.pm

estruturas osseas regionais com morfologia and intensidade de sinal preservados. cartilagem articular com espessura and sinal
conservados. tendoes patelar and do quadriceps and demais estruturas do aparelho extensor sem alteracoes. ligamentos cruz
ados anterior and posterior , bem como os colaterais lateral and medial com forma , orientacao and intensidade de sinal habitua
is. meniscos medial and lateral , com forma , contornos and intensidade de sinal normais fossa poplitea sem alteracoes.

-----

luana@luana-laptop:~/Desktop/Elruconi$

```

Figura 13: Resultado do módulo *RemoveAccent.pm*

5.6.3. Resultados do Módulo *Start.pm*

O resultado da execução do módulo *Start.pm* é ilustrada na figura 14. Esse módulo é responsável por pré-processar os textos de laudos e enviar seu resultado ao módulo *E-rad.pl*. Esse resultado é composto por um conjunto de gramas e vírgulas ordenados tais quais suas respectivas palavras de origem, assim como por *breaks* (separadores) identificando o término de cada sentença. Na figura 14, o retângulo azul ressalta a ocorrência de um *break*, enquanto a sentença sublinhada em laranja destaca a alta complexidade semântica existente nessa sentença. Esta terá sua estruturação apresentada em resultados seguintes.



```

luana@luana-laptop: ~/Desktop/Elruconi
Arquivo  Editar  Ver  Terminal  Abas  Ajuda

luana@luana-laptop:~/Desktop/Elruconi$ perl Elruconi.pl

Resultados do módulo Start.pm

    id : dentr limit normal

-----

Resultados do módulo Start.pm

    estrutur osse region com morfolog and intens sinal preserv : cartilag articul com espessur and s
inal conserv : tendo patel and quadriceps and dem estrutur aparelh extens sem alterac : ligament cruz
anteri and poster , bem com colater lateral and medial com form , orientaca and intens sinal habitu :
menisc medial and lateral , com form , contorn and intens sinal norm foss poplite sem alterac :

-----

```

Figura 14: Resultado do módulo *Start.pm*

5.6.4. Resultados do Módulo *TakeTermSentence.pm*

O resultado das verificações de identificação dos termos existentes na sentença, realizadas pelo módulo *TakeTermSentence.pm*, é composto por um conjunto de termos ordenados conforme seu grama final (o último elemento que compõe este termo) e seu respectivo tipo conceitual. Na figura 15 é possível observar estes resultados.

```

luana@luana-laptop: ~/Documentos/ERad
Arquivo Editar Ver Terminal Abas Ajuda
Use of uninitialized value $sentences in split at Elruconi.pl line 59.
saída do módulo:
identificador dos termos:
conceito dos termos:

Use of uninitialized value $sentences in string ne at Elruconi.pl line 70.
saída do módulo:
identificador dos termos: 149 388
conceito dos termos:regiao_anatomica observacao

Use of uninitialized value $oldSentences in string ne at Elruconi.pl line 70.
Use of uninitialized value in numeric ne (!=) at ProgressBar.pm line 58.
saída do módulo:
identificador dos termos: 203 204 114 371 37 299 437
conceito dos termos: entidd_anatomica qualificador presenca caract_anatomica aditivo caract_anatomica observacao

saída do módulo:
identificador dos termos: 82 114 201 37 495
conceito dos termos: entidd_anatomica presenca caract_anatomica aditivo caract_anatomica

saída do módulo:
identificador dos termos: 540 37 541 37 211 488 30
conceito dos termos: entidd_anatomica aditivo entidd_anatomica aditivo quantitativo negacao observacao

saída do módulo:
identificador dos termos: 332 37 333 658 62 329 37 330 114 241 658 393 37 299 676
conceito dos termos: entidd anatomica aditivo entidd anatomica virgula aditivo entidd anatomica aditivo e
ntidd anatomica presenca caract anatomica virgula caract anatomica aditivo caract anatomica observacao

saída do módulo:
identificador dos termos: 366 37 365 658 114 241 658 122 37 299 387 250 488 30
conceito dos termos: entidd_anatomica aditivo entidd_anatomica virgula presenca caract_anatomica virgula
caract_anatomica aditivo caract_anatomica observacao entidd_anatomica negacao observacao

luana@luana-laptop:~/Documentos/ERad$

```

Figura 15: Resultados do módulo *TakeTermSentence.pm*

Como observado na figura 15, foram destacados tanto o conjunto de identificadores de termos (sublinhado laranja) quanto seus respectivos tipos de conceito (sublinhado azul) da sentença “Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais.”.

Como observado na mesma figura, o conceito dos termos contidos nesta sentença foi corretamente listado. Destaca-se o fato de que os conceitos referentes à palavra “e” e ao símbolo de vírgula também foram preservados e denominados, respectivamente, por aditivo e vírgula.

5.6.5. Resultados do Módulo *MakeSentences.pm*

A fim de manter a semântica do exemplo¹⁷ estudado, torna-se necessária o processamento da sentença recebida, resultando na geração de novas sentenças. Pois, como relatado neste trabalho, se apenas a aplicação da abordagem *bag-of-words* for realizada a esta sentença, não

¹⁷ “Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais”

será possível determinar a quais entidades anatômicas a característica “orientação” estará relacionada.

Os identificadores de termos sublinhados em laranja na figura 16 relatam a geração de algumas destas novas sentenças de termos, a partir do correto relacionamento identificado entre os elementos existentes na sentença de origem.

Lembrando que o identificador 332 refere-se ao ligamento cruzado anterior, 333 ao ligamento cruzado posterior, 114 à palavra “com”, 241 à característica forma, 393 à orientação, 299 à característica intensidade de sinal e 676 à observação habitual, o resultado apresentado por este módulo consiste nas seguintes sentenças:

- Ligamento cruzado anterior com forma habitual
- Ligamento cruzado anterior com orientação habitual
- Ligamento cruzado anterior com intensidade de sinal habitual
- Ligamento cruzado posterior com forma habitual
- Ligamento cruzado posterior com orientação habitual
- Ligamento cruzado posterior com intensidade de sinal habitual

É relevante ressaltar que o relacionamento entre os outros termos existentes no exemplo estudado também foram realizados, contudo não pode ser visualizados devido à dimensão da figura.

Por fim, como relatado na figura 16, as novas sentenças assim como as sentenças que não precisaram ser processadas são inseridas na base de dados.

```

luana@luana-laptop: ~/Desktop/Elruconi
Arquivo Editar Ver Terminal Abas Ajuda
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
203 204 114 371 437
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
203 204 114 299 437
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
82 114 201
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
82 114 495
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
540
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
211 488 30
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
541
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
211 488 30
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
332 114 241 676
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
332 114 393 676
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
332 114 299 676
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
333 114 241 676
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn
333 114 393 676
Resultados do módulo MakeSentences.pm
Sentença de Termos a ser inserida no RadOn

```

Figura 16: Resultado do módulo *MakeSentences.pm*

5.7. Resultado de Utilização da Interface de Busca do *E-Rad*

A interface de busca da ferramenta *E-Rad* é de simples manipulação, sendo necessário apenas preencher o campo de busca com os dados que se deseja encontrar em um laudo. Caso se deseje realizar uma busca mais específica em relação ao tipo de modalidade e exame que o laudo se refere, o usuário deve selecionar, dentre as opções, o tipo de exame do laudo. A seguir, serão apresentados alguns exemplos e resultados de busca, realizados nesta interface, assim como a justificativa de sua demonstração.

5.7.1. Busca Utilizando Uma Palavra

A busca ilustrada na figura 17 foi realizada utilizando uma única palavra na sentença de busca. No caso, “cisto” foi escolhido para testar sua identificação no laudo e de possíveis termos com o mesmo conceito, como imagem cística.

Identificador do laudo	número RIS do exame	Descrição Textual
10	204-1	1. Ruptura do menisco medial. 2. <u>Cisto</u> sinovial com sinais de inflamação.
44	2464-1	1. Osteonecrose do femur e da tibia com perda de subânfancia óssea no femur. 2. Lesão posterior do menisco medial (<u>cisto</u> intra-meniscal ?). 3. Derrame articular com cisto de Baker.
64	4250-1	1- Ruptura horizontal do menisco lateral com <u>cisto</u> para-meniscal. 2- Tendinite patelar discreta.
20	396-1	Ausência cirúrgica da patela. Aumento de intensidade adjacente ao tendão quadriceps patelar. Tendões patelar e do quadriceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Presença de área de hipersinal nas sequencias com densidade de prótons, no interior dos meniscos, sem extensão para a superfície. Imagem cística hipointensa em T1 e hiperintensa em T2, na fossa poplitea com área de continuidade com a sinóvia. Pequena quantidade de liquido livre intra articular.
47	2766-1	Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadriceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Traço de hipersinal horizontal na substância do corno posterior do menisco medial do joelho direito, sem atingir a superfície. Ausência da imagem do corno anterior do menisco lateral em sua topografia habitual, notando imagem de isossinal ao menisco junto ao corpo e corno posterior do menisco lateralmente sugerindo fragmento do corno anterior. Imagem cística herniando da fossa poplitea entre o ventre medial do músculo gastrocnêmio e o tendão do músculo semimembranoso.
53	4193-1	Alteração de sinal do côndilo femural e platô tibial mediais. Pequena área de aspecto cístico subcondral no fêmur. Presença de osteófitos na patela e no compartimento medial e eminências intercondilleanas. Cartilagem articular de contornos irregulares com áreas de afilamento. Tendões patelar e do quadriceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Meniscos lateral, com forma, contornos e intensidade de sinal normais. Menisco medial deformado de dimensões reduzidas e contornos irregulares sem áreas evidentes de fratura nos segmentos remanescentes. Fossa poplitea sem alterações.

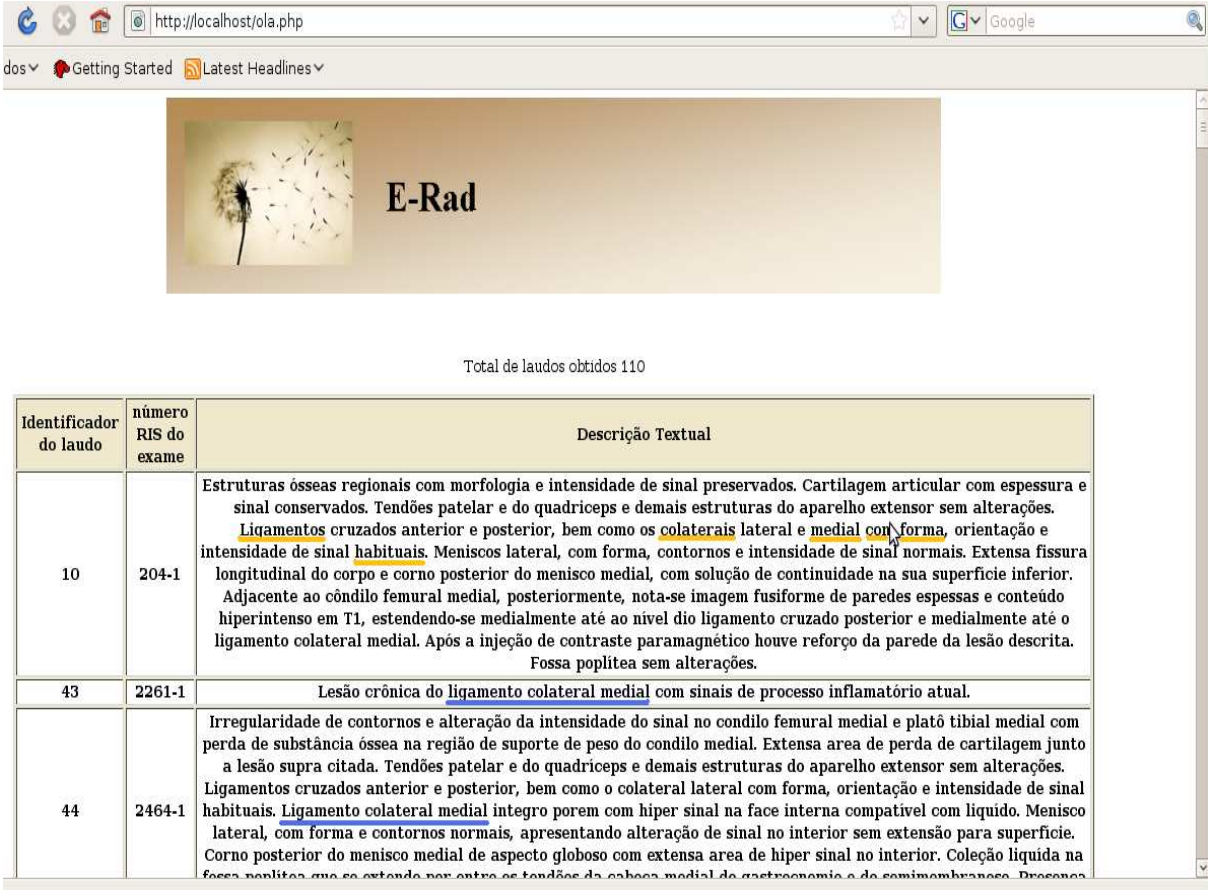
Figura 17: Resultado da busca por laudo que relatam a ocorrência de cisto

Observa-se que foram obtidos seis resultados, todos contendo relatos de achados de cisto, porém nota-se 3 resultados peculiares, no entanto, previstos, a serem: o quarto e quinto resultados com um achado de Imagem cística, destacado em laranja, e o sexto resultado apresentando um achado de área de aspecto cístico, sublinhado em laranja.

Estes resultados demonstram que as buscas pelos próprios termos e seus sinônimos estão ocorrendo conforme as expectativas.

5.7.2. Busca com um Conjunto de Palavras

Após a busca por uma palavra, testou-se a busca por um conjunto de palavras. Foi escolhido o conjunto “ligamento colateral medial com forma habitual” para identificar esses termos e/ou suas semânticas nos laudos. O resultado pode ser observado na figura 18.



Total de laudos obtidos 110

Identificador do laudo	número RIS do exame	Descrição Textual
10	204-1	Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadriceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Meniscos lateral, com forma, contornos e intensidade de sinal normais. Extensa fissura longitudinal do corpo e corno posterior do menisco medial, com solução de continuidade na sua superfície inferior. Adjacente ao côndilo femural medial, posteriormente, nota-se imagem fusiforme de paredes espessas e conteúdo hiperintenso em T1, estendendo-se medialmente até ao nível do ligamento cruzado posterior e medialmente até o ligamento colateral medial. Após a injeção de contraste paramagnético houve reforço da parede da lesão descrita. Fossa poplítea sem alterações.
43	2261-1	Lesão crônica do ligamento colateral medial com sinais de processo inflamatório atual.
44	2464-1	Irregularidade de contornos e alteração da intensidade do sinal no condilo femural medial e platô tibial medial com perda de substância óssea na região de suporte de peso do condilo medial. Extensa área de perda de cartilagem junto a lesão supra citada. Tendões patelar e do quadriceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como o colateral lateral com forma, orientação e intensidade de sinal habituais. Ligamento colateral medial íntegro porém com hiper sinal na face interna compatível com líquido. Menisco lateral, com forma e contornos normais, apresentando alteração de sinal no interior sem extensão para superfície. Corno posterior do menisco medial de aspecto globoso com extensa área de hiper sinal no interior. Coleção líquida na fossa poplítea que se estende por entre os tendões da cabeça medial do gastrocnêmio e da semitendíneo. Processo

Figura 18: Busca para teste do módulo *MakeTermSentence.pm*

Os resultados obtidos são convenientes à expectativa de busca, totalizando 110 laudos identificados por possuírem algum dos termos da sentença de busca ou termos de mesmo conceito.

Uma vez que a sentença de busca é composta por 4 termos, a serem: “ligamento colateral medial”, “com”, “forma” e “habitual”, o sistema deverá retornar todos os laudos que possuam tais termos ou termos de mesmo conceito. Como observado na figura 18, nota-se no primeiro resultado a presença de todos os termos, sublinhados em laranja. Vale ressaltar que, por mais que o conjunto de palavras “ligamento colateral medial” se encontre separados, porém na mesma sequência, o termo “ligamento colateral medial” foi identificado pelo

sistema. Já o segundo laudo, identificado pela busca, apresenta, somente, o primeiro termo da busca (“ligamento colateral medial”). Contudo a semântica presente na sentença desse termo não é a mesma da semântica de busca, uma vez que a primeira relata um processo inflamatório nessa entidade anatômica, enquanto a segunda retrata o estado (habitual) dessa característica anatômica (forma).

Após estudo e rastreamento do algoritmo, verificou-se que a ocorrência destes resultados são conseqüentes à estratégia tomada durante a implementação do sistema de busca para reduzir a complexidade computacional. Como a intenção do usuário é identificar todos os laudos que possuam a mesma semântica da sentença de busca e não os laudos que possuam somente um dos termos dessa sentença, como apresentado pelo segundo laudo identificado na figura 18, propõe-se que futuros trabalhos sejam realizados para melhorar esta ferramenta.

5.7.3. Busca Baseada em um Conceito

Nessa etapa, buscou-se por laudos que contenham o termo “hipointensidade”. Esta busca foi realizada sabendo-se que esse termo possui um ou mais termos com o mesmo conceito armazenados na base de dados. Portanto, o objetivo desta etapa foi o de verificar a capacidade da ferramenta em encontrar laudos a partir do conceito dos termos de busca, e não das palavras que os compõem. Validando-se, dessa forma, a necessidade do desenvolvimento e da utilização de uma ontologia.

Na figura 19 são apresentados os nove resultados da busca por “hipointensidade”.

Identificador do laudo	número RIS do exame	Descrição Textual
4	122-1	Imagem arredondada medindo 0,8cm de diâmetro em extremidade distal femural na região intercondileana, hipointenso em T1 e hiperintenso em T2 com supressão de gordura. osteófitos marginais discretos em côndilos femorais. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Em topografia do ligamento cruzado anterior nota-se imagem hiperintensa de contornos borrados. Ligamentos cruzados posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Aumento da intensidade de sinal dos meniscos medial e lateral com direção horizontal sem atingir a superfície articular. Fossa poplíteia sem alterações.
11	210-1	Pequena área arredondada hiperintensa nas sequências pesadas em T2 e hipointensas em T1, de limites bem definidos e contornos discretamente irregulares, medindo 0,7cm, localizada na medula do 1/3 distal da diáfise femural. Discreta hiperintensidade no côndilo femural lateral. Cartilagem articular com espessura e sinal conservados. Discreta hiperintensidade ao redor da inserção patelar dos tendões do quadríceps e patelar, os quais apresentam espessura e intensidade de sinal normais. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Meniscos medial e lateral, com forma, contornos e intensidade de sinal normais Fossa poplíteia sem alterações.
15	292-1	Discreta irregularidade de contornos com área de sinal heterogêneo no côndilo femural medial, corpúsculo ósseo livre adjacente, associada a lesão condral. Área de afilamento com aumento de sinal na porção central da cartilagem patelar. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Meniscos medial e lateral, com forma, contornos e intensidade de sinal normais. Presença de coleção laminar hipointensa em T1 e hiperintensa em T2 posterior ao côndilo femural medial e fossa poplíteia.
20	396-1	Ausência cirúrgica da patela. Aumento de intensidade adjacente ao tendão quadríceps patelar. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Presença de área de hipersinal nas sequências com densidade de prótons, no interior dos meniscos, sem extensão para a superfície. Imagem cística hipointensa em T1 e hiperintensa em T2, na fossa poplíteia com área de continuidade com a sinóvia. Pequena quantidade de líquido livre intra articular.
34	714-1	Imagens de contornos irregulares hipointensas em T1 e hiperintensas em T2, nas epífises distal do fêmur e proximal da tibia. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Aumento de sinal de aspecto globular nos cornos posteriores de ambos os membros. Fossa poplíteia sem alterações.

Figura 19: Resultado de busca pelo conceito “hipointensidade”

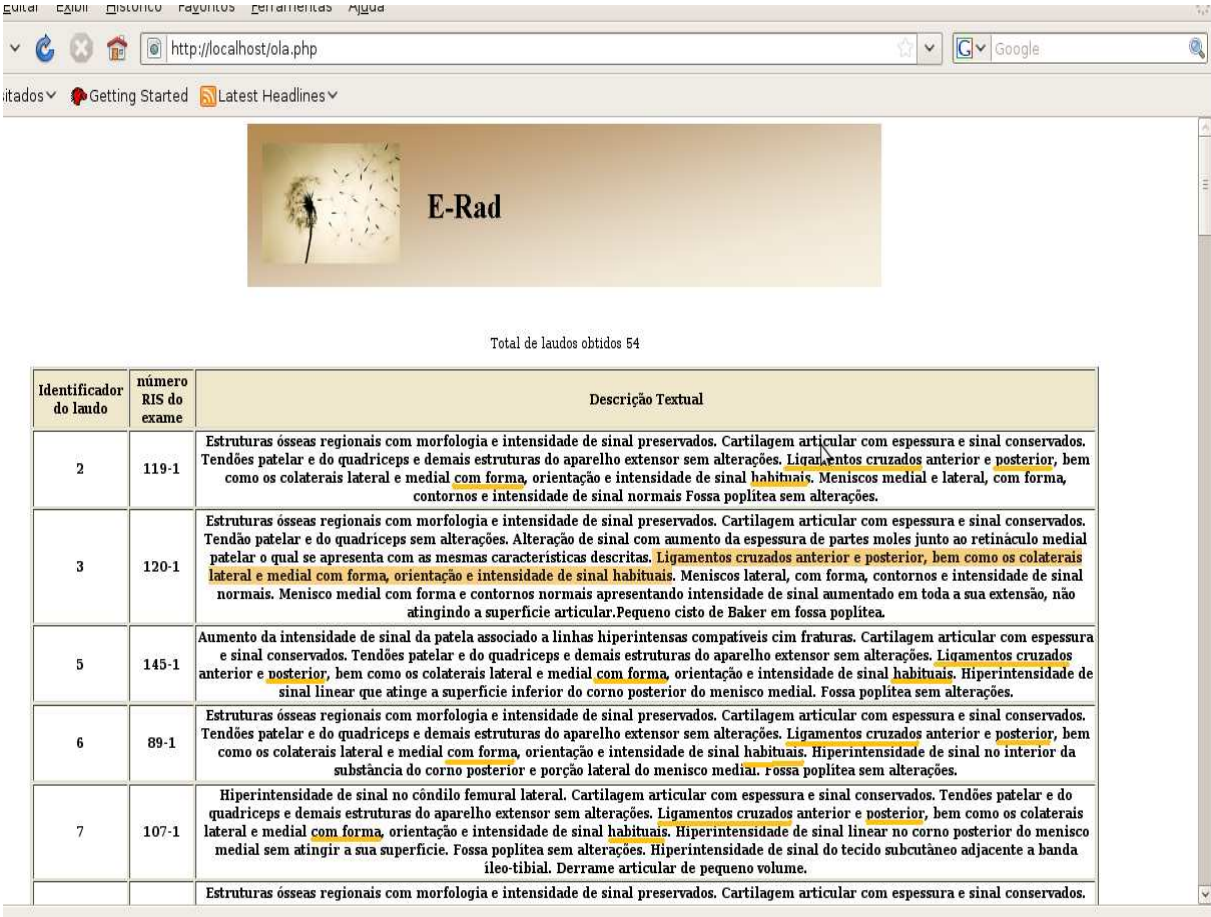
A partir das palavras sublinhadas na figura 19, é destacada a identificação de termos com mesmo conceito do elemento solicitado na busca, ou seja, enquanto a busca foi por “hipointensidade”, os resultados gerados foram “hipointensas”, “hipointenso”, “hipointensa”, palavras diferentes, porém com conceitos iguais. Pode-se, dessa forma, afirmar que o desenvolvimento e a utilização da ontologia fez-se necessário e benéfico, pois auxilia no processo de busca, dando mais flexibilidade ao usuário e, conseqüentemente, aumentando de forma consistente o número de laudos identificados para esta busca.

5.7.4. Busca Baseada em Dois Conceitos

Por fim, foi realizada uma pesquisa a partir de uma sentença de busca composta por um conjunto de termos: “LCP com forma preservada”. Dentre esses termos, dois são distintos entre si e possuidores de sinônimos que estão armazenados na base de dados. Além do mais, esta sentença é composta por um conjunto de termos que para sua identificação, em uma mesma sentença, será necessário a execução correta do módulo *MakeTermSentences.pm*.

Esta pesquisa foi realizada para verificar a capacidade da ferramenta em trabalhar sob várias condições conflituosas ao mesmo tempo. Na figura 20, são apresentados os cinquenta e quatro resultados da pesquisa por “LCP com forma preservada”.

Nota-se a correta identificação do termo “habitual”, por se tratar de um sinônimo de “preservada”. O mesmo fato ocorre na identificação do termo “ligamento cruzado posterior” como um sinônimo do termo “LCP”. Vale ressaltar que a ferramenta efetuou o mesmo procedimento apresentado na sub-seção 5.7.2 a qual demonstra a identificação do termo “ligamento colateral medial” com as palavras separadas, porém com mesma ordem sequencial. Na busca pelo termo “LCP com forma preservada”, o sinônimo “ligamento cruzado posterior” é apresentado com suas palavras separadas, porém sequencialmente ordenadas.



Total de laudos obtidos 54

Identificador do laudo	número RIS do exame	Descrição Textual
2	119-1	Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Meniscos medial e lateral, com forma, contornos e intensidade de sinal normais Fossa poplitea sem alterações.
3	120-1	Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilagem articular com espessura e sinal conservados. Tendão patelar e do quadríceps sem alterações. Alteração de sinal com aumento da espessura de partes moles junto ao retináculo medial patelar o qual se apresenta com as mesmas características descritas. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Meniscos lateral, com forma, contornos e intensidade de sinal normais. Menisco medial com forma e contornos normais apresentando intensidade de sinal aumentado em toda a sua extensão, não atingindo a superfície articular. Pequeno cisto de Baker em fossa poplitea.
5	145-1	Aumento da intensidade de sinal da patela associado a linhas hiperintensas compatíveis com fraturas. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Hiperintensidade de sinal linear que atinge a superfície inferior do corno posterior do menisco medial. Fossa poplitea sem alterações.
6	89-1	Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Hiperintensidade de sinal no interior da substância do corno posterior e porção lateral do menisco medial. Fossa poplitea sem alterações.
7	107-1	Hiperintensidade de sinal no côndilo femoral lateral. Cartilagem articular com espessura e sinal conservados. Tendões patelar e do quadríceps e demais estruturas do aparelho extensor sem alterações. Ligamentos cruzados anterior e posterior, bem como os colaterais lateral e medial com forma, orientação e intensidade de sinal habituais. Hiperintensidade de sinal linear no corno posterior do menisco medial sem atingir a sua superfície. Fossa poplitea sem alterações. Hiperintensidade de sinal do tecido subcutâneo adjacente a banda íleo-tibial. Derrame articular de pequeno volume.
		Estruturas ósseas regionais com morfologia e intensidade de sinal preservados. Cartilagem articular com espessura e sinal conservados.

Figura 20: Resultado da busca por “LCA com forma preservada”

CAPÍTULO 6. Discussão e Conclusões

6.1. Considerações Finais

Este trabalho foi realizado com o intuito de estudar e aplicar técnicas e conceitos de mineração de texto com a finalidade de estruturar laudos radiológicos. A estruturação de laudos radiológicos faz-se necessária para aperfeiçoar os recursos e aplicabilidades do SIR no HCFMRP, sem interferir no modo com que os profissionais da saúde realizam suas tarefas de descrição dos laudos, ou seja, permitindo que estes continuem preenchendo os laudos e diagnósticos a partir de textos abertos; e habilitando seus dados a serem utilizados em futuras ferramentas baseadas em Aprendizado de Máquina, Inteligência Artificial, assim como para sistemas complexos de busca.

Este trabalho iniciou com a revisão realizada pela graduanda em anatomia humana, anatomia clínica e radiologia. E, em seguida, foram realizados estudos com uma quantidade representativa de laudos referentes à Ressonância Magnética de Joelho, com o intuito de obter o máximo de informação referente às características de preenchimento existentes nestes tipos textuais. Com a mesma finalidade foram realizadas visitas a CCIFM no HCFMRP e à clínica Documenta no hospital São Francisco. Durante este estudo, gerou-se um léxico contendo as características de preenchimento analisadas e esse léxico foi avaliado por diversos radiologistas durante as visitas aos hospitais. Também durante esta etapa, foi possível observar que a maioria dos textos de descrições de laudos segue um modelo estabelecido pelo hospital, tendo o caráter detalhista, ou seja, relatam tanto anormalidades observadas quanto as normalidades.

Após toda esta contextualização e estudo do domínio de conhecimento da radiologia, foi realizado o estudo e aplicação de técnicas e conceitos de mineração de texto, utilizando as ferramentas *PreText* versão 1 e versão 2. Com estas aplicações observou-se que a abordagem *bag-of-words* não é adequadamente utilizada para a estruturação dos textos, pois essa abordagem consiste na listagem dos termos existentes nos textos, perdendo o relacionamento entre os termos e, conseqüentemente, perdendo as semânticas existentes nas sentenças. Por outro lado, as técnicas de remoção de *stopwords* e *stemming* são de grande valia para a presente proposta, pois diminui o conjunto de palavras a serem trabalhadas e, assim, diminui a complexidade dos processos.

Como estratégia para manter o relacionamento entre os termos, a graduanda realizou atividades extras de estudo e modelagem de uma ontologia, a qual foi utilizada como base para a modelagem da base de dados, *RadOn*, que é utilizada neste trabalho para registrar os elementos necessários para a estruturação dos laudos - como gramas, termos, conceitos - além de armazenar o resultado da estruturação realizado por este trabalho.

Por fim, para implementar todos estes estudos com o intuito de estruturar os laudos e adaptar os textos para a ferramenta de mineração de texto, tornou-se necessário o desenvolvimento de uma ferramenta, *E-Rad*. Esta realiza um pré-processamento dos textos removendo acentos e letras maiúsculas para aperfeiçoar a obtenção de gramas pela ferramenta *PreText* versão 2. Os gramas são gerados utilizando, principalmente, as técnicas de remoção de *stopwords* e *stemming*. A partir dos gramas, a ferramenta *E-Rad*, acessando a base de dados *RadOn*, identifica os termos que compõem cada sentença e processa estas com o objetivo de manter sua semântica pelo relacionamento entre componentes. A estruturação é finalizada com a alimentação da base de dados, relatando quais os termos existentes em determinada sentença e em qual descrição textual essa sentença está presente.

Aditivamente, uma interface de busca foi desenvolvida baseada na mesma lógica de processamento da ferramenta *E-Rad*, para que pesquisas por laudos fossem realizadas baseadas em sua estruturação.

6.2. Discussão

Se tomarmos como base os processos de busca efetuados atualmente no HCFMRP (buscas manuais dos laudos digitais), o sistema de busca aqui desenvolvido gerou resultados de grande valia, uma vez que a busca tornou-se mais eficiente e rápida, além de poder ser facilmente utilizada em futuros trabalhos de desenvolvimento de sistemas de auxílio ao diagnóstico, devido à estruturação resultante.

A eficiência na obtenção destes laudos a partir de uma sentença de busca é diretamente proporcional ao tamanho da base de dados, ou seja, quanto maior for a quantidade de termos existentes na base de dados, maior será a eficiência da ferramenta, uma vez que a ferramenta não identifica termos não armazenados na base de dados.

Também foi observado no presente trabalho que a base de dados *RadOn* está apta a receber dados referentes a outros tipos de exames, pois foi desenvolvida baseando-se numa ontologia radiológica. Portanto, mesmo que o domínio de conhecimento estudado neste trabalho seja referente a exames de Ressonância Magnética de Joelho, as classes da ontologia,

assim como os tipos conceituais existentes na tabela *Conceito* são significativas para uma relativa parte do domínio de conhecimento da Radiologia.

Em relação à alimentação da base com novos termos, não se considera adequado aplicar técnicas de auto-alimentação, pois, como é necessário um conhecimento prévio para delegar a que tipo conceitual cada termo se refere, torna-se necessário que esta atividade seja realizada por especialistas ou conhecedores da área.

Trabalhos futuros poderão ser desenvolvidos com o intuito de melhorar o sistema de busca a partir de laudos estruturados, regrando condições mais expressivas em relação ao conjunto de termos buscados, ou seja, deve-se realizar um estudo mais detalhado para criar condições de busca quanto à existência de termos de negação, ou mesmo retornar os laudos que possuam uma quantidade mínima de termos semelhantes aos termos existentes na sentença de busca, tentando-se manter a semântica da sentença.

Outros trabalhos que poderão ser desenvolvidos consistem na alimentação da base de conhecimento por termos, conceitos, um_grama e descrições textuais referentes a exames de Ressonância Magnética em Joelho, assim como para outros tipos de exames.

Aconselha-se também que futuros trabalhos relacionados ao desenvolvimento de sistemas de auxílio ao diagnóstico, utilizando técnicas de Aprendizado de Máquina e/ou Inteligência Artificial, sejam realizados a partir dos laudos estruturados por este trabalho, utilizando a indicação dos termos existentes como atributos destas técnicas.

6.3. Conclusão

Nos sistemas informatizados voltados à geração de laudos em radiologia, predomina-se a aquisição e o armazenamento da informação na forma de anotação textual aberta. Um grande desafio enfrentado no desenvolvimento destes sistemas reside na inadequabilidade de ambientes caracterizados por interfaces de preenchimento padronizados e pré-definidos, que restringem fortemente a liberdade do médico na geração de seus relatos; contrastando com a ineficácia de outros ambientes que trabalham com textos abertos, restringindo fortemente as possibilidades de análise futura da informação.

A ferramenta *E-Rad* propõe minimizar a ineficácia de ambientes que utilizam dados do tipo texto aberto. Pois, ao ser desenvolvida baseada em computação não intrusiva, essa ferramenta mantém a liberdade dos usuários, preservando a forma habitual de trabalho e não

restringindo as tarefas de laudagem de exames a partir do preenchimento padronizado e pré-definido de campos. Ao mesmo tempo em que, de forma invisível aos usuários, essa ferramenta processa os dados informados e os estruturam. Permite-se que, dessa forma, sistemas de busca e sistemas inteligentes sejam desenvolvidos a partir desses dados estruturados.

Logo, esta ferramenta torna-se necessária e aconselhável, por viabilizar que técnicas baseadas em dados estruturados sejam utilizadas pelos usuários sem que sejam obrigados a se adaptarem às condições de fornecimento de dados, e sim que o sistema se adapte às condições de trabalho do usuário.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Caritá, E. C., Matos, A. L. M., Azevedo-Marques, P. M. Ferramentas para Visualização de Imagens Médicas em Hospital Universitário. *Radiol Brás* 2004; 37(6):437-440
- [2] Blois, M.S., Shortliffe, E.H. (1990), “The Computer Meets Medicine: Emergence of a Discipline”. In: Shortliffe, E.H., Perreault, L.E. (eds). *Medical Informatics: Computer Applications in Health Care*. New York: Addison-Wesley Publishing, 1990. p.3-36.
- [3] Costa, C.G.A. (2001), “Desenvolvimento e Avaliação Tecnológica de um Sistema de Prontuário Eletrônico do Paciente, Baseado nos Paradigmas da World Wide Web e da Engenharia de Software”. Faculdade de Engenharia Elétrica e de Computação, Unicamp, Campinas, SP, uma opinião”. *Revista Medicina Interna*; vol 10, N.4.
- [4] Kock Jr.; N.F.; McQueen, R. J.; Corner, J. L. *The Nature of Data, Information and Knowledge Organizations: A Critical Analysis of Four Contemporary myths*. The Learning Organization. 1996, p. 31-40.
- [5] Aldershot: Gower. *Harrod’s Librarian’s Glossary of Term Used in Librarianship, Documentation and the Book Craft and Reference Book*. 6th edition. 1989, p281
- [6] *Dictionnaire encyclopédique de l’information et de la documentation*. 2^{ème} edition. Paris : Nathan. 2001, p297
- [7] Dodebei, V. L. D. *Tesauro: Linguagem de representação da memória documentária*. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002.
- [8] Sebastiani, F.; *Machine learning in automated text categorization*. *ACM Computing Surveys*, March - 2002.
- [9] Feldman, R.; Dagan, I. *Knowledge Discovery in Textual Database (KDT)*. In *First International Conference on Knowledge Discovery (KDD’95)*, Montreal, August, 1995.
- [10] Lewis, D. D., Jones, K. S. *Natural Language Processing For Information Retrieval*. *Communication of the ACM*, 39(1):92-101, 1996
- [11] Kamber, M.; Han, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

- [12] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.. From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press, 1996, 611 p.p.11-34.
- [13] Silva, C. F., Vieira, R., Osório, F. S., *Uso de Informação Lingüísticas na etapa de pré-processamento em Mineração de Texto*, Universidade do vale do Rio Dos Sinos Ciências Exatas e Tecnológicas, Programa interdisciplinar de Pós-Graduação em Computação Aplicada – PPIPCA, Fevereiro 2004.
- [14] Dixon, M. *Na Overview of Document Mining Technology*. 1997.
- [15] Corrêa, A. C. G., Vieira, M. T. P., Santos, M. T. P., *Recuperação de Documentos baseada em Informação Semântica no Ambiente AMMO*, Universidade Federal de São Carlos, Centro de Ciências Exatas e de Tecnologia, Programa de Pós-Graduação em Ciência da Computação, Agosto 2003.
- [16] Allen, J., “*Natural Language Understanding (2nd ed.)*”. The Benjamin/Cummings Publishing Company. 1995. p.20.
- [17] Martin, J. H. *Speech and Language Processing*, prentice-hall. 2000, p 1-18
- [18] Lucena, P., *Dissertação de Mestrado*. Instituto de Ciências Matemáticas e de Computação – ICMC-USP, Título: *SemanticAgent*, uma plataforma para desenvolvimento de agentes inteligentes, 2003
- [19] Dias, M. A. L., Malheiros, M. G.; *Extração Automática de Palavras-chave de Textos da Língua Portuguesa*. Centro Universitário UNIVATES. 2005.
- [20] Orengo, V.M., Huyck, C. R., *A Stemminh Algorithm for The Portuguese Language*. Proceedings of the SPIRE Conference. 2001
- [21] Porter, M., *An algorithm for suffix stripping*. *Program* 14(3), 130-137, 1980.
- [22] Matsubara, E. T., Martins, C. A., Monard, M. C., *PreTextT: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words*, Instituto de Ciências Matemáticas e de Computação, USP-SC, Agosto 2003
- [23] Martins, C. A., *Dissertação de Doutorado*. Instituto de Ciências Matemáticas e de Computação – ICMC. Universidade de São Paulo, São Carlos, Brasil. Título: *Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado*, Ano de Obtenção: 2003.

- [24] Luhn, H. P., The automatic creation of literature abstracts. IBM Journal of Research and Development, 1958
- [25] Soares, M. V. B.; Prati, R. C.; Monard, M. C., PreTexT: A Reestruturação da Ferramenta de Pré-processamento de Textos. Instituto de Ciências Matemáticas e de Computação, USP-SC, Agosto, 2008
- [26] Gruber, T.R., A translation approach to portable ontology specifications. Knowledge Acquisition, 1993. 5: p. 199-220.
- [27] Novello, T. C., Ontologias, Sistemas Baseados em Conhecimento e Modelo de Banco de Dados. Universidade Federal do Rio Grande do Sul.
- [28] Moore, K. L.; Dalley A. F.; Anatomia Orientada para a Clínica, 1994, 4a edição.

APÊNDICES

Apêndice 1 – Arquivo de Configuração da ferramenta *PreText*

```
<?xml version="1.0" encoding="utf-8"?>
<PreText
  lang="pt"
  dir="joelho"
  silence="on">

  <maid>
    <html/>
    <simbols/>
    <stoplist dir="stoplist">
      <stopfile>port.xml</stopfile>
    </stoplist>
    <stemming/>
  </maid>

</PreText>
```

Tabela 8: Arquivo de configuração da ferramenta *PreText*

ANEXOS

Anexo A – Conjunto de palavras que foram removidas do arquivo *Stoplist*

<stopword>abaixo</stopword>	<stopword>menos</stopword>
<stopword>acima</stopword>	<stopword>muita</stopword>
<stopword>antes</stopword>	<stopword>muitas</stopword>
<stopword>apos</stopword>	<stopword>muitissimo</stopword>
<stopword>bastantes</stopword>	<stopword>muito</stopword>
<stopword>bem</stopword>	<stopword>muitos</stopword>
<stopword>bom</stopword>	<stopword>nao</stopword>
<stopword>contudo</stopword>	<stopword>nem</stopword>
<stopword>depois</stopword>	<stopword>nenhum</stopword>
<stopword>durante</stopword>	<stopword>nenhuma</stopword>
<stopword>embaixo</stopword>	<stopword>porem</stopword>
<stopword>entre</stopword>	<stopword>pouca</stopword>
<stopword>entretanto</stopword>	<stopword>poucas</stopword>
<stopword>exceto</stopword>	<stopword>pouco</stopword>
<stopword>exceto</stopword>	<stopword>poucos</stopword>
<stopword>mais</stopword>	<stopword>praticamente</stopword>
<stopword>mal</stopword>	<stopword>sem</stopword>
<stopword>mas</stopword>	<stopword>tambem</stopword>

Tabela 9: Palavras desconsideradas *stopwords*