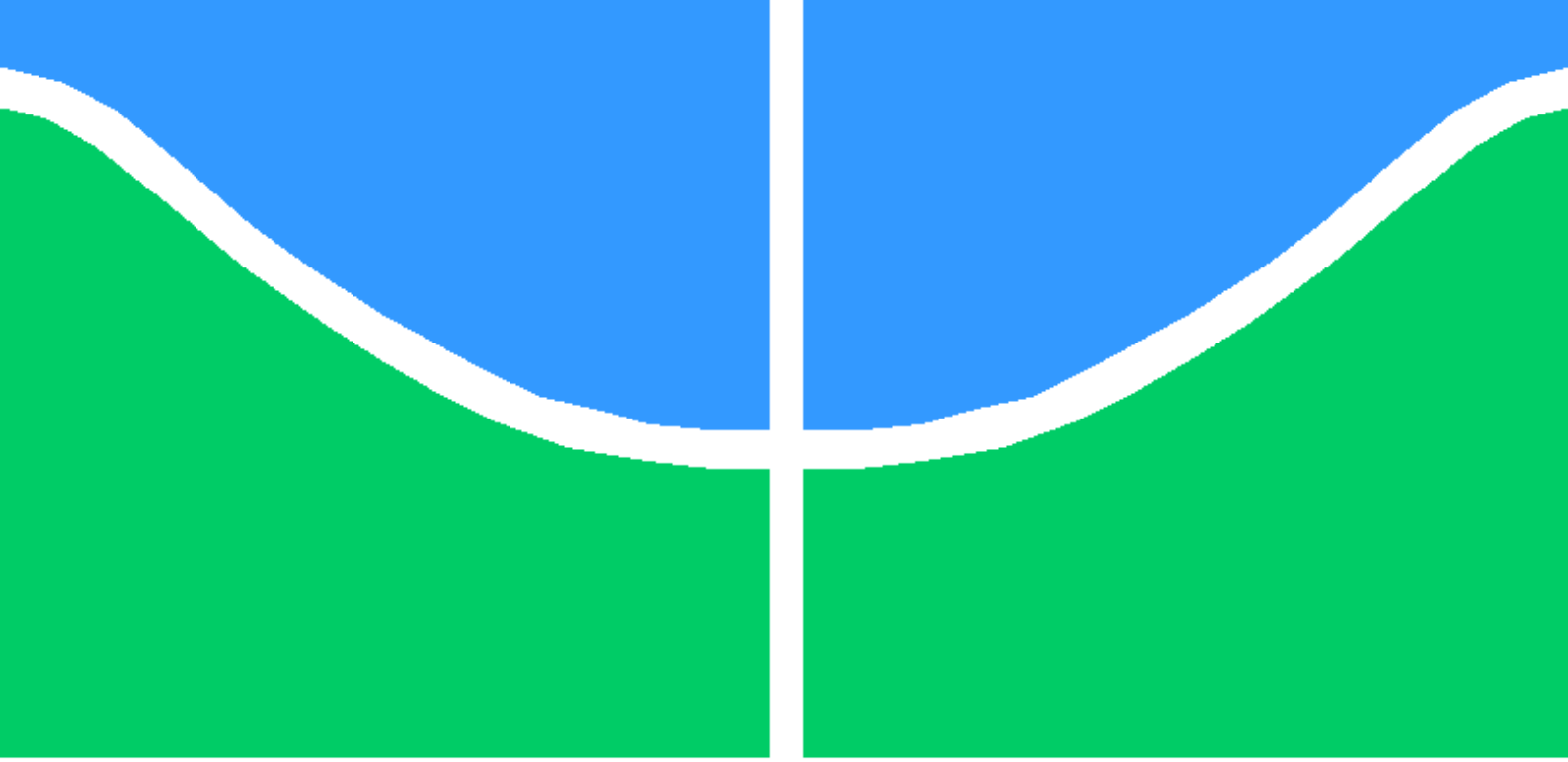


CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO

VANDERLEI JANDIR ALLES

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO

VANDERLEI JANDIR ALLES

Orientador: PROF. DR. WILLIAM FERREIRA GIOZZA , ENE/UNB
Co-Orientador: PROF. DR. ROBSON DE OLIVEIRA ALBUQUERQUE , ENE/UNB

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO PPGENE.DM - 714/2018
BRASÍLIA-DF, 17 DE DEZEMBRO DE 2018.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO

VANDERLEI JANDIR ALLES

DISSERTAÇÃO DE MESTRADO ACADÊMICO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

APROVADA POR:

Prof. Dr. William Ferreira Giozza , ENE/UnB
Orientador

Prof. Dr. Rafael Timóteo de Sousa, ENE/UnB
Examinador interno

Prof. Dr. Rafael Rabelo Nunes, ENE/UnB
Examinador externo

BRASÍLIA, 17 DE DEZEMBRO DE 2018.

FICHA CATALOGRÁFICA
VANDERLEI JANDIR ALLES
CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO
2018xv, 73p., 201x297 mm
(ENE/FT/UnB, Mestre, Engenharia Elétrica, 2018)
Dissertação de Mestrado - Universidade de Brasília
Faculdade de Tecnologia - Departamento de Engenharia Elétrica

REFERÊNCIA BIBLIOGRÁFICA

VANDERLEI JANDIR ALLES (2018) CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO. Dissertação de Mestrado em Engenharia Elétrica, Publicação 714/2018, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 73p.

CESSÃO DE DIREITOS

AUTOR: Vanderlei Jandir Alles
TÍTULO: CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO.
GRAU: Mestre ANO: 2018

Vanderlei Jandir Alles
Qd Central Cj B Bloco A, Apto 203- Sobradinho-DF.

Agradecimentos

A dissertação não poderia chegar ao final sem o apoio de algumas pessoas que fizeram a diferença e devem ser mencionadas.

Em primeiro lugar ao meu orientador Professor Doutor William Ferreira Giozza, pelo acompanhamento e paciência nas correções e dicas que foram cruciais para o desenvolvimento deste trabalho.

Agradecer ao co-orientador, que também chamo de Professor, Doutor Robson de Oliveira Albuquerque, com sua dedicação com conselhos de ajuda em algumas horas difíceis que foram cruciais para o desenvolvimento deste trabalho.

Gostaria de agradecer também a minha querida esposa, Reni, pelo apoio energias positiva que não me deixou desistir jamais nesta caminhada.

E aos meus colegas e amigos do trabalho da DPU, que colaboram com suas energias positivas e em especial ao amigo Amoz Felipe pelas contribuições deste trabalho.

Durante o desenvolvimento de seu mestrado, o autor foi bolsista de projeto entre à Defensoria Pública da União e a Universidade de Brasília (TED 066/2016 DPU-FUB), agradecendo assim às instituições pelo suporte ao presente trabalho.

Resumo

CONSTRUÇÃO DE UM *CORPUS* PARA EXTRAIR ENTIDADES NOMEADAS DO DIÁRIO OFICIAL DA UNIÃO UTILIZANDO APRENDIZADO SUPERVISIONADO

Autor: Vanderlei Jandir Alles
Orientador: Dr. William Ferreira Giozza
Programa de Pós-graduação em Engenharia Elétrica
Brasília, 17 de Dezembro de 2018

O entendimento da estrutura gramatical de uma frase é um passo importante para que os computadores sejam capazes de compreender o significado pretendido em um texto.

Esta dissertação faz um estudo de quatro ferramentas que realizam processamento de linguagem natural. O trabalho explora conceitos envolvendo ferramentas que realizam PLN e utiliza uma metodologia de construção de um *corpus* específico que auxilie o reconhecimento de entidades da fonte de dados textual (DOU), processando o entendimento linguístico das palavras em um texto e depois comparando a quantidade e qualidade das entidades que foram reconhecidas nos textos processados.

Assim, a OpenNLP foi escolhida e construiu-se um novo *corpus*, utilizando o aprendizado supervisionado, para que fosse elaborada uma proposta de construção de um *corpus* específico para extrair Entidades Nomeadas com melhor qualidade em comparação com os resultados obtidos com os *corpus* disponíveis.

Uma arquitetura foi desenvolvida para compreender um conjunto de atividades a serem executadas na extração de Entidades Nomeadas, identificando e descrevendo a organização dos módulos, visando a codificação e especificação de cada um deles.

Palavras chaves: Aprendizado de Máquina, *Corpus*, Diário Oficial da União do Brasil, Entidades Nomeadas, Processamento de Linguagem Natural.

Abstract

CONSTRUCTION OF A CORPUS TO EXTRACT NAMED ENTITIES OF THE UNION OFFICIAL DIARY USING SUPERVISED LEARNING

Author: Vanderlei Jandir Alles
Supervisor: Dr. William Fereira Giozza
Post-Graduation Program on Eletrical Engineering
Brasília, December 17th, 2018

Understanding the grammatical structure of a sentence is an important step for computers to be able to understand the meaning intended in a text.

This dissertation makes a study of four tools that perform natural language processing. The work explores concepts involving tools that perform PLN and uses a methodology of construction of a specific corpus that helps the recognition of entities of the textual data source (DOU), processing the linguistic understanding of the words in a text and then comparing the quantity and quality of the entities that were recognized in the texts processed.

Thus, OpenNLP was chosen and a new corpus was constructed, using supervised learning, to elaborate a proposal to build a specific corpus to extract named entities with better quality in comparison to the results obtained with the available corpus.

An architecture was developed to understand a set of activities to be performed in the extraction of named entities, identifying and describing the organization of the modules, aiming at the coding and specification of each one of them.

Keywords: Machine Learning, Corpus, Union Official Diary of the Brazil, Named Entities, Natural Language Processing.

SUMÁRIO

Lista de Termos e Siglas

CRF	Conditional Random Fields
DARPA	*Defense Advanced Research Projects Agency*
DOU	Diário Oficial da União
EN	Entidades Nomeadas
ENME	Entidade Nomeada com Máxima Entropia
IA	Inteligência Artificial
IE	Extração da Informação
JSON	*JavaScript Object Notation*
ME	Máxima Entropia
MUC	*Message Understanding Conference*
NLTK	*The Natural Language Toolkit*
PDF	*Portable Document Format*
PLN	Processamento de Linguagem Natural
REN	Reconhecimento de Entidades Nomeadas
SN	Sintagmas Nominais
SQL	*Structured Query Language*
UUID	*Universally unique identifier*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os últimos anos foram marcados por profundas mudanças no campo das ciências e tecnologia ocasionando, entre outras mudanças, a popularização de equipamentos e dispositivos tecnológicos que promoveram a difusão da web em praticamente todas as regiões do planeta. As informações produzidas a todo instante são conhecidas e compartilhadas na mesma velocidade, assim de qualquer lugar, estando conectado à rede, é possível não só ter acesso mas também produzir informações. Entretanto, por outro lado, o grande volume de informação disponível tende a dificultar a extração de informação que seja relevante [Souza and Claro 2014].

Mais especificamente, neste trabalho, é estudado o comportamento das ferramentas de extração de informação em termos do processo de aprendizagem de máquina para o processamento de linguagem natural. O trabalho explora conceitos envolvendo ferramentas que realizam PLN e utiliza uma metodologia de construção de um *corpus* específico que auxilie o reconhecimento de entidades da fonte de dados textual (DOU), processando o entendimento linguístico das palavras em um texto e depois comparando a quantidade e qualidade das entidades que foram reconhecidas nos textos processados.

Aqui entendemos como informação relevante aquela atribuída de significado e veracidade, que seja possível a sua utilização e análise em determinado contexto contribuindo para a construção de conhecimentos. A busca automatizada por essas informações ainda é limitada mesmo quando dispomos de métodos específicos disponibilizados pelas principais ferramentas de extração de informação. A complexidade advinda de uma estrutura textual formada por um léxico (conjunto de palavras usadas numa língua, ou num texto ou por um autor) e a gramática que trata das regras da linguagem constituem desafios importantes a serem superados.

O entendimento da estrutura gramatical de uma frase é um passo importante para que os computadores sejam capazes de compreender o significado pretendido em um texto. O Processamento de Linguagem Natural (PLN), subárea da Inteligência Artificial, é um campo de estudo de automatização computacional, entendimento e organização gramatical de uma

linguagem não estruturada em aplicações como, por exemplo, tradução automática, processamento e sintetização de textos em linguagem natural, reconhecimento de fala, sistemas especialistas e extração de significado, entre outras [Finatto et al. 2015].

Nesse sentido o Processamento de Linguagem Natural (PLN) é visto como o estudo de algoritmos e métodos para construção de modelos computacionais capazes de analisar e executar textos expressos na língua natural em diversos idiomas. Desta maneira, o PLN permite interpretar e processar a informação por meio de linguagem natural [Manning et al. 2014]. A construção de dicionários e gramáticas, é fundamental na elaboração de ferramentas de extração de informação que apresentem qualidade em termos de extração de conteúdo significativo.

Tendo os princípios discutidos acima levados em consideração, este trabalho apresenta inicialmente um estudo comparativo entre quatro ferramentas de extração de informação: *OpenNLP* [OpenNLP 2017], *CoreNLP* [Manning et al. 2014], *Syntaxnet* [Alberti et al. 2017] e *NLTK* [Bird and Loper 2004], utilizadas para extrair Entidades Nomeadas do Diário Oficial da União (DOU), utilizando-se de dicionários previamente disponíveis para cada ferramenta. Como resultado desse estudo preliminar, identificou-se a importância dos dicionários para o bom desempenho de cada ferramenta de extração e concluiu-se pela necessidade de elaboração de um dicionário próprio, melhor adaptado à aplicação pretendida e capaz de prover uma extração de Entidades Nomeadas (EN) com um desempenho superior aos conseguidos com os dicionários previamente disponibilizados pelas ferramentas de extração.

Este trabalho tem como foco a realização de extração de Entidades Nomeadas do Diário Oficial da União (DOU). A ferramenta de extração escolhida foi a *OpenNLP* porque ela fornece maneiras de construir os modelos de forma mais simples, e possuir documentações, com exemplos e procedimentos que auxiliam na construção do modelo. Entretanto, para obter um desempenho melhor do que o obtido com o dicionário previamente disponível com a *OpenNLP* foi necessário elaborar um *corpus* que são conjuntos de dados textuais legíveis e compilados com o propósito de servirem como fonte para diferentes tarefas de processamento linguagem natural [Chiele et al. 2015] específico por meio da coleta prévia de trechos do DOU, até a fase de treinos e submissão do mesmo à ferramenta.

1.1 Motivação

O Diário Oficial da União (DOU) contém várias informações que descontextualizadas perdem o sentido e talvez não sejam de interesse dos órgãos públicos nem dos servidores que o utilizam. Então, promover uma busca de informação mais objetiva ao que seria relevante no DOU é um problema em que o processamento de linguagem natural poderia contribuir oferecendo soluções eficazes.

Do ponto de vista computacional, verifica-se a importância de se reconhecer as unida-

des de informação do DOU por meio do uso de PLN o que facilitaria a busca por dados e informações de interesse. O uso de PLN no DOU poderia facilitar bastante encontrar, por exemplo, quais foram as pessoas exoneradas em 2017 em um determinado órgão e quem foi nomeado na sequência.

Contudo, para que o processamento da informação seja automatizado precisa-se preparar a informação para ser extraída do DOU. É importante por exemplo prover técnicas para ajudar a ferramenta de extração na busca da informação correta e este é o primeiro passo deste trabalho: estudar as técnicas de extração de entidades nomeadas, termo comumente conhecido por Reconhecimento de Entidades Nomeadas.

1.2 Objetivos

Este trabalho, tem por objetivo explorar os conceitos envolvendo as ferramentas que realizam PLN, construir um *corpus* específico que permita realizar de maneira otimizada o reconhecimento de entidades em textos do DOU e desenvolver uma aplicação que apresente os resultados da extração de informação.

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é propor e avaliar métodos de extração de Entidades Nomeadas utilizando o DOU como *dataset* de informação de extração.

1.2.2 Objetivos Específicos

No que se refere aos objetivos específicos deste trabalho, tem-se:

- Estudo e teste de ferramentas de Processamento de Linguagem Natural;
- Construção de um corpus específico (*DOU-Corpus*);
- Desenvolvimento da *engine* que treine e extraia Entidades Nomeadas do DOU utilizando a *OpenNLP*;
- Desenvolvimento de uma ferramenta de extração de Entidades Nomeadas do DOU;
- Avaliação da capacidade do método de extração proposto em termos de resultados qualitativos e quantitativos.

1.3 Metodologia de Desenvolvimento

O trabalho inicia-se com um estudo das principais ferramentas de extração de informação existentes. Escolhida a ferramenta de extração mais adaptada aos objetivos do trabalho, passa-se à etapa de a construção de um *corpus* específico (*DOU-Corpus*) para extrair a Entidades Nomeadas do DOU. Por fim, para auxiliar na verificação e avaliação do método de extração proposto foi implementada uma aplicação para apresentação dos resultados das extrações.

1.4 Estrutura do Trabalho

Este trabalho está estruturado da seguinte forma:

- O Capítulo 2 apresenta uma revisão bibliográfica e o estado da arte sobre aprendizagem de máquina e o processamento de linguagem natural e suas técnicas.
- O Capítulo 3 descreve a extração de informações do DOU e propõe uma solução baseada em aprendizado de máquina.
- O Capítulo 4 apresenta e discute a implementação da solução proposta para extração de informações do DOU, a sua prototipagem e os resultados dos experimentos de extração realizados.
- O Capítulo 5 apresenta as conclusões deste trabalho, as contribuições e possíveis trabalhos futuros.

Capítulo 2

Estado da Arte e Trabalhos Relacionados

Neste capítulo é apresentada uma breve análise sobre Aprendizado de Máquina e seus principais conceitos relacionados ao Processamento de Linguagem Natural (PLN). Para melhor compreensão do trabalho proposto, também serão abordadas as técnicas de PLN para a construção de um *corpus*.

Em seguida, são apresentados os principais trabalhos de pesquisa abordando conceitos que sejam pertinentes ao tema deste trabalho. Este capítulo descreve também diversos conceitos relevantes a respeito do processamento da informação procurando, sempre que possível, estabelecer uma relação entre os temas estudados.

Ao final do capítulo é apresentada uma análise das ferramentas de extração que fazem uso de PLN.

2.1 Inteligência Artificial

Uma das principais características da inteligencia e aprendizagem é a capacidade de se comunicar, compreender e aprender a resolver problemas e conflitos e de adaptar-se a novas situações. A Inteligência Artificial (IA) é uma área de conhecimento computacional que se propõe a automatizar um comportamento inteligente das máquinas, visando desenvolver mecanismos e dispositivos tecnológicos que possam simular o raciocínio humano em termos de capacidade na adquisição, representação e manipulação do conhecimento [Lustosa 2010]. Segundo o estudioso John McCarthy [Sellitto 2002], a IA é "a ciência e engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes". Mas isso constitui um desafio que requer habilidades em termos de planejamento, compreensão da linguagem, reconhecimento de objetos e sons, aprendizado e resolução de problemas.

Nos últimos anos, o volume do conhecimento produzido tem sido enorme, formatado ou e desorganizado, e continua mudando constantemente. Um dos objetivos da IA é identificar e resolver problemas de processamento de informações úteis. É um campo do conhecimento que oferece modelos de apoio à decisão e ao controle, com base em fatos reais e conhecimen-

tos empíricos e teóricos, permitindo a organização, a solução de problemas e a compreensão de ideias complexas, usando a linguagem natural, de maneira adaptativa a novas situações.

Observa-se que não é uma tarefa trivial tornar o conhecimento informal em termos formais exigidos pela notação lógica. Mesmo assim, a IA vem ganhando significativa importância na automação e organização do comportamento inteligente.

Pode-se dizer que a IA, com a utilização de processos computacionais, propõe a elaboração de dispositivos que simulem a capacidade de raciocinar, tomar decisões através de experiências acumuladas. Um dos requisitos básicos para qualquer comportamento inteligente é aprender e não há inteligência sem aprendizado. Assim, o objetivo da IA é desenvolver sistemas computacionais que possam ser capazes de identificar objetos do cotidiano que tenham alguma orientação, na tentativa de construir sistemas de raciocínio computacional [Russell and Norvig 2016]. Dentre as tecnologias, que precisam ser programadas e integradas aos sistemas inteligentes, tem-se:

- *aprendizado de máquina* para permitir o computador adaptar-se a novas circunstâncias e detectar e extrapolar padrões, baseado no treinamento de dados e análise estatística para determinar padrões e previsões;
- *representação do conhecimento* para colocar o conhecimento em uma forma que um computador possa "raciocinar";
- *aprendizagem profunda* para um tipo particular de aprendizagem de máquina que emprega algoritmos que utilizam redes neurais com múltiplas camadas de abstração, simulando o processamento do cérebro humano.

A Figura 2.1 ilustra a estrutura das subáreas componentes da área de Inteligência Artificial.

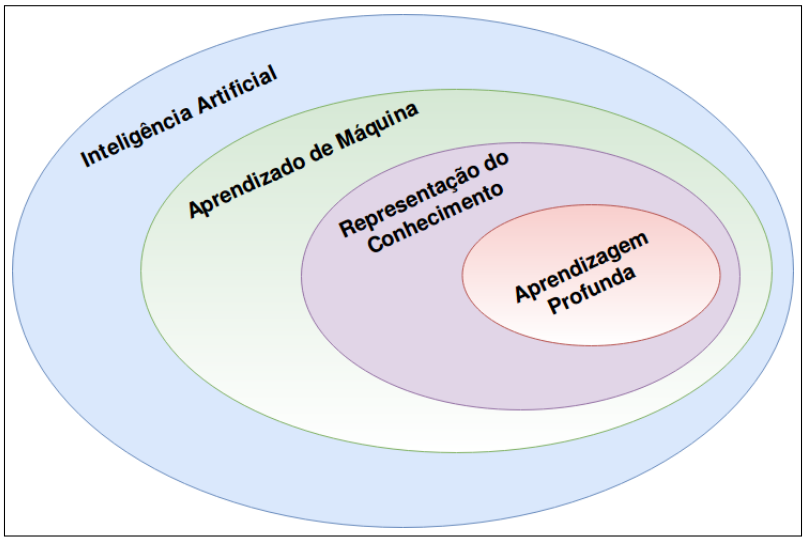


Figura 2.1: Estrutura das subáreas da Inteligência Artificial

Portanto, as questões principais a serem abordadas pelo projetista de um sistema de IA são: aquisição, representação e manipulação de conhecimento e, geralmente, adoção de uma estratégia de controle ou máquina de inferência que determina os itens de conhecimento a serem acedidos, as deduções a serem feitas, e a ordem dos passos a serem usados. A Figura 2.2 representa estas questões, mostrando a inter-relação entre os componentes de um sistema clássico de IA [VICCARI 1990].

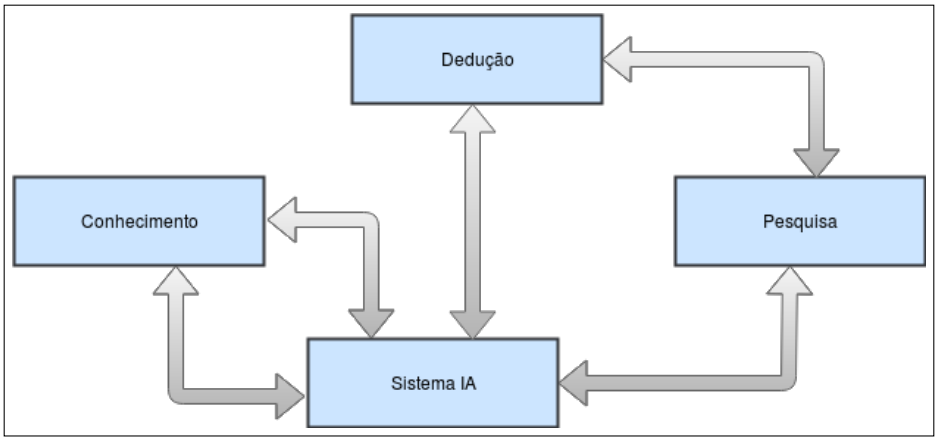


Figura 2.2: Inter-relação entre componentes de um sistema clássico de IA

2.2 Aprendizado de Máquina

O Aprendizado de Máquina utiliza diferentes conjunto de técnicas e ferramentas, para detectar padrões e descobrir soluções de problemas complexos. A característica mais importante do comportamento inteligente é o aprendizado. Assim, as pesquisas e experimentos em Aprendizado de Máquina tem sido uma busca constante na solução de problemas de domínios do mundo real. E a medida que aplicações bem sucedidas surgem, aumenta a aceitação de seu uso. É possível realizar previsões baseadas em um paradigma que pode se referir ao aprendizado de experiências passadas e utilizadas para treinar o algoritmo, por meio da aquisição e otimização de dados e construção de modelo. Desta maneira, se viabiliza que computadores processem tarefas específicas e de complexidade elevada.

Porém, o aprendizado de máquina, por si só, não pode ser visto como uma solução para os problemas relacionados no campo da IA, haja vista que o mesmo implica em intervenções humanas e intuição para desenvolver os algoritmos.

No aprendizado de máquina é dada ênfase em técnicas que empregam um princípio denominado como indutivo, que consiste em uma conclusão genérica a partir de um conjunto particular de treinos. Os processos externos de aprendizado podem ser classificados nos seguintes três importantes subgrupos de algoritmos indutivos [Jordan and Mitchell 2015].

- *Supervisionado*: para esse tipo de algoritmo é dado um conjunto de dados, e previamente rotulados e já conhecido. O algoritmo é ensinado por meio de um conjunto de

dados de entradas e respostas previamente conhecidas. É o modelo mais comumente utilizado e portanto, adotado neste trabalho.

- *Semi-supervisionado*: neste tipo de aprendizado, é apenas um pequeno número de exemplos rotulados encontra-se disponível e passam a ser considerados como supervisionados com a diferença que o algoritmo de forma inteligente decide uma rotulagem nova dos dados, e que desta forma consiga aprender com dados anotados e não anotados, uma vez que dados não rotulados existem em abundância e exemplos rotulados são geralmente escassos.

- *Não-supervisionado*: neste tipo de aprendizado o algoritmo é responsável por tentar descobrir padrões e estruturas intrínsecas, por exemplo, extraindo inferências em um conjunto de dados que possui respostas não rotuladas ou conhecidas. O objetivo neste caso é descobrir atividades úteis e desejadas por meio de tentativa-e-erro e processos auto-organizáveis, não utilizando informações das variáveis de saída. Este tipo de aprendizado ainda não possui uma escala de implementação tão ampla e difundida quanto à do aprendizado supervisionado.

A Figura 2.3 ilustra os tipos de aprendizado de máquina baseados no aprendizado indutivo.

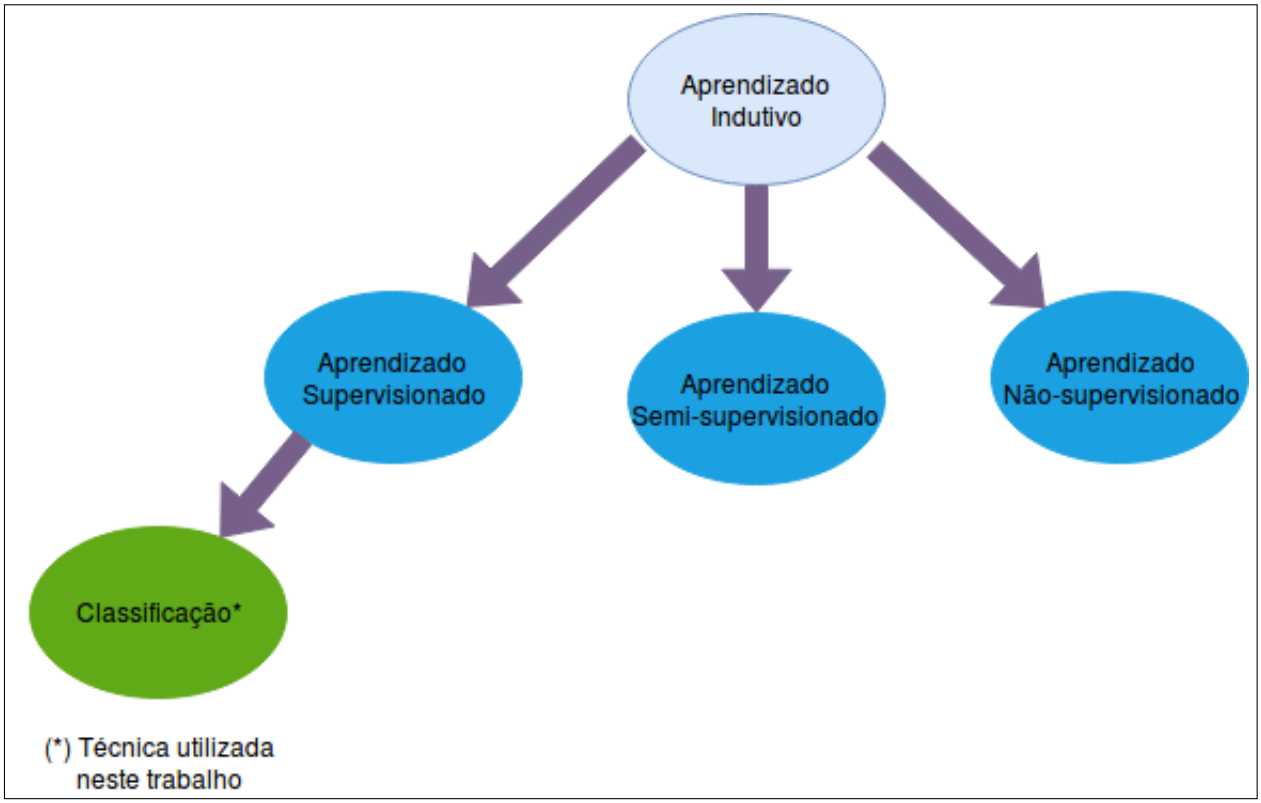


Figura 2.3: Tipos de aprendizado de máquina baseados no aprendizado intuitivo (adaptado de [Monard and Baranauskas 2003])

Neste trabalho será utilizado o aprendizado supervisionado com o intuito de classificar

entidades tais como: pessoas, lugares e a própria organização do DOU que é descrito com mais riqueza os detalhes na Seção 3.1.

Diversos trabalhos mencionam técnicas e práticas do uso do aprendizado de máquina nos mais variados setores para tomada de decisões na criação de sistemas inteligentes.

Em [Kotsiantis et al. 2007] os autores, descrevem algumas técnicas de aprendizado supervisionado de classificação e concluem que a escolha do algoritmo que deve ser usado no aprendizado é um passo crítico. Eles salientam que não necessariamente um determinado método possa superar outro, mas que sob condições diferenciadas um determinado método pode apresentar um desempenho significativamente superior na solução de um determinado problema.

Mostram também que o objetivo do aprendizado supervisionado é construir um modelo preciso para a tarefa de rotular entidades e que o classificador resultante possa ser usado a partir do conjunto de regras pré-definidas. O processo de aprendizado adotado em [Kotsiantis et al. 2007] é descrito na Figura 2.4.

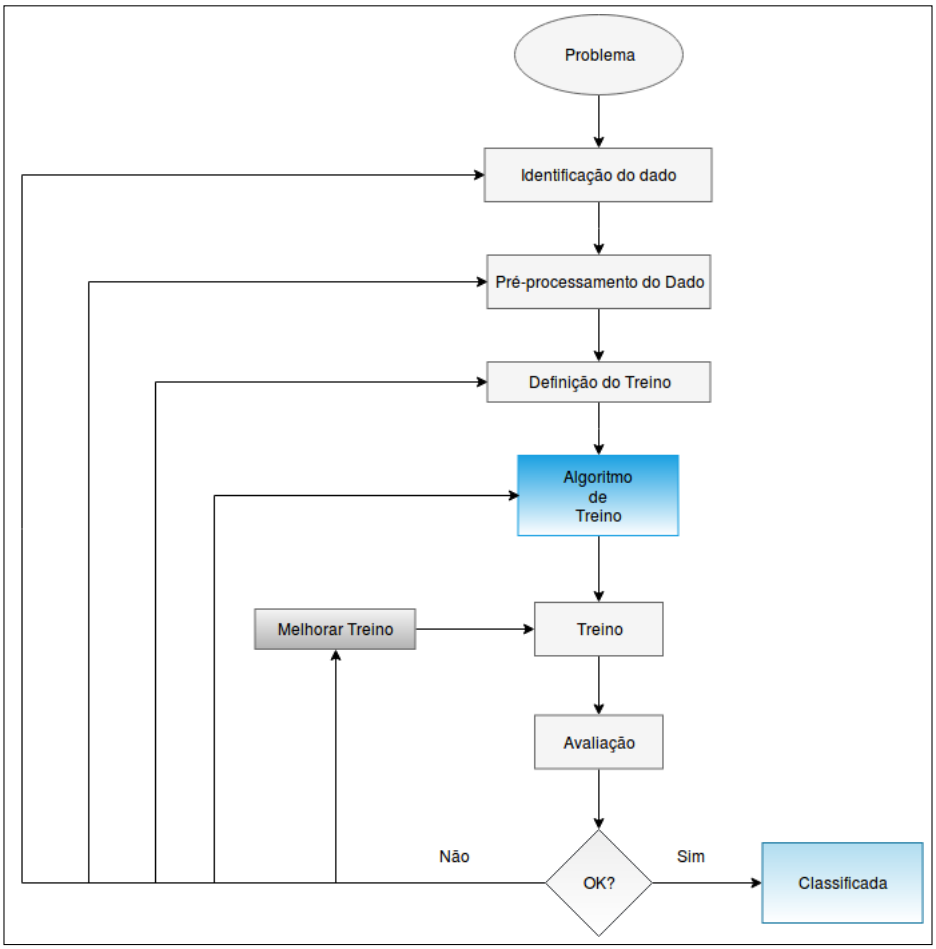


Figura 2.4: O processo do Aprendizado Supervisionado (adaptado de [Kotsiantis et al. 2007]

Segundo Kotsiantis e colegas, a solução de resolver um determinado problema utilizando aprendizado supervisionado necessita que se execute os seguintes três passos iniciais:

- O primeiro passo é identificar o conjunto de dados que será usado como conjunto de treino.
- O segundo passo é preparar e pre-processar o conjunto de dados identificando e removendo dados que não sejam relevantes para o processo de treino.
- O terceiro passo é a escolha de um algoritmo de treino

Os autores em [Jordan and Mitchell 2015] enfatizam que o aprendizado é um processo complexo que passa por muitas decisões. Explora a contextualização de três diferentes paradigmas de aprendizado de máquina, supervisionado, não-supervisionado e por esforço. Destacam que o aprendizado semi-supervisionado pode ser usado mesclando o aprendizado supervisionado com o não-supervisionado, permitindo completar a tarefa de forma mais otimizada possível. Já o aprendizado por reforço usa um algoritmo que em um ambiente dinâmico, tem foco na criação de agentes capazes de tomar decisões para um determinado objetivo sem especificar como a tarefa é realizada [Ottoni et al. 2016]. As informações de treinamento são intermediárias entre o aprendizado supervisionado e o não supervisionado. Esse trabalho apresenta os mecanismos que fundamentam cientificamente o aprendizado de máquina e destaca que os algoritmos de aprendizado dependem das suas estruturas de dados nas teorias de aprendizagens para entender um problema em particular.

O trabalho divulgado em [Das and Behera 2017] se concentra em explicar o conceito e a evolução do Aprendizado de Máquina. Explorando alguns dos algoritmos mais populares, esse trabalho discute como a aprendizagem refere-se à modificação ou melhoria do algoritmo com base em experiências passadas automaticamente sem qualquer assistência externa de humano.

Em [Qiu et al. 2016] os autores abordam uma perspectiva interessante sobre a utilização de técnicas de Aprendizado de Máquina em grandes volumes de dados destacando os desafios para o processamento em diferentes contextos na solução do problema do *big data*. Em particular, os autores destacam que as técnicas de aprendizado supervisionado e de aprendizado não-supervisionado focam em análise de dado enquanto que o aprendizado por reforço é direcionado para os problemas de tomada de decisão.

2.3 Representação do Conhecimento

Nos últimos anos com a crescente expansão dos meios eletrônicos, a quantidade de conhecimento tem sido ampliada de forma exponencial e de maneira contínua, dificultando ainda mais a tarefa de recuperação de informações relevantes.

Nesse contexto de grandes volumes de informação é preciso disponibilizar ferramentas automatizadas para estruturar o conhecimento. Nesse sentido, torna-se necessária a busca de formas mais eficazes de representar o conhecimento, visando sua posterior recuperação, estabelecendo critérios de seleção, de organização e de representação da informação, e criando

estruturas de representação mais adequadas para alcançar maior sucesso na recuperação da informação.

Os autores, em [Russell and Norvig 2016] descrevem a representação do conhecimento por meio de uma proposta que introduz a ideia de uma ontologia geral, que organiza tudo em uma hierarquia de categorias. Ao aprofundar os estudos sobre os detalhes de como uma pessoa representa uma variedade de conhecimento, os autores pretendem dar ao leitor uma noção de como as bases reais de conhecimento são construídas e um sentimento para as questões filosóficas que surgem. Os principais pontos destacados nesse trabalho são os seguintes:

- A representação do conhecimento em grande escala requer uma ontologia de propósito geral para organizar e vincular os vários domínios específicos do conhecimento.
- Uma ontologia de propósito geral precisa cobrir uma ampla variedade de conhecimentos e deve ser capaz, em princípio, de lidar com qualquer domínio.
- Construir uma grande ontologia de propósito geral é um desafio significativo que ainda precisa ser totalmente realizado, embora as estruturas atuais pareçam ser bastante robustas.
- Os autores por fim propõem uma ontologia superior baseada em categorias e no cálculo de eventos, incluindo categorias, subcategorias, partes, objetos estruturados, medidas, substâncias, eventos, tempo e espaço, mudança e crenças.

Resumidamente, a representação do conhecimento pode ser realizada por meio da lógica de primeira ordem, que pressupõe que o mundo consiste em objetos com certas relações entre eles que são válidas ou não-válidas [de Souza 2008].

2.4 Aprendizagem Profunda

A Aprendizagem Profunda, conhecida em Inglês como *Deep Learning* é uma das muitas abordagens para o aprendizado de máquina. Os algoritmos de aprendizagem profunda são baseados na aprendizagem de múltiplos níveis de representações/abstrações, às vezes sendo chamados apenas de redes neurais profundas [Goodfellow et al. 2016].

Esse tipo de algoritmo de rede neural utiliza metadados como entrada e processa os dados através de algumas camadas da transformação não linear dos dados de entrada para calcular a saída. Uma rede neural pode ter apenas uma única camada de dados, enquanto que uma rede neural profunda tem duas ou mais. As camadas podem ser vistas como uma hierarquia aninhada de conceitos relacionados ou de árvores de decisão. A resposta a uma pergunta leva a um conjunto de questões relacionadas mais profundas.

Os algoritmos de aprendizagem profunda possuem um recurso exclusivo que é a extração automática de recursos. Isso significa que esse tipo de algoritmo automaticamente capta os recursos relevantes necessários para a solução do problema mas precisa ver grandes quantidades de itens para ser treinado. Desta forma a carga sobre o programador para selecionar os recursos explicitamente fica reduzida, permitindo uma maior dedicação para, por exemplo, resolver desafios supervisionados, não supervisionados ou semi-supervisionados.

Numa rede neural de aprendizagem profunda, cada camada de dados oculta é responsável por treinar o conjunto exclusivo de recursos com base na saída da camada anterior. À medida que o número de camadas ocultas aumenta, a complexidade e a abstração de dados também aumentam. Isso forma uma hierarquia de recursos de baixo nível para recursos de alto nível. Com isso, torna-se possível que um algoritmo de aprendizagem profunda possa ser usado para resolver problemas de maior complexidade, consistindo em um grande número de camadas transformacionais não-lineares.

Hoje, o reconhecimento de imagens por máquinas treinadas via Aprendizagem Profunda em alguns cenários é melhor que o reconhecimento humano, e isso varia de simples figuras como imagens de um animal a indicadores de identificação de câncer em sangue e tumores em exames de ressonância magnética. Um exemplo de algoritmo de aprendizagem profunda é o *AlphaGo* do *Google* [Gibney 2016] que aprendeu o jogo GO, treinando por meio de sua rede neural jogando contra si mesmo repetidas vezes. A Figura 2.5 ilustra o processo de aprendizagem profunda.

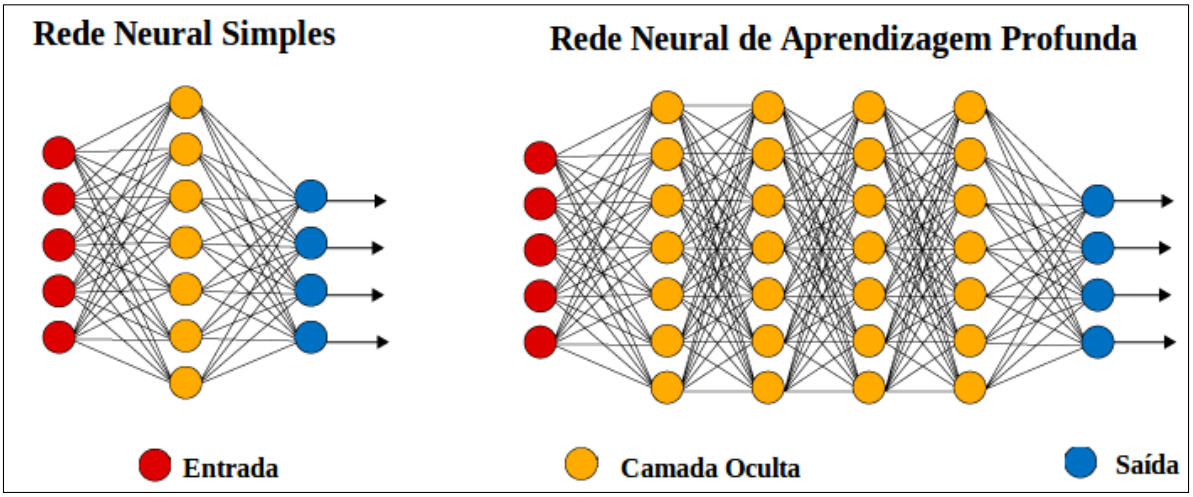


Figura 2.5: O processo do aprendizado profunda (adaptado de [Edwards 2018])

2.5 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (conhecido pela sigla PLN) escrita usa os conhecimentos léxicos, sintáticos e semânticos de um idioma bem como outras informações reais necessárias para determinar os significados das palavras e seus sentidos no contexto aplicado.

O PLN é considerado um dos mais importantes componentes da IA e da linguística computacional. Fornece uma interação perfeita entre computadores e seres humanos, dando aos computadores, a capacidade de entender a fala humana com a ajuda do aprendizado de máquina, propiciando uma interação entre a linguagem natural e o computador [Chopra et al. 2016].

O PLN fornece aos computadores a capacidade de reconhecer um contexto por meio de técnicas linguísticas para sua execução, envolvendo o uso de conhecimentos linguísticos no estudo do problema de geração e compreensão automática de linguagens naturais, extraindo informação, interpretando os sentidos, analisando sentimentos e até aprendendo conceitos com os textos processados.

O termo entidade, hoje amplamente utilizado em PLN, foi pela primeira vez utilizado em 1995, para a conferência *Message Understanding Conference* (MUC-6) [Grishman and Sundheim 1996], iniciada e financiada pela DARPA (Agência de Projetos de Pesquisa Avançada de Defesa) a fim de promover o desenvolvimento da tarefa de identificação e classificação de entidades nomeadas.

Desta iniciativa, constituíram-se três grandes classes de entidades: "*timex*" (datas e horas), "*numex*" (expressões numéricas) e "*enamex*" (que continham organizações, pessoas e localizações). Naquele momento, a MUC estava focada nas tarefas de Extração de Informação (IE).

2.5.1 Estudos Relacionados ao Processamento de Linguagem Natural

A tarefa de Reconhecimento Entidades Nomeadas (REN) consiste em identificar entidades nomeadas a partir de textos não-estruturados de forma livre e classificá-las dentro de um conjunto de tipos de categorias pré-definidas, tais como pessoa, organização e localização. Ferramentas e métodos computacionais descrevem, em larga escala os diferentes usos na extração da informação [Finatto et al. 2015]. Vários estudos e projetos relacionados a PLN têm sido realizados para extrair entidades nomeadas de textos.

Alberti e colegas em [Alberti et al. 2017], destacam a ferramenta Syntaxnet [Andor et al. 2016] como uma ferramenta que pode ser treinada para trabalhar com qualquer idioma, uma vez que a sintaxe da língua é ensinada para a ferramenta. O objetivo da *Google* (mantenedora e desenvolvedora do Syntaxnet) é fazer com que os computadores e sistemas consigam entender os diferentes idiomas que os seres humanos falam.

Bird e Loper, em [Bird et al. 2009], apresentam uma coleção de módulos e *corpora*, lançadas sob uma licença de código aberto, que permite o aprendizado e a elaboração de pesquisas em PLN usando a ferramenta *NLTK* [Bird and Loper 2004]. Destacam que a vantagem mais importante de usar essa ferramenta é que ela é totalmente autossuficiente, ou seja, não só fornece funções pré-definidas que podem ser usadas como blocos de construção para tarefas comuns de PLN, como também, fornece versões brutas e pré-processadas de

corpus padrão encontradas na literatura sobre PLN.

O trabalho de Aprosio e Moretti [Aprosio and Moretti 2016], apresenta a ferramenta Tint, um PLN simples para o italiano, baseada na ferramenta CoreNLP da Universidade de *Stanford* [Manning et al. 2014]. Segundo os autores, a ferramenta *Tint* fornece uma interface fácil para ampliar a anotação para novas tarefas e/ou idiomas comuns de processamento de linguagem natural.

Ribeiro e Medeiros, em [Ribeiro and Medeiros 2016], apresentam uma abordagem para extração das entidades nomeadas por meio da técnica da estatística da Máxima Entropia (ME) utilizando o *corpus* público chamado de Amazônia, como dicionário para treino. Esse trabalho faz uso da ferramenta OpenNLP [OpenNLP 2017] para extração das entidades e mostra que ela pode ser eficiente quando se considera um *corpus* específico para o domínio de pesquisa.

O trabalho de Weber e colegas, em [Weber et al. 2015], relata a construção de um modelo utilizando bases do *Wikipedia* e a *Dbpedia*. A construção do modelo é realizada por meio da classificação de entidades nomeadas relacionando *wikilinks* da *Wikipedia* (base textual) com instâncias e classes da *Dbpedia* (base de dados estruturada).

Zaccara, em [Zaccara 2012], apresenta uma plataforma para anotação e classificação automática de entidades nomeadas para notícias escritas em português brasileiro. O desenvolvimento consistiu na exposição de uma interface web para classificar e anotar o modelo de treinamento, e classificação automática de treinamento supervisionado.

Drury e colegas, em [Drury et al. 2017], descrevem a construção de um modelo que consiste em notícias coletadas na *Internet*, no período de 1996 a 2016 relacionadas à agricultura, escritas em português brasileiro, que anota informações de temporais, causais, entidades nomeadas e sentimentos em notícias agrícolas.

2.6 Principais ferramentas de PLN e características

Na última década muitas ferramentas de PLN foram disponibilizadas publicamente nas mais variadas linguagens de programação, como *Java* , *Python*, *R* e *C++*. Entretanto a seleção de uma ferramenta de PLN adequada aos objetivos deste trabalho, envolve critérios específicos que requerem um conhecimento prévio dos recursos de extração, das possibilidades de extensão do *corpora*, dos documentos disponíveis, bem como a condição de aplicar os conceitos específicos de PLN ao idioma português do Brasil mais especificamente.

A seguir serão apresentadas as ferramentas, baseadas na linguagem *Python* e *Java* que serviram como estudo na abordagem de extrair Entidades Nomeadas, bases na linguagem *Python* e *Java*.

2.6.1 Syntaxnet

Uma abordagem estocástica relativamente nova é o da ferramenta *SyntaxNet*, da *Google*, uma implementação de código aberto do método discutido em [Andor et al. 2016], que integra a estrutura do *TensorFlow* e acompanha um analisador sintático *ParseyMcParseface*. Os modelos são pré-tratados e treinados em conjuntos de dados da *Universal Dependencies* fornecidos gratuitamente em vários idiomas e que ajudam as máquinas a entender a linguagem, em um processo chamado Compreensão da Linguagem Natural. O Syntaxnet pode ser treinado e avaliado em qualquer um desses *corpora*.

O modelo de análise sintática *Parsey Universal* da Syntaxnet [Xhafa et al. 2017] faz uso de sete tipos de recursos diferentes:

- **palavras**: a palavra em análise juntamente com uma janela de três palavras em torno dela;
- ***tags***: *tags* atribuídas (previstas) às palavras vizinhas na janela e precedendo a palavra atual em análise;
- **sufixos**: terminações de todas as palavras que ocorrem na janela considerada;
- **prefixos**: início de todas as palavras que ocorrem na janela considerada;
- **capitalização**: transformação de todos os caracteres em minúsculas;
- ***char ngram***: subconjunto de caracteres (chamados de *ngram*) em todas as palavras que ocorrem na janela considerada;
- **outro**: qualquer símbolo que não inclua um caractere alfabético (por exemplo, pontuação, hifens, dígitos)

Segundo documentação [Syntaxnet 2017], a ferramenta Syntaxnet precisa ser instalada e configurada para se utilizar. Assim, todos os procedimentos são realizados utilizando-se linha de comando.

2.6.2 NLTK

A ferramenta NLTK (*Natural Language Toolkit*) [Bird et al. 2009] é amplamente documentada, e considerada relativamente fácil de aprender e simples de usar. Foi implementada como uma grande coleção de módulos minimamente interdependentes, organizados em uma hierarquia superficial.

A ferramenta de extração NLTK tornou-se popular nos últimos anos no ensino e na pesquisa. É uma ferramenta de código aberto, que possibilita trabalhar com dados de linguagem humana usando um conjunto de ferramentas de linguagem natural de ampla cobertura que

fornece uma estrutura simples, extensível e uniforme para atribuições, demonstrações e projetos.

A NLTK possui um conjunto de módulos escritos em *Python* que fornece muitas tarefas de processamento de PLN, *corpora* e ferramentas de análise de gráficas, utilizadas para explicar os conceitos básicos mediante análise gráfica e para mostrar como o algoritmo funciona na análise do processo de uma única sentença. [Bird et al. 2009].

O processo de extração de informação com a ferramenta NLTK começa com a entrada de um texto qualquer onde ele é dividido em várias sentenças menores, usando um segmentador. Cada sentença é então subdividida em palavras usando um *tokenizador*. A partir deste ponto cada *token* é marcado com uma etiqueta (*tag*) de acordo com a semântica, o que é considerado útil nas próximas etapas de extração das entidades. E no final é realizada a detecção de prováveis relações entre as diferentes entidades encontradas.

A Figura 2.6 ilustra os procedimentos de extração de informação compondo a arquitetura da ferramenta NLTK.

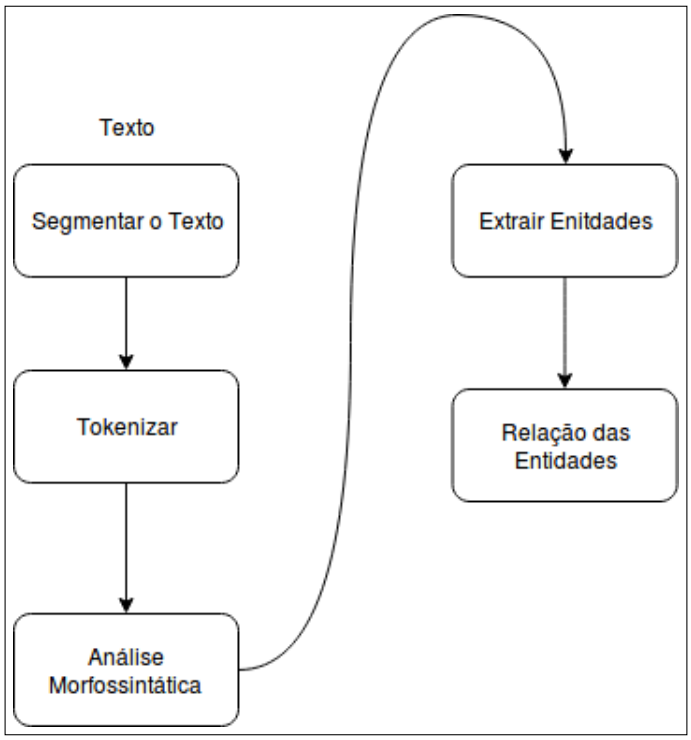


Figura 2.6: Arquitetura da ferramenta NLTK

Para ajudar no Processamento de Linguagem Natural normalmente são usados grandes *corpus* de dados linguísticos. A ferramenta NLTK já tem alguns disponíveis para uso, dependendo da linguagem. Especificamente para Português existe um *corpus* chamado de *macmorpho* formado por artigos publicados no jornal **Folha de São Paulo**, 1994 [Hamada and Neto 2015].

2.6.3 CoreNLP

A ferramenta CoreNLP [Manning et al. 2014], projetada pela equipe de desenvolvimento da Universidade de Stanford (*Stanford NLP Group*), fornece um conjunto de ferramentas de processamento de linguagem humana altamente flexível e extensível, que se integram facilitando a análise linguística de um texto. A arquitetura da CoreNLP inclui etiquetagem, reconhecimento de entidades nomeadas, analisadores e resolução de relação entre duas palavras ou frases nas quais ambas se referem à mesma entidade.

A configuração da CoreNLP é baseada em propriedades pré-definidas, que ajudam no conjunto que se deseja processar. Os modelos disponíveis são treinados principalmente para uso no idioma inglês para todas as funcionalidades. A CoreNLP é baseada em *Java*, e podendo ser executada por linha de comando com os seguintes anotadores:

- *tokenize*: o texto é subdividido em *tokens* individuais, ou seja, palavras, sinais de pontuação, etc.
- *ssplit*: fragmenta um texto em sentenças
- *pos*: o módulo *Stanford Part of Speech Tagger* atribui etiquetas da classe da palavra a cada *token* de acordo com um esquema do modelo e do *annotation*.
- *lema*: a lematização é o ato de deflexionar uma palavra para determinar o seu lema, ou seja, método de encontrar a raiz das palavra. O lematizador fornece o lema ou a forma de base para cada *token*.
- *ner*: o módulo Reconhecedor de Entidade Nomeada identifica os símbolos que são nomes próprios como membros de classes específicas, como nome de pessoa (al), nome da organização, etc.
- *parse*: o módulo *Parser* analisa e anota a estrutura sintática de cada sentença no texto. Esse módulo, na verdade, não é apenas um analisador sintático, mas também analisa a estrutura da frase e analisa a dependência.
- *dcoref*: o módulo *CorefAnnotator* implementa resolução de referência nominal e pronominal.

Existem alguns modelos disponíveis da CoreNLP para outras línguas, tais como chinês, alemão e francês, mas que ainda não abrangem todas as funcionalidades.

2.6.4 OpenNLP

A principal características da ferramenta OpenNLP [Chiele et al. 2015] é a disponibilização de diferentes recursos, tais como *tokenização*, segmentação de sentença, etiquetagem morfossintática, extração de entidade nomeada, extração de sintagmas, análise sintática e

resolução de correferência, sendo que esta última ainda não se encontra muito bem desenvolvida na ferramenta.

Modelada e escrita em *Java*, a OpenNLP é uma ferramenta de código, assim como outras, com licenciamento aberto mas diferentemente da ferramenta da (CoreNLP), que também é escrita em *Java* o código da OpenNLP não possui licença comercial.

A ferramenta OpenNLP contém um conjunto de componentes que podem ser utilizados na extração de informação para diferentes técnicas e para diferentes idiomas. Existem modelos disponíveis, pré-treinados, para alguns idiomas que ajudam na extração de informação para diferentes tipos de técnicas, tais como alemão, dinamarquês, holandês, inglês, sueco e português europeu. Além disso, a ferramenta OpenNLP também permite o treinamento de modelos para outros idiomas, por meio de uma coleção de dados que consiste em um conjunto de textos anotados com *tags*.

A interface de programação fornecida pela OpenNLP é composta por um conjunto de classes e/ou bibliotecas que podem ser utilizadas como linha de comando ou podem ser integradas utilizando-se chamadas das bibliotecas no desenvolvimento de uma aplicação que queira utilizar PLN.

2.6.5 Características da PLN

No PLN reconhecer uma entidade em um conjunto de dados não-estruturado é uma forma de extração de informações na qual busca-se classificar cada palavra deste conjunto como sendo de uma determinada categoria, tais como: pessoa-nome, organização, localização, data, hora, valor monetário, porcentagem ou nenhuma das opções.

Essa tarefa tem sua importância nos mecanismos de busca na Internet, tradução e indexação automática de documentos e em trabalhos que contenham informações mais complexas tarefas de extração.

Os avanços do PLN, associados com um poder computacional cada vez maior têm permitido o uso de novas técnicas de aprendizagem com cargas de processamento mais leves, diferente por exemplo do uso de algoritmos de aprendizado de máquinas mais antigos, como as árvores de decisão, baseados em sistemas de regras rígidas, semelhantes às regras existentes na escrita à mão.

Por exemplo com o uso da chamada marcação da fala, conhecida no PLN como *Part-of-Speech*, começou-se a ter estudos baseados em modelos estatísticos, usando probabilidade na atribuição de pesos reais aos recursos que compõem os dados de entrada.

Em [Borthwick and Grishman 1999] os autores, descrevem um novo sistema de reconhecimento estatístico chamado entidade (ou seja, "nome próprio") conhecido como ENME "(Entidade Nomeada com Máxima Entropia), ou em Inglês, "MENE"(*Maximum Entropy Named Entity*).

Outro trabalho nessa mesma direção é o estudo do Algoritmo de *Perceptron* [NETO and BONINI 2010] inventado em 1957 por Frank Rosenblatt nos EUA, que é a forma mais simples de rede neural artificial sendo utilizada como classificador linear, simulando um neurônio com entradas e pesos que, ajustados teriam a capacidade de aprender a se comportar de determinada forma.

Em [do Amaral and Vieira 2014], os autores apresentam uma ferramenta baseada na utilização do método probabilístico chamado de *Conditional Random Fields (CRF)*. Nesse trabalho é realizado um comparativo entre o CRF e modelo chamado de *Markov* de Máxima Entropia, ou simplesmente de Máxima Entropia.

2.7 Reconhecimento de Entidades Nomeadas

O REN defini-se como uma tarefa de identificar e classificar as entidades nomeadas para a compreensão do dado textual atribuindo uma categoria semântica para essas entidades e que é amplamente utilizada no PLN.

O REN [Nadeau and Sekine 2007] é a tarefa de identificar e classificar termos relevantes para a compreensão de um dado textual. Os termos que podem ser considerados entidades variam de acordo com o domínio de interesse e são comumente atribuídos a nomes que referenciam pessoas, organizações, locais, entre outros.

O estudo na área de REN é aplicado a diversas áreas e pode ser divido em três tipos de categorias, no reconhecimento de entidades nomeadas: REN utilizando aprendizado de máquina, que faz uso de técnicas de aprendizagem automática, onde são criados modelos que conseguem alcançar previsão de desempenho de acordo com o domínio de interesse; REN baseado em regras, consiste em definir heurísticas na forma de expressões regulares, analisando diversas características dos termos e da forma como são organizados no texto [Chiticariu et al. 2010]; REN baseada em aboragem híbrida, que reúne regras e aprendizagem de máquina.

Neste trabalho é utilizada a abordagem baseada em aprendizado de máquina com base que o aprendizado de máquina adota uma abordagem probabilística usando dados e resultados históricos e não apenas a entrada, mas números de outros fatores fornecendo a probabilidade de resultados diferentes e devido aos modelos que podem se adaptar às tendências de mudança e à flexibilidade de ajustar os parâmetros envolvidos.

2.8 Síntese do Capítulo

Este capitulo apresentou uma síntese de conceitos trabalhados ao longo desta dissertação, tanto de uso prático como de somente estudos, no desenvolvimento utilizando Aprendizado de Máquina.

Discorreu também sobre os trabalhos relacionados que utilizam técnicas de PLN para extração de dados. Por fim, apresentou uma breve discussão sobre algumas ferramentas para extrair informação utilizando ferramentas de PLN.

Capítulo 3

Análise do Problema e Proposta da Solução

Muito mais que simplesmente descrever, o PLN se dedica a criar soluções plausíveis para problemas pontuais relacionados com o reconhecimento e a reprodução da linguagem humana em alguma escala. Assim, neste capítulo, é apresentada uma análise do conjunto de dados utilizado e uma proposta de construção do sistema capaz de realizar a extração de entidades nomeadas.

O estudo realizado é suportado por algumas ferramentas que utilizam recursos de PLN e uma proposta de que extraia diferentes entidades-chave utilizando técnicas de PLN aplicadas ao DOU brasileiro que possui uma estrutura de informação diferenciada.

As ferramentas baseadas em PLN possuem mecanismos para compreender frases em linguagem natural e, assim, executar comandos ou exibir dados. A escolha de ferramentas que possam se adequar na extração de informação em português brasileiro, tais como entidades nomeadas, classificação das entidades morfossintáticas deve levar em consideração características diferentes no tratamento da extração das informações.

A ideia é processar textos em português brasileiro e analisar o que cada ferramenta pode oferecer de acordo com o que se deseja. Nas próximas seções serão apresentados o conjunto de dados utilizado, a descrição do ferramentas utilizadas neste trabalho e a construção da proposta de solução.

3.1 Descrição do conjunto de Dados

O Diário Oficial da União (DOU) foi criado durante o período imperial e se mantém, até os dias atuais, como o veículo de comunicação oficial mais importante do país. A partir do artigo 37 da Constituição da República de 1988, que determina o princípio da publicidade dos atos da administração pública foi oficializada a sua importância como o “veículo de acesso

universal e validação dos atos administrativos do Estado". Observa-se por exemplo que desde a sua criação, publicou fatos determinantes na história do Brasil, como: Proclamação da República, em 16 de novembro de 1889 (Decreto Federal n 1), Lei de Anistia, de 28 de agosto de 1979 (Lei n 6.683), Consolidação das Leis do Trabalho (Decreto-lei nº 5.452, de 1º de maio de 1943), Fim da escravidão (Lei nº 3353, de 13 de maio de 1.888), Normas para licitações e contratos da Administração Pública (Lei nº 8.666, de 21 de junho de 1993). (http://portal.imprensanacional.gov.br)

A Impressa Nacional, subordinada à Presidência da República Federativa do Brasil, é o órgão responsável pela produção do DOU e por meio dele torna público assuntos acerca do âmbito federal, conforme as normas para a publicação contidas no decreto 9.215 de 29 de novembro de 2017 (Os atos que a administração deve publicar nos diários oficiais estão enumerados em lista no art. 2o do Decreto n.o 4.520/02, sendo todos atos de interesse público [4520] e [Nacional]).

O DOU é publicado de segunda a sexta-feira, uma vez por dia, salvo nos feriados nacionais ou pontos facultativos da administração pública federal. Além disso, para que haja publicações em dias não previstos, é preciso autorização da Presidência da República ou da Casa Civil. Para quem deseja ter informações por assuntos, os conteúdos são oferecidos e estruturados em três seções que auxiliam a busca, a saber:

- Seção 1 – Destinada a publicação de atos normativos de interesse geral, tais como,leis, decretos, resoluções, instruções normativas, portarias e outros atos normativos.
- Seção 2 – Destinada à publicação de atos de interesse dos servidores da Administração Pública Federal.
- Seção 3 – Destinada à publicação de contratos, editais, avisos e atos governamentais e licitações

O processo de informatização do DOU teve início em 1994, três anos depois começou a ser disponibilizado de forma parcial na *Internet* e integralmente apenas em 2000. Esse processo teve a sua culminância em 25 de outubro de 2017 quando foi anunciado o fim da edição impressa marcada para 1 de dezembro do mesmo ano. Assim, atualmente o DOU é disponibilizado,gratuitamente, apenas em meio eletrônico.

3.1.1 Considerações sobre o Diário Oficial da União

Uma publicação que abriga diversos atos oficiais e normas possui aspectos particulares que devem ser compreendidos para quem deseja desvendar o Diário Oficial da União. De acordo com Cecília Andreotti Atienza, autora do livro "Documentação jurídica: introdução à análise e indexação de atos legais", [Atienza 1979] devem ser levados em consideração os seguintes aspectos:

- Título e histórico;
- Periodicidade: frequência da publicação;
- Numeração dos fascículos;
- Paginação;
- Finalidade;
- Arranjo ou organização: partes do diário, ordem e sequência dos atos publicados, de acordo com o órgão de onde emanam tais atos; e
- Suplementos: numeração e conteúdo dos suplementos e observações sobre como tomar conhecimento da existência dos suplementos.

O arranjo ou organização do DOU possui, normalmente, a seguinte sequência de atos: Atos do Poder Executivo, Presidência da República, Ministérios, Conselho Nacional do Ministério Público, Ministério Público da União, Tribunal de Contas da União, Poder Legislativo, Poder Judiciário, Entidades de Fiscalização do Exercício das Profissões Liberais.

Cabe destacar que, nem sempre, uma edição do DOU possui todos os tipos de atos elencados no arranjo, podendo ocorrer variações. É possível, por exemplo, que uma edição não contenha atos do Congresso Nacional, apenas os do Poder Executivo ou Judiciário. Por este motivo, a consulta ao sumário da edição é fundamental.

A extração de entidades nomeadas do DOU apresenta uma questão a ser tratada em um primeiro momento que é referente a forma como o mesmo é escrito. Embora tenha tido durante muito tempo, nas edições impressas a forma de jornal, a linguagem utilizada não é a jornalística, ou seja, não encontramos textos analíticos, crônicas ou notícias. Trata-se, exclusivamente, de informações dispostas em seções definidas cuja publicação em muitos casos é aguardada pelo órgão interessado ou servidor público.

Desta maneira, são dirimidas as possibilidades de erros ou diversas inferências para que não haja danos aos órgãos envolvidos e aos servidores públicos. As informações publicadas podem ser revistas, alteradas e até canceladas por meio de outras publicações posteriores, mas cada publicação segue o princípio de não promover dúvida a quem a lê.

Outro ponto a ser destacado na análise do DOU é o fato que este embora não seja um documento exclusivamente jurídico, apresenta terminologia e siglas deste ramo e se observa que essas podem variar ao longo do tempo. Essa característica tornou o processo deste trabalho diferenciado, por ter regras gramaticais com recursos expressivos nos textos com expressões jurídicas, que procuram alcançar o conceito de precisão e objetividade na exposição da informação nos textos do DOU. Isso significa que os textos podem variar os formatos quanto à disposição da informação dentro de um arquivo, assim como, a nomenclatura dos nomes dos arquivos. O DOU contém várias informações que podem não ser de interesse

numa dada pesquisa. Assim sendo, definir o que é relevante resulta em um problema que precisa ser equacionado para o processamento PLN do DOU.

3.2 Descrição do conjunto de ferramentas para a solução

Este trabalho também apresenta um estudo de quatro ferramentas de extração de informação que utilizam PLN. Mais especificamente as ferramentas OpenNLP, CoreNLP, Syntaxnet e NLTK, utilizando-se de dicionários previamente já disponíveis para cada ferramenta. Como resultado diferencial, foi elaborada uma proposta inicial de construção de um dicionário específico para extrair entidades nomeadas com melhor qualidade em comparação com os resultados obtidos com os dicionários disponíveis.

No caso das ferramentas OpenNLP e CoreNLP, com suas devidas adaptações, foi utilizado o corpus público Amazônia [1]. Para a ferramenta NLTK foi utilizado um corpus, também público, chamado de Nltk-Tagger-Portuguese disponível no GitHub[2]. Por sua vez, para a ferramenta Syntaxnet, que possui 40 corpora de vários idiomas pré-treinados, foi utilizado o Português brasileiro disponível no GitHub[3].

No estudo, as quatro ferramentas utilizadas apresentaram diferentes tipos de extração. A OpenNLP e CoreNLP realizaram uma extração que reconhece entidades nomeadas. Por outro lado, a NLTK e a Syntaxnet foram utilizadas para extração de classificações morfossintáticas.

Levando-se em consideração das abordagem realizadas, percebeu-se que, com o dicionário existente, ocorreram muitos erros de identificação de entidades e classificações equivocadas de termos, optando-se por criar um dicionário próprio utilizando OpenNLP para extração das entidades nomeadas.

3.3 Construção de um *Corpus*

Um *corpus* é um conjunto de enunciados representativos em linguagem natural construído com um propósito específico. É um conjunto linguístico que por meio de uma seleção de dados coletados aleatoriamente serve de base para a extração de informação. A qualidade de um *corpus* depende do seu tamanho, ou seja, quanto mais treinamento, maior é a quantidade de textos anotados o que implica em informações extraídas com mais precisão.

A construção de um *corpus* requer como ponto de partida a seleção e classificação de textos. Uma grande quantidade textos é imprescindível pois fornecem subsídios para a categorização, entretanto, a sua qualidade deve ser levada em consideração para que seja possível

[1] https://www.linguateca.pt/floresta/corpus.html
[2] https://github.com/fmaruki/Nltk-Tagger-Portuguese
[3] https://github.com/tensorflow/models/blob/master/research/syntaxnet/g3doc/universal.md

um treino efetivo. Assim, é preciso ter uma grande quantidade de textos devidamente selecionados que forneçam os subsídios necessários para a correta extração de entidades nomeadas.

Desta maneira, utilizar um *corpus* já existente é em geral inviável, uma vez que pode gerar resultados diversos dos pretendidos e não ser efetivo pois o processo de seleção de textos leva em conta o objetivo pretendido e um contexto específico, que embora possam apresentar semelhanças com a situação em qual se busca a aplicação, talvez não consiga extrair as entidades requeridas. Assim muitas vezes o problema da extração de determinados tipos de textos pode tornar o processo trabalhoso e demorado, por que o *corpus* sugerido não consegue extrair suficientemente a informação desejada.

No aprendizado de máquina as ferramentas possuem diferente recursos para realizar o processo de extração. O aprendizado automatizado e o uso de técnicas de treinamento permitem que grande parte do conhecimento relevante seja adquirido diretamente dos dados. Neste trabalho é descrito o resultado da construção de um corpus específico, utilizando os recursos da ferramenta OpenNLP para geração de um modelo.

Uma alternativa aos programas de rotulação de corpus é a anotação manual por meio da inserção dos sintagmas entre parênteses angulares <>. A anotação criteriosa do corpus pode acrescentar insumos valiosos para a identificar e interpretar os padrões textuais.

Na anotação de um *corpus* são fornecidos indicadores do que é relevante sobre um conjunto de dados em linguagem natural, esses indicadores muitas vezes vêm na forma de anotações - metadados que fornecem informações adicionais sobre o texto. No entanto, para ensinar um computador de maneira eficaz, é importante fornecer os dados certos e ter dados suficientes para a aprendizagem. Embora haja softwares para anotação de *corpora*, nem sempre eles acomodam as categorias de análise de uma determinada busca textual, especialmente quando se lida com diferenciadas análises discursivas como o DOU.

3.3.1 Classificação das categorias

O campo Reconhecimento de Entidades Nomeadas tem prosperado nos últimos anos, destinando-se a extrair e classificar as menções de designadores, a partir de textos, como nomes próprios, organizações, lugares e expressões financeiras e temporais, nos mais variados idiomas.

No entanto alguns idiomas como Português brasileiro, assim como termos jurídicos deste ainda precisam ser explorados. O formalismo nos textos do DOU mostra que mesmo com o uso de *corpus* de trabalhos já realizados pode-se ter extrações que não evidenciam o que é real no sentido da análise sintática do contexto.

Um dos pontos de partida deste estudo é a definição de categorias a serem utilizadas para anotações e posteriores interpretações. A seguir são descritas as fases de aprendizagem, construção e funcionamento do corpus específico para extração de informações no DOU. Normalmente, o processo de construção de um corpus é realizado levando-se em considera-

ção as seguintes atividades:

1. Identificar quais são as possíveis categorias que precisam ser descobertas. No caso em particular da proposta deste trabalho envolvendo o DOU, as seguintes categorias de entidades foram identificadas: Cargo, Data, Evento, Lei, Lugar, Número, Organização, Pessoa, Processo e Valor-monetário.

2. Construção de um arquivo com trechos de informação de interesse. No caso deste trabalho, o DOU é composto por 3 seções. Logo o *corpus* deve conter trechos das 3 seções de vários diários oficiais.

3. Processar o arquivo, fazendo interferência manual em algumas categorias que são consideradas subjetivas e não padronizadas no arquivo. No caso do DOU, por exemplo, categorias como Pessoa, Organização ou Cargo são relevantes do ponto de vista de busca de dados. Em outro momento processar através de utilização de expressão regular e anotar expressões que sejam padrões como, por exemplo, Valor-monetário, Número ou Data.

3.3.2 Regras de Anotação

A etiquetagem por *tags* é especialmente projetada para o uso na fase de treinamento de um algoritmo de aprendizado. As técnicas de aprendizado de máquina supervisionado exigem como entrada um *corpus* de treino com exemplos corretamente identificados do problema que se deseja aprender a resolver.

As unidades mínimas que possuem uma relação de determinação são chamadas de sintagmas conhecidas como *tokens*. Na extração de informações a partir de textos, a ideia é que o sistema saiba escolher sequências de palavras com alto poder discriminatório e potencial informativo. Para isso os sintagmas nominais (SN) apresentam-se como candidatos naturais, pois, de um ponto de vista linguístico, eles tipicamente carregam significado substantivo, desempenham papéis semânticos e geralmente trazem o tema do enunciado.

Em [Perini 2017], o autor explica que um sintagma consiste em um conjunto de elementos que constituem uma unidade significativa dentro da sentença e que mantêm entre si relações e dependência e de ordem.

O processo de aprendizagem inicia-se com a criação de um arquivo de treino com uma coleção de sentenças que constituem em palavras anotadas com *tags* de acordo com a classificação de sua categoria, inferindo um conjunto de atributos que caracterizam cada *token* existente no texto.

O próximo passo, é um resultado obtido de um *corpus* construído por meio do treino do arquivo utilizando a classe *NameFinderME*, da biblioteca OpenNLP, que conta com características que auxiliam na identificação de padrões, classificação e probabilidades associadas das entidades e resume as regras de identificação de entidades nomeadas.

Para finalizar, o *corpus* é submetido para os textos do DOU onde são extraídas as entida-

des nomeadas utilizando a classe *TokenNameFinder*, da biblioteca OpenNLP, reconhecendo então as entidades existentes nos textos do DOU.

3.4 Proposta de solução

O Aprendizado de Máquina é uma técnica computacional que apresenta características de aprendizagem automática, baseadas em decisões nas experiências acumuladas de soluções bem sucedidas de problemas anteriores, e que melhoram automaticamente com a experiência, com o uso de uma variedade de algoritmos que interativamente aprendam com os dados para melhorar o resultado.

A aprendizagem pode ser diferenciada com base nos algoritmos utilizados, que podem ser divididos em diferentes técnicas de aprendizado. No processo de extração das Entidades do DOU é utilizado a técnica de aprendizado supervisionado para a tarefa de classificação das entidades-chaves dos textos. Para este tipo de técnica existem alguns algoritmos que expressam o resultado a partir de representação de atributos previamente rotulados. Para este trabalho é utilizada o algoritmo de Máxima Entropia para problemas de processamento de textos.

Levando em consideração o problema descrito nos estudos já realizados este trabalho propõe uma solução envolvendo um conjunto de elementos que extraiam entidades-chave do DOU brasileiro utilizando ferramenta de PLN. Uma solução que atue, de forma transparente, desde a coleta do DOU até a visualização dos resultados da extração das entidades-chaves.

Uma especificação arquitetural é essencial para analisar e descrever propriedades de sistemas complexos, permitindo ter uma visão geral completa do trabalho. Nesta arquitetura, são definidas a organização das funcionalidades, os artefatos produzidos e os procedimentos executados. Além disso, é analisada a aplicação do ferramental desenvolvido com o objetivo de avaliar os resultados em uma pequena amostra, exemplificando a utilização.

A arquitetura da solução proposta inclui um fluxo de processos que pode ser alterada ou adaptado de acordo com mudanças que possam a vir acontecer, tais como: leiaute do DOU, entidades adicionadas, e ajuda na organização dos componentes que possa impactar sobre a qualidade apresentada por ele.

Esta seção apresenta os passos para identificar, marcar e extrair entidades nomeadas utilizando a OpenNLP como ferramenta de PLN e o modelo de Máxima Entropia, conhecido como *Maxen*t, uma técnica de estatística indutiva. A OpenNLP é uma ferramenta *open source*, incluindo Apache OpenNLP nesse desafio como seu suporte à língua portuguesa.

Os experimentos de extração e classificação de informações serão realizados experimentos com os modelos existentes e um com modelo novo, desenvolvido de forma personalizada para os problemas específicos do *dataset*, que nesse caso é o DOU.

3.4.1 *Package* OpenNLP

A ideia motivadora por trás da Máxima Entropia (ME) é que se deve preferir os modelos mais uniformes que também satisfazem quaisquer restrições. O modelo ME pode facilmente combinar diversos recursos, por isso é amplamente adotado para muitas línguas naturais, no processamento e nas tarefas de conversão de texto, como marcação de parte da fala (POS), reconhecimento de entidade nomeado e extração de relações[Sun et al. 2007].

A OpenNLP oferece vários modelos pré-construídos para diferentes idiomas, usando técnicas de aprendizado de máquina, com os algoritmos de *Maxent* e *Perceptron*, com fontes e exemplos de textos anotados usados nos treinos de novos modelos. A documentação fornecida no site é clara e estruturada, juntamente com comentários nos códigos brutos [OpenNLP 2017].

A estrutura da Máxima Entropia estima probabilidades com base no princípio de fazer o mínimo possível de suposições, além das restrições impostas. Tais restrições são derivadas de treinamentos com dados, expressando alguma relação entre características e resultados. A distribuição de probabilidade que satisfaz a propriedade acima é aquela com a maior entropia [Chieu and Ng 2002].

A ME pode envolver, de maneira conveniente, recursos avançados para classificação e melhor desempenho de classificação em comparação com outros classificadores. A análise detalhada de um *corpus* mostra que também é uma tarefa importante e desafiadora.

O modelo de Máxima Entropia, também conhecido como *Maxent*, trata-se da técnica de estatística indutiva com regressão logística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica. Neste caso, a entropia é definida como uma medida única para incerteza, ou seja, mede a quantidade de informação contida em uma variável aleatória [Ribeiro and Medeiros 2016].

3.5 Reconhecimento de Entidades Nomeadas do DOU

O Reconhecimento de Entidades Nomeadas (REN) é um passo importante, que envolve detectar entidades e, em seguida, classificá-las quando se trata de extração de informação. É possível defini-lo como uma tarefa cujo objetivo é identificar as entidades nomeadas bem como sua posterior classificação e de acordo com os critérios de delimitação estabelecidos para uma categoria semântica. As categorias mais comuns incluem nomes, locais, organizações e coisas [Nadeau and Sekine 2007].

3.5.1 Entidades e suas denominações

Nesta seção são apresentados os critérios de delimitação das entidades mencionadas, nomeadamente os que dizem respeito às categorias Cargo, Evento, Data, Lei, Lugar, Número, Organização, Pessoa, Processo e Valor-monetário.

As Entidades Nomeadas (EN) são tratadas do ponto de vista das palavras e não dos termos, que apresentam um ou mais desígnios. Isso significa que são consideradas as palavras, e suas respectivas classes, de cada EN identificada no *corpus*, seja ela um termo simples. As EN consideradas no contexto deste trabalho são as seguintes:

- **Cargo**: se trata da palavra atribuída ao cargo que uma pessoa exerce em uma organização ou partição pública qualquer.
- **Evento**: uma categoria é classificada como Evento quando ocorre algum episódio que tenha algum significado, ou ação que possa ter gerado algo de novo e são descritos no DOU. Esse tipo de categoria é muito comum na Seção 2, onde acontecem por exemplo, exoneração ou nomeação de um(a) servidor(a) ou aposentadoria de um servidor.
- **Data**: a Categoria Data é classificada desde que satisfaça uma das expressões descritas a título de exemplo a seguir: 25 de novembro de 2017, 19 de agosto, 17/03/2017, 05/06/16.
- **Lei**: a classificação da categoria Lei é dada quando encontra algo como as seguintes expressões a seguir: Lei n 5548/97, LEI 1865/89.
- **Lugar**: o Lugar considera alguns endereços, locais ou regiões ou municípios que sejam mencionados, tais como: Rua Jorge Alberto Flores-205 , Boa Vista do Buricá, Campus Rio Branco, Pernambuco.
- **Numero**: a categoria Número descreve todo numerário explorado nos textos do DOU.
- **Organizacao**: a Organização é uma categoria onde são classificados todos os órgãos possíveis e existentes no ambiente do governo em geral. Elas podem ser autarquias, ministérios, tribunais, universidades e institutos federais entre outras repartições que possam vir ser classificadas como Organização. Além da classificação do setor público tem aquelas que podem ser empresas privadas que eventualmente são mencionadas, principalmente na Seção 3 do DOU.
- **Pessoa**: a categoria Pessoa é classificada para qualquer indivíduo que seja nomeado e como nome próprio. Essa é uma das categoria bem difíceis de ser mapeada em razão de sua subjetividade.
- **Processo**: a classificação de um processo ocorre quando se tem uma expressão que possua a palavra Processo com um sequencial de números de um certa formatação.

- **Valor-monetario**: o Valor-monetário é uma expressão onde são expressos valores de dinheiro ou moeda.

3.6 Descrição do processo arquitetural

Neste capítulo serão abordados a solução da arquitetura desenvolvida para extrair as Entidades Nomeadas, os procedimentos que foram realizados e a descrição dos artefatos produzidos.

Para gerar o *corpus* específico (DOU-*Corpus*), foi desenvolvida uma ferramenta de construção de *corpus* a qual recebe as informações do DOU, processa a informação de acordo com dados fornecidos. Os instrumentos utilizados são um conjunto de ferramentas desenvolvidas em linguagem de programação Java que se integram com o objetivo de extração de Entidades Nomeadas.

Há necessidade de estruturas que permitam mostrar os resultados que foram construídos visando atender uma necessidade da transparência das informações. Um sistema de análise, ademais, foi desenvolvido para extração de estruturas a partir de texto analisado pelo DOU-*Corpus*.

Na sequencia, são descritas as etapas de desenvolvimento que seguiram um planejamento que visava atender uma necessidade de construção do modelo da extração, processamento e armazenagem dos resultados para validar a extração das entidades-chaves uma proposta de uma arquitetura é desenhada que extraia as entidades presentes no DOU e mostre, no final do processo, uma interface os resultados. A construção do modelo basicamente seguiu as seguintes fases:

- Coleta do DOU - Esta é a fase inicial que precisa ser realizar para começar o desenvolvimento da solução deste trabalho. Assim os diários foram extraídos de um repositório disponível no link (http://grafica.ufes.br/diario-oficial-da-uniao) onde estão disponíveis, em arquivos pdf, todos desde 2012.

- Extrator do DOU - Esta fase é necessária para que os dados do DOU seja manipuláveis. Esse Conversor é disponibilizado no repositório da UNB.

- Criar Arquivo de Treino - Já nesta fase é realizada uma junção de vários trechos de textos aleatórios dos diários que compreendem entre os meses de agosto de 2015 à outubro de 2017.

- Treino do Arquivo - Este passo é realizando utilizando a ferramenta OpenNLP como que treina um modelo como saída.

- Extração das entidades-chave - Nesta fase o modelo gerado é submetido a diários já convertidos pelo conversor, utilizando a biblioteca do OpenNLP.

- Avaliação do Modelo de Treino - Esta é uma fase que é avaliado o resultado.
- Armazenagem no MongoDB - As entidades chaves são armazenadas em uma base do *MongoDB* configurada.
- Construção de Interface de Usuário - Esta é a fase que é mostrado o resultado das extrações realizadas.

A Figura 3.1 apresenta um desenho do fluxo das etapas do sistema proposto desde a criação do dicionário até a extração das entidades-chave, que serão detalhadas nas próximas seções.

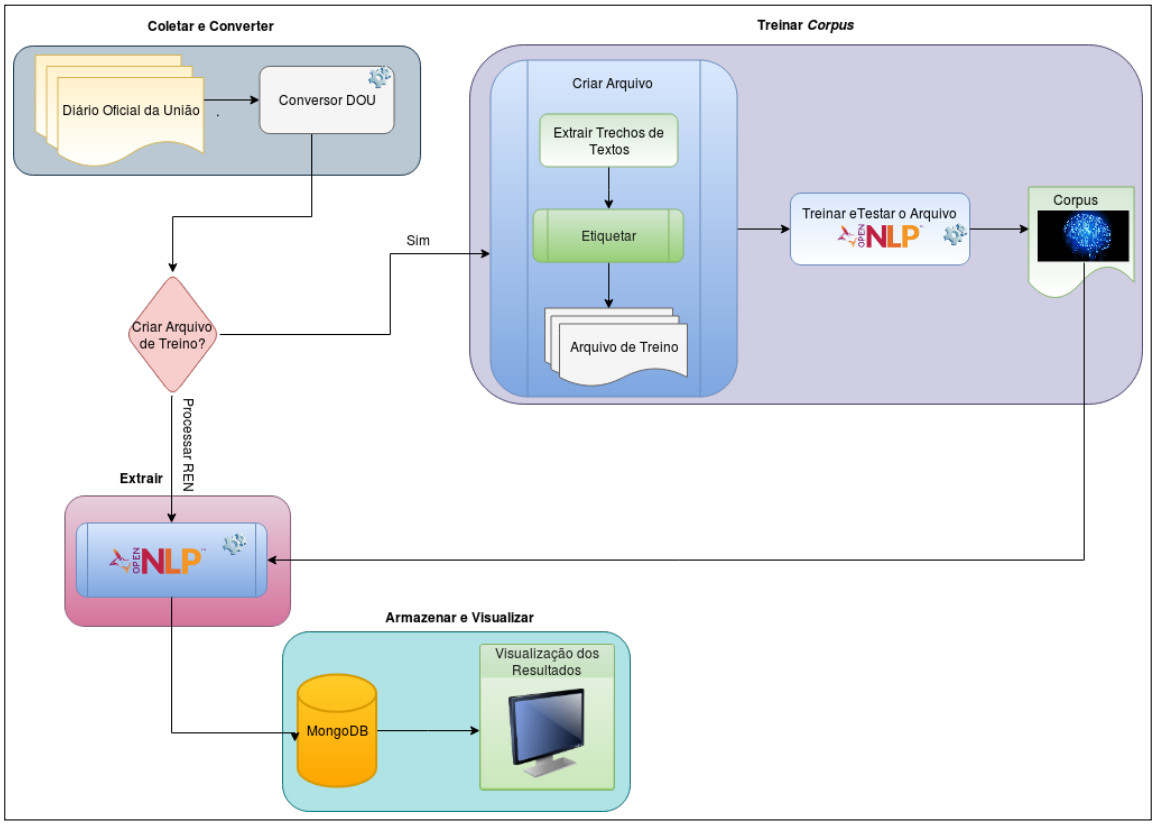


Figura 3.1: Arquitetura do sistema proposto para extração de Entidades Nomeadas

A Figura 3.1 mostra as dependências entre os diferentes módulo. A análise básica consiste na coleta dos DOU's. Por sua vez, é o insumo para extração das entidades nomeadas, passando pela conversão dos DOU's, criação de arquivos de treino chamado de *corpus*, treino de um modelo e extração de entidades nomeadas.

3.6.1 Coleta do DOU

A coleta e a preparação dos dados foram os primeiros passos realizados e consistiram, basicamente, no colhimento de diários. Embora seja uma tarefa cuja complexidade seja menor, ela apresenta grande relevância para o desenvolvimento das etapas posteriores, haja

vista que fornece os subsídios necessários para o planejamento, desenvolvimento e experimentos. Esse primeiro momento pode ser considerado o norteador das etapas seguintes. Para manter uma organização na coleta, os arquivos foram separados e classificadas por ano de publicação, já que a primeira recuperação dos arquivos consistia entre os anos de 2015 a 2017.

As coletas forneceram dados na construção inicial do arquivo de treino e geraram insumos de dados do *corpus*, assim como para extração das entidades nomeadas. Todos os dados utilizados para este trabalho foram obtidos através do link `(http://grafica.ufes.br/diario-oficial-da-uniao`.

3.6.2 Conversor do DOU

Levando-se em consideração que os arquivos em formato *pdf* precisam ser convertidos para um formato de arquivo editável, foi utilizado um conversor para realizar a conversão dos diários. Esse conversor trata-se de um aplicativo, desenvolvido em *Java*, que converte arquivos do DOU em arquivos textuais.

Visando atender as peculiaridades desse trabalho, foram feitas adaptações no conversor para auxiliar a leitura dos textos. Assim, originalmente a ferramenta transforma os dados do *pdf* em um arquivo texto, mas nesse caso foi necessária um ajuste para que os textos fossem extraídos e salvos em um formato em modo *JSON*(*JavaScript Object Notation*)[4].

Na prática, *JSON* é um modelo para armazenamento e transmissão de informações no formato texto e que é bastante utilizado por aplicações web. O arquivo *JSON* continua sendo um arquivo em texto puro mas a informação é representada atribuindo um nome ou rótulo que descreve o seu significado e a seguir o seu valor que é uma notação derivada do *Javascript* para representar informações. Por exemplo, para representar a Organização do Ministério das Cidades utiliza-se a seguinte sintaxe: “organizacao”: “Ministério das Cidades”.

3.6.3 Criando o *Corpus*

Neste trabalho o *corpus*, chamado como DOU-*Corpus*, é uma coleção de textos produzidos naturalmente a partir do DOU, com uma dimensão considerável e em formato eletrônico, podendo ter informações anotadas ou não. Observa-se, então que se trata de uma etapa para identificar e classificar as entidades do DOU. Cabe ressaltar, que a identificação das entidades implica em delimitá-las, ou seja, estabelecer onde uma entidade começa e termina. Paralelamente, classifica-se a categoria correspondente da entidade identificada.

A ambiguidade da entidade, onde uma pode ser categorizada e representada por mais de um significado representa uma complexidade que extrapola o escopo deste trabalho e por este motivo é apenas aqui mencionada essa possibilidade.

[4]https://www.json.org/

A OpenNLP foi a ferramenta de PLN adotada. Os estudos preliminares, mostraram que utilizando a OpenNLP o processo de criar e treinar um *corpus* fica mais simples. Isso, por que existe uma documentação e trabalhos tais como do [Ribeiro and Medeiros 2016] que explicam como realizar o processo de criação e extração de entidades.

Portanto a escolha da OpenNLP justifica-se primeiro pela existência de documentações e blogs, como [Kart 2017], que explicam utilizando a ferramenta OpenNLP com exemplos demonstrando a criação de *corpus* e extração de entidades nomeadas, tornando a ferramenta mais clara e a construção de um *corpus* mais simples.

O treino neste caso, pode-se dizer que é uma compilação do *corpus* utilizando a OpenNLP onde a saída é em arquivo com a extensão do tipo *.bin* com um dicionário de todas categorias mapeadas.

Contudo, para que a compilação funcione corretamente alguns cuidados precisam ser observados na preparação dos dados de treinamento. As anotações devem ser fornecidas para entidades nomeadas no arquivo de treinamento usando o seguinte formato: *<START:**categoria**> Entidade <END>*.

Importante ressaltar, caso houver apenas um tipo de entidade nomeada no arquivo de treinamento, como por exemplo, "pessoa" não será necessário mencionar o tipo específico da categoria. Exemplo: *<START> João <END> e <START> Carlos <END> são irmãos*.

Vários tipos podem ser dados em um único arquivo de treinamento. Um exemplo para treinar frases com vários tipos é: *<START:pessoa> João <END> e <START:pessoa> Carlos <END> trabalham na <START:organizacao> UNB <END>*.

O formato do *corpus* anotado com as categorias bedece às seguintes regras:

- Uso das *Tag*: uma categoria está localizada entre duas *tags*: <START:?> Entidade <END>
- Espaço entre as *Tags*: deverá ter pelo menos um espaço em branco entre as *tags*. Por exemplo, no trecho "...no <START:lei> art. 24 <END> ,..."pode-se observar um espaço entre palavra ***no*** e a *tag* ***<START:lei>***; o mesmo acontece antes da *tag* ***<END>***.

Segue-se na Figura 3.2 um exemplo mais detalhado do processo de anotação de entidades nomeadas na sintaxe da atual versão da ferramenta na extração entidades nomeadas.

```
Ratifico a dispensa de licitação, com fulcro no <START:lei> art. 24 <END>
, inciso X, da <START:lei> Lei n o 8.666 <END> / <START:numero> 93 <END> ,
referente à locação  dos imóveis situados à Rua Rui Barbosa, no 1.535 e no
1.555, no <START:local> Município de Campo GrandeMS <END> , no valor
mensal respectivo de <START:valor-monetario> R9.241, 00 <END> e
<START:valor-monetario> R$7.295,00 <END> , por <START:data> 30 meses <END>
, a contar de <START:data> 8 de junho de 2017 <END> , sendo locador
<START:pessoa> ELIAS PANAGIOTIS KONTOS <END> , CPF no <START:numero>
403.491.871-34 <END>.
```

Figura 3.2: Processo de anotação de entidades nomeadas

Para criar o *corpus*, treinar e extrair as entidades nomeadas duas etapas foram realizadas. A primeira etapa dessa solução foi realizar a extração do texto do DOU utilizando-se de um conversor, que processa um arquivo em *pdf* e extrai as informações em formato de texto. A segunda etapa foi utilizar uma sequência de adições de textos já treinados da primeira etapa.

Os primeiros resultados para criação do *corpus* se deram utilizando um escopo delimitado de trechos de textos que poderia servir como base para o arquivo de treino. Essa primeira fase do treino serviu como processo de estudo da construção do arquivo de treino e utilizou um total de 5960 trechos de textos extraídos aleatoriamente do DOU durante o período de 2017, dentre os meses de julho até outubro.

Na segunda etapa, se deu a continuidade da construção do *corpus*, adicionando novos trechos de textos. Mas desta vez foram adicionados trechos de um período maior, compreendido entre os anos de 2015 a 2017. A extração dos trechos foi realizada de forma automatizada e aleatória, utilizando um programa construído em *Java*. Assim obteve-se no final um arquivo com 46.690 linhas como trechos de textos extraídos aleatoriamente. Observa-se que nesta segunda etapa vários ciclos de ajustes de anotação manual e automática no arquivo de treino foram necessários, para se obter a melhor precisão possível das informações extraídas.

3.6.4 Treino do DOU-*Corpus*

A extração de Entidades Nomeadas com a OpenNLP, passa por técnicas que precisam ser executadas. Primeiro, é necessário realizar a construção do *corpus* conforme descrito na Seção 3.2. onde são apresentados os passos para treinar e validar a criação do modelo. A construção do *corpus*, e o seu tamanho, tem influência direta nos resultados extraídos de Entidades Nomeadas, i.e., na qualidade das informações.

Os sistemas de PLN possuem diversos usos no ambiente corporativo. Eles são implementados para otimizar o atendimento a clientes, para criação de aplicativos mais eficientes e até mesmo de ferramentas de acessibilidade. As técnicas descritas a seguir foram exploradas na realização deste trabalho.

3.6.4.1 *Tokenização*

A segmentação de palavras é conhecida em inglês como *tokenization* (em inglês, *tokenization*), que é o processo de cortar uma dada sentença em partes menores que são chamadas de *tokens*. O *token* é considerado como a menor unidade que pode ser identificada dentro do contexto da linguagem natural.

A identificação de *tokens*, ou *tokenização* é uma importante etapa do pré-processamento para extrair unidades mínimas de textos. Na *tokenização* das sentenças dadas em fragmentos mais simples, a biblioteca OpenNLP [OpenNLP 2017] fornece três classes diferentes:

- *SimpleTokenizer* - Esta classe *tokeniza* o texto bruto fornecido usando classes de ca-

racteres.

- *WhitespaceTokenizer* - Esta classe usa espaços em branco para *tokenizar* o texto fornecido.
- *TokenizerME* - Esta classe converte o texto bruto em *tokens* separados, usando o algoritmo da Máxima Entropia para tomar suas decisões.

O processo de *tokenização* é usado em várias tarefas tais como verificação ortográfica, processamento de buscas, classificação de documentos, etc. Neste trabalho, a *tokenização* é utilizado para realizar o Reconhecimento de Entidades Nomeadas (REN) aproveitando um modelo já disponível na OpenNLP[5] [Fonseca et al. 2015]. A Figura 3.3 ilustra um exemplo de resultado do processo de *tokenização* com a OpenNLP.

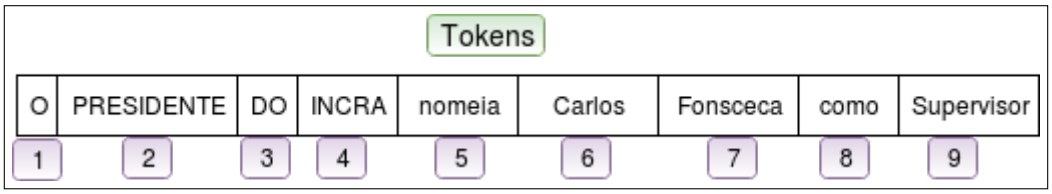


Figura 3.3: Tokenização de uma sentença

O resultado do processo de *tokenização* são *tokens* com seus valores semânticos da categoria, suas posições inicial e final dentro da sentença e suas probabilidades em termos de acurácia.

3.6.4.2 Processo de Etiquetagem

A sentença depois de ser *tokenizada* passa pelo processo de etiquetagem que consiste na busca e classificação de entidades nomeadas de acordo o modelo de treino passado como recurso.

É o procedimento de anotar cada termo do texto em alguma categoria morfológica mapeada. Entre as informações associadas aos itens lexicais, encontram-se a categoria gramatical do item, tais como substantivo e valores morfossintáticos, como pessoa, moedas, números, datas, lugares e organizações. Um exemplo do processo da etiquetagem pode ser visualizado na Figura 3.4

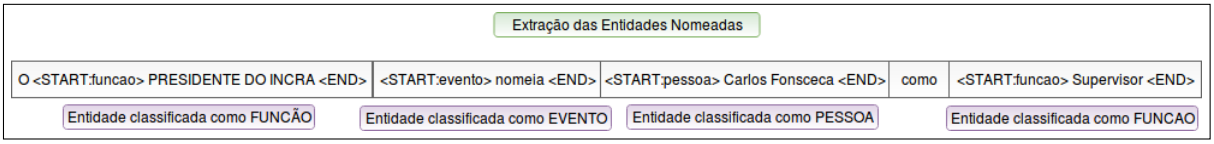


Figura 3.4: Exemplo de resultado do processo de etiquetagem

[5]http://opennlp.sourceforge.net/models-1.5/pt-token.bin

3.6.5 Armazenamento

O ciclo da extração de entidades nomeadas, conforme mostrado na arquitetura do sistema proposto (Fig. 3.1) se fecha na criação de um objeto de formato *JSON* para cada diário processado, contendo os dados como: data do diário, seção, as sentenças sem anotações e com anotação e a coleção de *tokens* do resultado da extração das entidades nomeadas.

Esse objeto *JSON* é armazenado em uma base de dados que serve como insumo para amostragem. E para armazenar foi escolhido um banco de dados *NoSql*. Pela rapidez e facilidade de se buscar os resultados o MongoDB foi considerado como uma alternativa para armazenar os resultados. A Figura 3.5 mostra parte de um resultado de um diário treinado.


```
{
    "id": "5b94372c1dbc05430c1c7f17",
    "identificador": "2017_11_23_assinado_do2",
    "dou": 1511402400000,
    "edicao": "do2",
    "sentencas": [
          {
           "sentenca": "Nº 1069 - NOMEAR VANESSA REGINA RIBEIRO PIMENTA, para exercer o cargo de Chefe da Assessoria de Comunicação Social do Gabinete do Ministro de Estado Chefe da
Secretaria de Governo da Presidência da República, código DAS 1015 ELISEU LEMOS PADILHA",
           "titulo": "PORTARIA Nº 1 9, DE Wed Nov 22 00:00:00 BRST 2017 (PORTARIAS DE 22 DE NOVEMBRO DE 2017)",
           "sentenca_anotada": "Nº 1069 - <START:evento> NOMEAR <END> <START:pessoa> VANESSA REGINA RIBEIRO PIMENTA <END> , para exercer o cargo de Chefe da Assessoria de Comunicação
Social do Gabinete do Ministro de Estado Chefe da Secretaria de Governo da <START:funcao> Presidência da República <END> , código DAS 1015 ELISEU LEMOS PADILHA",
           "tokens": [
               {
                   "start": "3",
                   "token": "NOMEAR",
                   "tipo": "evento",
                   "prob": "0.8052533155977102",
                   "end": "3"
               },
               {
                   "start": "4",
                   "token": "VANESSA REGINA RIBEIRO PIMENTA",
                   "tipo": "pessoa",
                   "prob": "0.9690723101837213",
                   "end": "7"
               },
               {
                   "start": "32",
                   "token": "Presidência da República",
                   "tipo": "funcao",
                   "prob": "0.9846741824064017",
                   "end": "34"
               }
           ],
           "precisao": "Precisão:1.0#Frequência: 0.333#Qualidade geral: 0.5"
       }
```


Figura 3.5: Resultado parcial de um treino em formato *JSON*

3.6.6 Interface de Usuário

O procedimento de armazenar levou em conta a possibilidade de consulta dos resultados na busca dos dados. Todo procedimento de treino para cada diário foi armazenado em uma base de dados na forma de documentos no *MongoDB*. Assim uma consultas poderá ser realizadas conforme a necessidade. A Figura 3.6 mostra um exemplo de uma interface de usuário ilustrando um processo de etiquetagem.

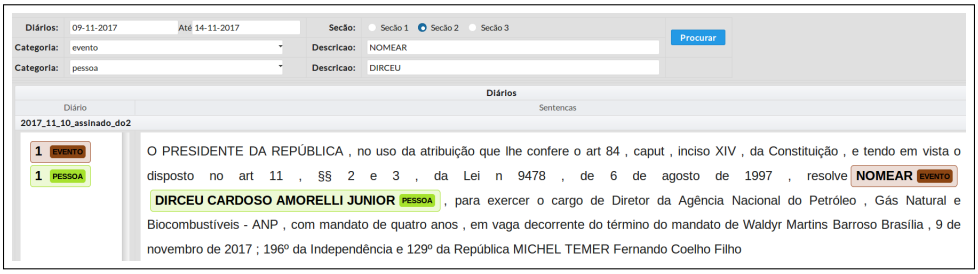


Figura 3.6: Exemplo interface de usuário mostrando a etiquetagem de uma sentença

3.7 Síntese do Capítulo

Este capítulo descreve a arquitetura do sistema proposto para extrair entidades nomeadas do Diário Oficial da União, baseado em ferramenta de PLN.

O capítulo discute os conceitos envolvendo o processo de construção do corpus e o processo de reconhecimento de entidades nomeadas, descrevendo as categorias de informações extraídas e usadas neste trabalho.

Capítulo 4

Implementação da solução

Tal como descrito anteriormente, este trabalho tem como objetivo o desenvolvimento de um conjunto de módulos que permitam realizar tarefas de conversão, treino e extração de entidades nomeadas do DOU, resultando em uma ferramenta que permita uma análise de forma quantitativa e qualitativa das informações extraídas.

A seguir são apresentados detalhes sobre a implementação e a abordagem seguida no desenvolvimento de cada uma das ferramentas componentes do sistema proposto.

Observa-se que não houve uma ordem de execução sequencial dos módulos componentes. Os módulos passaram por várias iterações de refinamento até serem considerados como ideal para continuidade. Tanto que foram desenvolvidos separadamente como forma de possuir uma independência de execução.

4.1 Tecnologias acessórias utilizadas para o desenvolvimento

De modo a tornar a implementação mais rápida e eficiente, foi preciso adotar alguns arcabouços e ferramentas que facilitassem o processo de desenvolvimento do protótipo.

O desenvolvimento de aplicações passa por escolhas de ferramentas, linguagem utilizada, componentes, armazenamento de dados, configurações, decisões sobre as estruturas que formarão o sistema, controles e outros atributos de qualidade, que envolvam a arquitetura de um sistema que podem impactar na qualidade no contexto na qual foi proposto.

Durante as etapas de desenvolvimento do sistema buscou-se o uso e melhorias de arcabouços disponíveis, identificando e descrevendo a organização dos módulos, visando a codificação e especificação de cada um deles. As ferramentas e arcabouços utilizadas neste trabalho foram as seguintes:

- *Intellij* como ambiente integrado de desenvolvimento [Krochmalski 2014]

- Linguagem *Java* versão 8 como linguagem de programação [Spell 2015]
- *Spring Boot* como recurso de desenvolvimento de micro serviços [Gutierrez 2016]
- *Apache Maven* como ferramenta de controle de bibliotecas [Porter et al. 2017]
- *Apache OpenNLP* como ferramenta de PLN [OpenNLP 2017]

4.2 Fluxo do Processo

O processo de desenvolvimento da solução de extração de entidades nomeadas compreende um conjunto de atividades a serem executadas. A Figura 4.1 mostra uma representação abstrata das atividades do sistema de extração, ilustrando seu funcionamento.

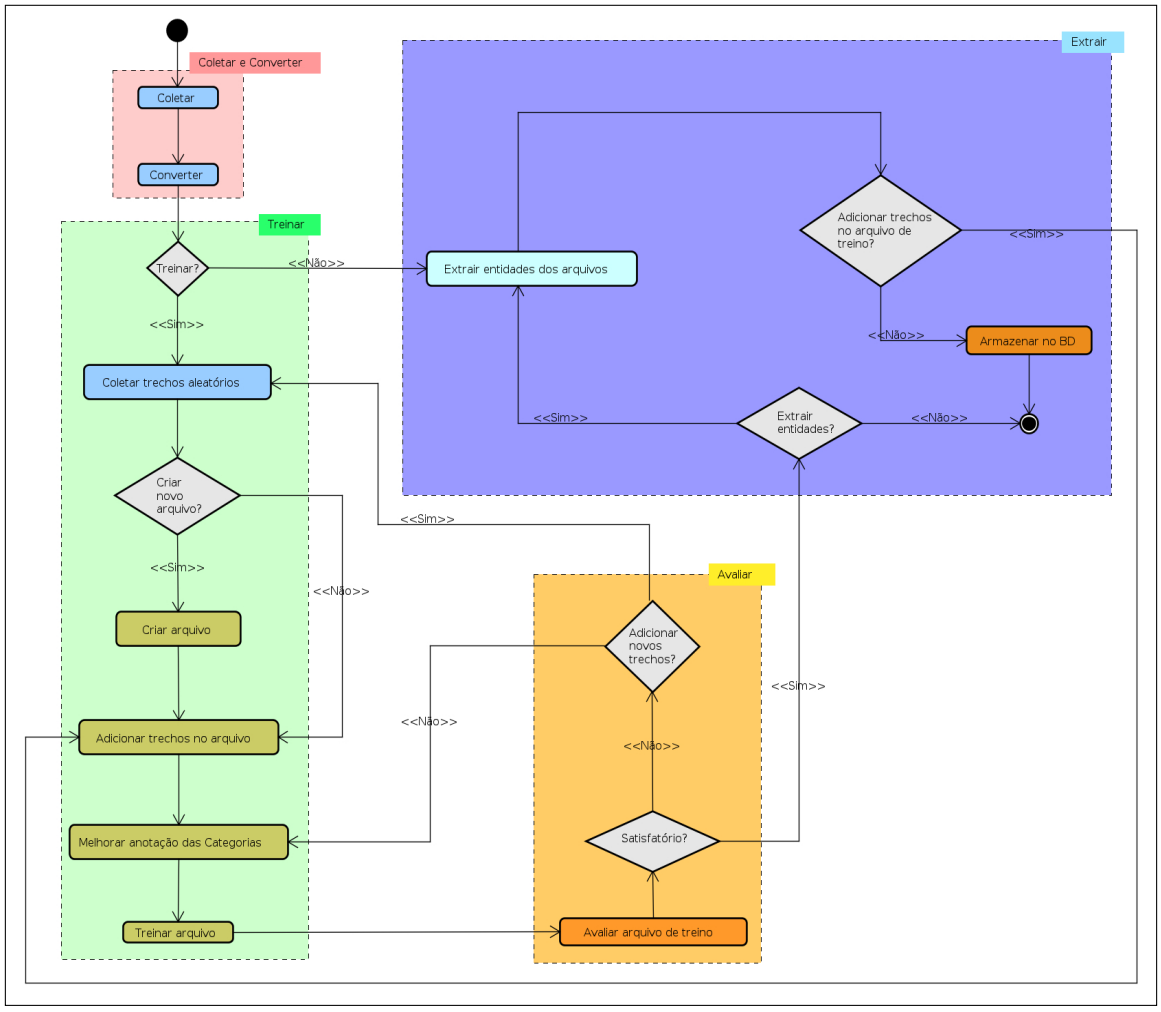


Figura 4.1: Fluxo do processo para extração de entidades nomeadas

Essencialmente, a Figura 4.1 mostra um fluxograma, que tem a função de representar graficamente o processo de extração de entidades nomeadas, mostrando as atividades realizadas durante o desenvolvimento deste trabalho. O fluxograma ilustra as etapas do processo,

a ordem e a interação entre as atividades evidenciando de forma simplificada o que está sendo executado.

Na Figura 4.1 estão detalhadas quatro atividades macros, que são utilizadas durante o processo de extração de extrair entidades nomeadas. A coleta e o conversor são as atividades iniciais que servem de insumos para as próximas atividades. O treino fornece um ciclo que tem como atividades de criar um *corpus*, treinar e melhorar as anotação das categorias. A avaliação do corpus é uma atividade que tem por objetivo verificar a qualidade do *corpus*. A extração é a atividade que processa os diários e extrai as entidades utilizando o modelo gerado apartir do *corpus* e armazenando na base de dados.

4.2.1 A Coleta e o Conversor DOU

A primeira atividade macro do sistema de extração é a coleta dos diários, que nada mais é do que realizar o *donwload* de arquivos do DOU, um passo estritamente manual de busca no link (http://www.grafica.ufes.br/diario-oficial-da-uniao).

Uma saída atual do DOU é ilustrada na Figura 4.2, que mostra um exemplo com três colunas com informações de publicações de um determinado dia. Observa-se que o leiaute da disponibilização das informações pode sofrer mudanças de uma publicação para outra.

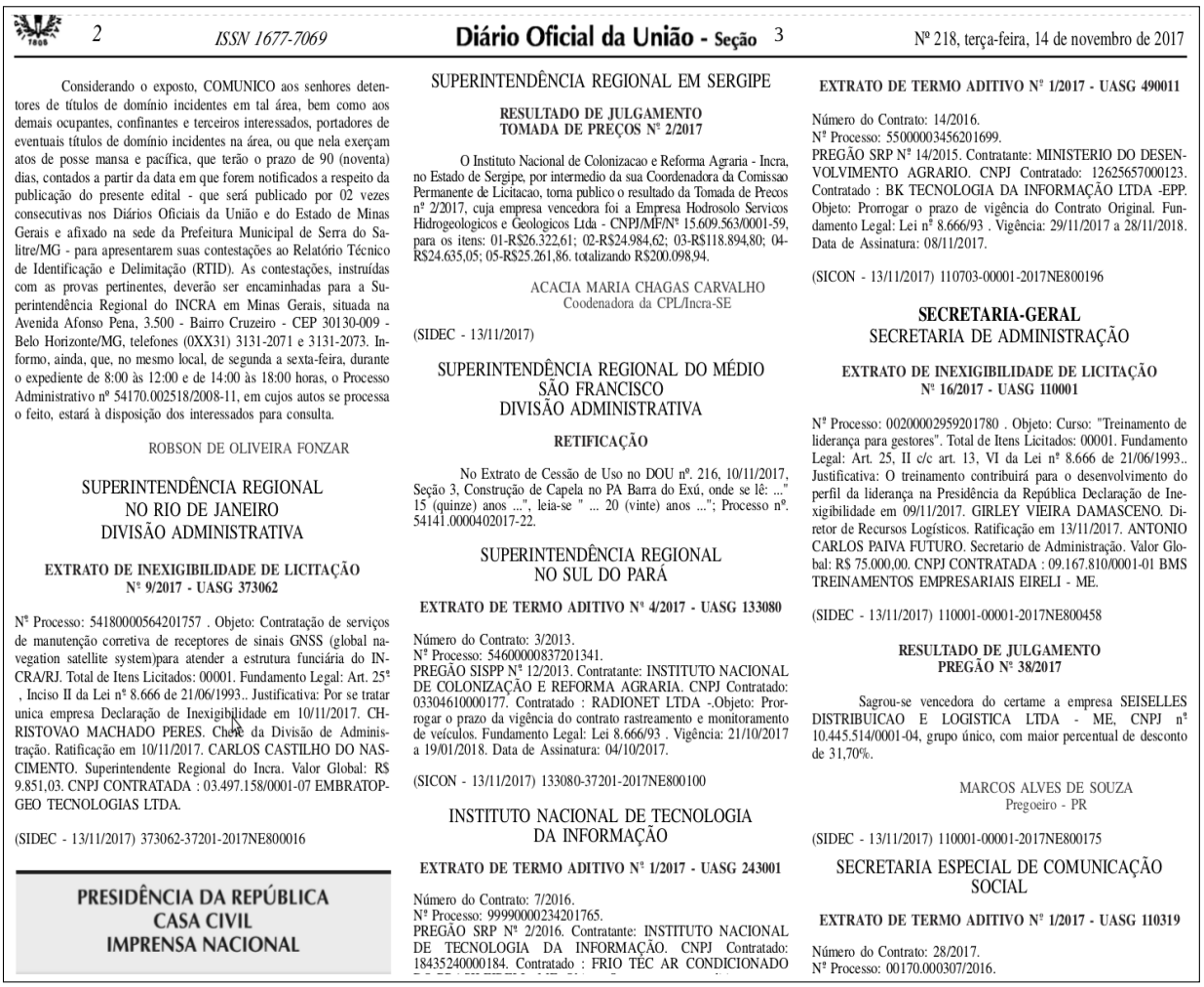
Realizada a coleta dos diários, o passo seguinte é o procedimento de converter os arquivos, salvos em *pdf*, para um modo editável. Este é um passo que inicialmente foi realizado convertendo os arquivos *pdf* para modo texto comum, ou seja, para *txt*.

Entretanto, na análise dos resultados preliminares da conversão , percebeu-se que as informações, agora editáveis, ainda continuavam difíceis de serem lidas e processadas. Então para melhorar a leitura dos dados, o Conversor foi adaptado para fornecer uma saída de arquivo em modo *JSON* que atribui um nome ou rótulo que descreve o seu significado e o seu valor.

Isso levou o processo de treino e extração ser mais eficiente na hora da leitura dos dados do DOU. Os resultados foram organizados de tal forma que tenham em cada arquivo *JSON* uma coleção de sentenças organizadas para cada DOU convertido, e uma espécie de cabeçalho com as relativas informações e características do próprio DOU, tais como: a data da publicação, seção e a seguir a coleção dos dados. Um exemplo de resultado de coleta com conversão para formato *JSON* pode ser visualizado na Figura 4.3.

A Figura 4.3 mostra uma organização dos itens dos textos de um Diário com *JSON*, facilitando a leitura das informações. Um Diário é classificado como único de acordo com uma data e a seção. Assim o leiaute do arquivo *JSON* para cada diário foi criado, com as seguintes propriedades:

- *dou*: possui a data do Diário.
- *secao*:representa a seção do Diário.



2 ISSN 1677-7069 Diário Oficial da União - Seção 3 Nº 218, terça-feira, 14 de novembro de 2017

Considerando o exposto, COMUNICO aos senhores detentores de títulos de domínio incidentes em tal área, bem como aos demais ocupantes, confinantes e terceiros interessados, portadores de eventuais títulos de domínio incidentes na área, ou que nela exerçam atos de posse mansa e pacífica, que terão o prazo de 90 (noventa) dias, contados a partir da data em que forem notificados a respeito da publicação do presente edital - que será publicado por 02 vezes consecutivas nos Diários Oficiais da União e do Estado de Minas Gerais e afixado na sede da Prefeitura Municipal de Serra do Salitre/MG - para apresentarem suas contestações ao Relatório Técnico de Identificação e Delimitação (RTID). As contestações, instruídas com as provas pertinentes, deverão ser encaminhadas para a Superintendência Regional do INCRA em Minas Gerais, situada na Avenida Afonso Pena, 3.500 - Bairro Cruzeiro - CEP 30130-009 - Belo Horizonte/MG, telefones (0XX31) 3131-2071 e 3131-2073. Informo, ainda, que, no mesmo local, de segunda a sexta-feira, durante o expediente de 8:00 às 12:00 e de 14:00 às 18:00 horas, o Processo Administrativo nº 54170.002518/2008-11, em cujos autos se processa o feito, estará à disposição dos interessados para consulta.

ROBSON DE OLIVEIRA FONZAR

SUPERINTENDÊNCIA REGIONAL NO RIO DE JANEIRO DIVISÃO ADMINISTRATIVA

EXTRATO DE INEXIGIBILIDADE DE LICITAÇÃO Nº 9/2017 - UASG 373062

Nº Processo: 54180000564201757 . Objeto: Contratação de serviços de manutenção corretiva de receptores de sinais GNSS (global navegation satellite system)para atender a estrutura funciária do INCRA/RJ. Total de Itens Licitados: 00001. Fundamento Legal: Art. 25º , Inciso II da Lei nº 8.666 de 21/06/1993.. Justificativa: Por se tratar unica empresa Declaração de Inexigibilidade em 10/11/2017. CHRISTOVAO MACHADO PERES. Chefe da Divisão de Administração. Ratificação em 10/11/2017. CARLOS CASTILHO DO NASCIMENTO. Superintendente Regional do Incra. Valor Global: R$ 9.851,03. CNPJ CONTRATADA : 03.497.158/0001-07 EMBRATOPGEO TECNOLOGIAS LTDA.

(SIDEC - 13/11/2017) 373062-37201-2017NE800016

PRESIDÊNCIA DA REPÚBLICA
CASA CIVIL
IMPRENSA NACIONAL

SUPERINTENDÊNCIA REGIONAL EM SERGIPE

RESULTADO DE JULGAMENTO TOMADA DE PREÇOS Nº 2/2017

O Instituto Nacional de Colonizacao e Reforma Agraria - Incra, no Estado de Sergipe, por intermedio da sua Coordenadora da Comissao Permanente de Licitacao, torna publico o resultado da Tomada de Precos nº 2/2017, cuja empresa vencedora foi a Empresa Hodrosolo Servicos Hidrogeologicos e Geologicos Ltda - CNPJ/MF/Nº 15.609.563/0001-59, para os itens: 01-R$26.322,61; 02-R$24.984,62; 03-R$118.894,80; 04-R$24.635,05; 05-R$25.261,86. totalizando R$200.098,94.

ACACIA MARIA CHAGAS CARVALHO
Coodenadora da CPL/Incra-SE

(SIDEC - 13/11/2017)

SUPERINTENDÊNCIA REGIONAL DO MÉDIO SÃO FRANCISCO DIVISÃO ADMINISTRATIVA

RETIFICAÇÃO

No Extrato de Cessão de Uso no DOU nº. 216, 10/11/2017, Seção 3, Construção de Capela no PA Barra do Exú, onde se lê: ..." 15 (quinze) anos ...", leia-se " ... 20 (vinte) anos ..."; Processo nº. 54141.0000402017-22.

SUPERINTENDÊNCIA REGIONAL NO SUL DO PARÁ

EXTRATO DE TERMO ADITIVO Nº 4/2017 - UASG 133080

Número do Contrato: 3/2013.
Nº Processo: 54600000837201341.
PREGÃO SISPP Nº 12/2013. Contratante: INSTITUTO NACIONAL DE COLONIZAÇÃO E REFORMA AGRARIA. CNPJ Contratado: 03304610000177. Contratado : RADIONET LTDA -.Objeto: Prorrogar o prazo da vigência do contrato rastreamento e monitoramento de veículos. Fundamento Legal: Lei 8.666/93 . Vigência: 21/10/2017 a 19/01/2018. Data de Assinatura: 04/10/2017.

(SICON - 13/11/2017) 133080-37201-2017NE800100

INSTITUTO NACIONAL DE TECNOLOGIA DA INFORMAÇÃO

EXTRATO DE TERMO ADITIVO Nº 1/2017 - UASG 243001

Número do Contrato: 7/2016.
Nº Processo: 99990000234201765.
PREGÃO SRP Nº 2/2016. Contratante: INSTITUTO NACIONAL DE TECNOLOGIA DA INFORMAÇÃO. CNPJ Contratado: 18435240000184. Contratado : FRIO TEC AR CONDICIONADO

EXTRATO DE TERMO ADITIVO Nº 1/2017 - UASG 490011

Número do Contrato: 14/2016.
Nº Processo: 55000003456201699.
PREGÃO SRP Nº 14/2015. Contratante: MINISTERIO DO DESENVOLVIMENTO AGRARIO. CNPJ Contratado: 12625657000123. Contratado : BK TECNOLOGIA DA INFORMAÇÃO LTDA -EPP. Objeto: Prorrogar o prazo de vigência do Contrato Original. Fundamento Legal: Lei nº 8.666/93 . Vigência: 29/11/2017 a 28/11/2018. Data de Assinatura: 08/11/2017.

(SICON - 13/11/2017) 110703-00001-2017NE800196

SECRETARIA-GERAL
SECRETARIA DE ADMINISTRAÇÃO

EXTRATO DE INEXIGIBILIDADE DE LICITAÇÃO Nº 16/2017 - UASG 110001

Nº Processo: 00200002959201780 . Objeto: Curso: "Treinamento de liderança para gestores". Total de Itens Licitados: 00001. Fundamento Legal: Art. 25, II c/c art. 13, VI da Lei nº 8.666 de 21/06/1993.. Justificativa: O treinamento contribuirá para o desenvolvimento do perfil da liderança na Presidência da República Declaração de Inexigibilidade em 09/11/2017. GIRLEY VIEIRA DAMASCENO. Diretor de Recursos Logísticos. Ratificação em 13/11/2017. ANTONIO CARLOS PAIVA FUTURO. Secretario de Administração. Valor Global: R$ 75.000,00. CNPJ CONTRATADA : 09.167.810/0001-01 BMS TREINAMENTOS EMPRESARIAIS EIRELI - ME.

(SIDEC - 13/11/2017) 110001-00001-2017NE800458

RESULTADO DE JULGAMENTO PREGÃO Nº 38/2017

Sagrou-se vencedora do certame a empresa SEISELLES DISTRIBUICAO E LOGISTICA LTDA - ME, CNPJ nº 10.445.514/0001-04, grupo único, com maior percentual de desconto de 31,70%.

MARCOS ALVES DE SOUZA
Pregoeiro - PR

(SIDEC - 13/11/2017) 110001-00001-2017NE800175

SECRETARIA ESPECIAL DE COMUNICAÇÃO SOCIAL

EXTRATO DE TERMO ADITIVO Nº 1/2017 - UASG 110319

Número do Contrato: 28/2017.
Nº Processo: 00170.000307/2016.

Figura 4.2: Leiaute de um Diário

- *dados* : este possui uma coleção dos itens do Diário.
  - *titulo* : o título do assunto em questão.
  - *autoridade* : o responsável pelo item.
  - *descricao*: o detalhe da divulgação do assunto em questão.

4.2.2 Treinar e Extrair

A construção de um *corpus* específico para o DOU passa pela criação de um arquivo que contenha vários trechos de textos retirados aleatoriamente dos diários. Mas é um processo onde novos trechos adicionados continuamente ao longo de sua construção até que as condições da qualidade do modelo sejam consideradas satisfatórias.

Um *corpus* tem a função de conter, como um dicionário, as informações de possíveis classificadores mapeados no arquivo de treino. Mas a qualidade do *corpus* depende necessariamente da quantidade de linhas e de classificadores mapeados no arquivo de treino. Portanto, a adição de trechos no arquivo de treino é um ciclo necessário até se atingir a

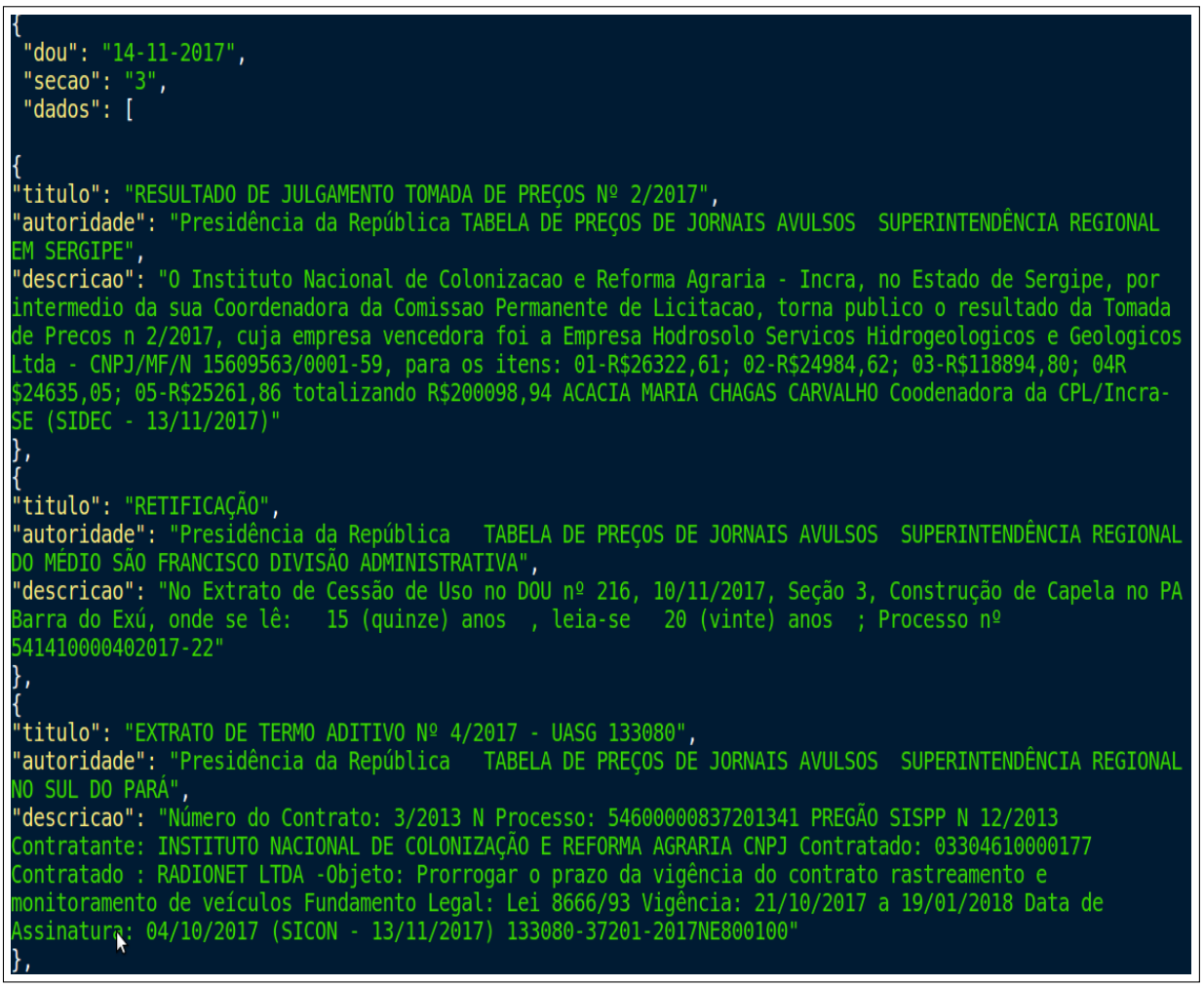


Figura 4.3: Resultado de conversão utilizando o Conversor de *pdf* para *JSON*

qualidade de extração presumida.

Neste trabalho, inicialmente foi criado um arquivo com 6 mil linhas com 400 mil palavras de textos retirados aleatoriamente do DOU no período entre 2015 e 2017, cobrindo as 3 seções. Mas à medida em que houve novas adições de diários se fez necessária uma nova atualização do arquivo de treino, adicionando-se trechos aleatórios retirados de DOUs publicados anos entre 2016 e 2017.

No começo do desenvolvimento deste trabalho, não se tinha informação de como funcionaria a extração já com o resultado com a sentença anotada com classificadores. Buscou-se então estudar mais detalhadamente o código da OpenNLP para entender sobre como trazer o resultado de uma sentença já anotada.

Assim, de forma aleatória, para cada arquivo foram tirados mais trechos de textos já treinados e adicionados ao arquivo de treino.

O passo seguinte foi realizar, de forma manual e visual, anotações que o *corpus* não conseguiu extrair.

No final do ciclo, obteve-se um arquivo com 60 mil linhas de trechos de texto com mais

de 4 milhões de palavras entre as anotadas e não anotadas do arquivo de treino. O manual da OpenNLP, sugere que pelo menos 15.000 sentenças devem estar disponíveis no arquivo de treinamento, para que o modelo treinado possa ter um bom desempenho [OpenNLP 2017].

O próximo passo foi extrair e classificar as entidades nomeadas de 470 arquivos. Para extrair as informações foi necessário o desenvolvimento de uma *engine* que preparasse uma solução onde a saída fosse também um arquivo *JSON* a ser armazenado na base de dados *NoSQL*.

Esta abordagem teve como princípio extrair e adicionar a sentença em JSON e assim posteriormente armazenar em um banco de dados.

Foi utilizado um banco de dados *NoSQL*, que é um banco de dados não relacional que não faz uso primário de tabelas e, no geral, não usa SQL para efetuar a manipulação de dados. E como a abordagem deste trabalho tem um resultado de documento *JSON*, o tipo de armazenamento adotado para esta forma foi a utilização do *MongoDB*, um banco de dados orientado a documentos, que utiliza conceitos de dados que são autocontidos, isto é, tem como característica conter todas as informações em único documento, onde o único pré-requisito é possui um identificador único universal (UUID) [Banker 2011].

O *MongoDB*, é considerado um banco de dados *NoSQL*, é um mecanismo de armazenamento e recuperação de dados que são modelados de formas diferentes das relações tabulares por ter ausência de *SQL* (*Structured Query Language*), ou seja, não traz a ideia de ter um modelo relacional e nem a linguagem SQL. Possui como característica ser de código-fonte aberto, ser multiplataforma e ter um conjunto de aplicativos *JSON* que são utilizados como sintaxe do *MongoDB*.

Então o resultado foi uma anotação na sentença descrita na Sub-seção do Conversor 4.2.1, onde são descritos os detalhes da conversão dos diários. Estas atividades são realizadas no bloco Treinar e Extrair ilustrado na Figura 4.1, que englobam as atividades de treino, de extração e no final do armazenamento do resultado da extração.

4.2.3 Avaliação

A atividade de avaliação de resultados de extração é um recurso que traz o desempenho do processo da qualidade do treino por meio do método de comparação dos resultados, que corresponde à uma análise sistemática das informações extraídas. No PLN, tarefa de a avaliação é uma medida que atesta se as necessidades atendem metas. Critérios e formas precisas de avaliação precisam ser definidos. A calibração ou precisão da avaliação depende de um conjunto suficiente de casos para os quais se sabe a resposta, ou de condições diferentes em que o sistema deve dar a mesma resposta, ou de precisão de resposta.

A avaliação permite observar se o resultado é satisfatório ou não, e pode ser realizada de duas formas :

- Primeiro é feita uma avaliação visual das informações extraídas, por exemplo sobre a precisão da extração.
- Segundo é realizada uma avaliação pela interface da OpenNLP extraindo dados numéricos de Precisão, Qualidade Geral e da Frequência do item extraído pelo *corpus*, que são as medidas básicas usadas na avaliação de eficiência no processo de extração.

Esta tarefa de avaliação é representada na Figura 4.1 pelo bloco Avaliar, que ilustra uma atividade que avalia o arquivo de treino, com condicional de satisfação dos resultados ou não.

4.3 Resultados Quantitativos

A abordagem quantitativa da avaliação, procura avaliar o processo de extração de informação com foco na efetividade da extração.

Nesta seção, são apresentados os resultados obtidos a partir da análise dos dados coletados nos experimentos realizados. Os resultados serão apresentados de forma quantitativa, comparando os dados extraídos do DOU com as quatro ferramentas testadas com a proposta de um novo modelo de treinamento para extrair entidades nomeadas utilizando a ferramenta OpenNLP.

Aqui também são analisados os erros frequentes que foram identificados, tais como problemas na conversão dos arquivos. O universo de dados deste trabalho é um conjunto de 470 diários do ano de 2017, que foram convertidos e processados para uma análise dos resultados.

A ideia é processar textos em português brasileiro e analisar o que cada ferramenta pode oferecer de acordo com o que se deseja. A seguir são apresentados e analisados os resultados quantitativos obtidos nos experimentos realizados com as ferramentas de extração utilizadas nesse trabalho.

4.3.1 Avaliação Quantitativa das ferramentas

Nesta seção são apresentados os resultados de um conjunto de dados quantitativos que foram extraídos do DOU, pelas quatro ferramentas estudas, utilizando os devidos *corpus* discutidos no item 3.2.

No caso das ferramentas OpenNLP e CoreNLP, com suas devidas adaptações, foi utilizado o *corpus* público Amazônia para extração de Entidades Nomeadas.

Para a ferramenta NLTK foi utilizado um *corpus*, também público, chamado de *Nltk-Tagger-Portuguese*. Por sua vez, para a ferramenta Syntaxnet, que possui 40 *corpora* de vários idiomas pré-treinados, foi utilizado o Português brasileiro.

No caso das ferramentas NLTK e Syntaxnet foram extraídos *tokens* que são morfossintaticamente reconhecidos. Toda informação de extração foi baseada em 2.605 páginas de informação do DOU, abrangendo o período específico entre 21 de agosto de 2017 até 25 de agosto de 2017.

4.3.1.1 Reconhecimento de Entidades Nomeadas

Nesta subseção é realizada uma comparação entre a (OpenNLP e a CoreNLP) no processo de extração e classificação de entidades nomeadas nos textos de linguagem natural do DOU. A extração, processa um conjunto de categorias pré-definidas, tais como pessoa, organização, local, data e outras referências específicas definidas.

Na Tabela 4.1 são apresentados os resultados da quantidade de *tokens* encontrados pelas duas ferramentas, conforme a distribuição das categorias. A ferramenta CoreNLP conseguiu processar um total de 2.370.305 *tokens* e a OpenNLP um total de 597.648.

Tabela 4.1: Resultados Comparativos de extração com OpenNLP e CoreNLP

Categoria	**Ferramentas**	
	OpenNLP	**CoreNLP**
Numérico	29.471	104.413
Organização	259.241	709.380
Pessoa	80.284	226.227
Lugar	97.387	714.530
Data	56.340	96.941
Outros	74.925	1.233.344

Neste estudo preliminar com um *corpus* público, observou-se algumas limitações quanto à sua qualidade na extração da categoria como por exemplo, o *token* "Ministério da Educação" considerado como Organização mas classificado como Pessoa.

A categoria Outros na Tabela 4.1 equivale à soma de todos as demais categorias para fins de resumo na apresentação destes dados, tais como, Evento, Desconhecido, Abstrato, Coisa.

4.3.1.2 Reconhecimento morfossintática – *POS-Tagger*

Esta subseção apresenta uma análise das ferramentas NLTK e Syntaxnet que realizam a anotação dos *tokens*, por meio de um valor morfossintático. Na Tabela 4.2 são apresentados os resultados da quantidade de *tokens* encontrados, conforme a distribuição das categorias morfossintáticas utilizando as ferramentas NLTK e Syntaxnet.

Assumindo que são utilizados *corpora* diferentes para cada ferramenta, a variabilidade da extração das categorias pode ser observada.

A Syntaxnet, por exemplo, extrai a categoria com "Nome Próprio" de um total de

Tabela 4.2: Resultados Comparativos de extração com NLTK e Syntaxnet

Categoria	**Ferramentas**	
	NLTK	**Syntaxnet**
Adjetivo	196.467	103.283
Substantivo	2.521.743	400.867
Numérico	178.302	398.380
Nome Próprio	-	1.206.436
Pronome	106.382	-
Verbo	295.585	152.545
Outros	2.070.990	1.062.982

1.206.436 *tokens* e que não teve extração pela NLTK, diferentemente da categoria "Pronome" que teve um total de 106.382 *tokens* extraídos utilizando NLTK e nenhum com a Syntaxnet.

4.3.2 Avaliação dos resultados quantitativos utilizando OpenNLP

Após realização do processo de treinamento, utilizando da ferramenta OpenNLP, o modelo passou a ser utilizado para realizar a extração das entidades nomeadas de 470 arquivos do DOU, visando extrair categorias e termos candidatos de acordo com as características textuais encontradas com a correlação textual do *corpus* treinado.

A compreensão dos fenômenos através da coleta de dados numéricos e a avaliação dos resultados da construção do modelo de treinamento é um importante processo na extração da informação. O resultado da extração obtida com a ferramenta OpenNLP é mostrado na Tabela 4.4

Tabela 4.3: Quantidade de Entidades-chaves do arquivo de Treino Atual

Categoria	**Quantidade**
DATA	1.202.600
EVENTO	73.807
FUNCAO	205.090
LUGAR	164.518
LEI	46.862
PROCESSO	10.805
NUMERO	200.523
ORGANIZACAO	457.133
PESSOA	235.539
VALOR-MONETARIO	8.581

O preparo das informações para extração das entidades nomeadas iniciou-se na identificação das sentenças dos diários para posterior organização. Ato contínuo, a partir de cada axioma, para cada diário, foi submetido o modelo construído para, em seguida, submeter o processo a um longo período de treino, resultando na extração de 2.605.458 entidades classi-

ficadas a partir dos 470 diários conforme mostrado na Tabela 4.3. Esse resultado de extração pode ser melhor ilustrado em um gráfico como mostrado na Figura 4.4, onde destacam-se os resultados como Data, Organização e Pessoa na extração de categorias

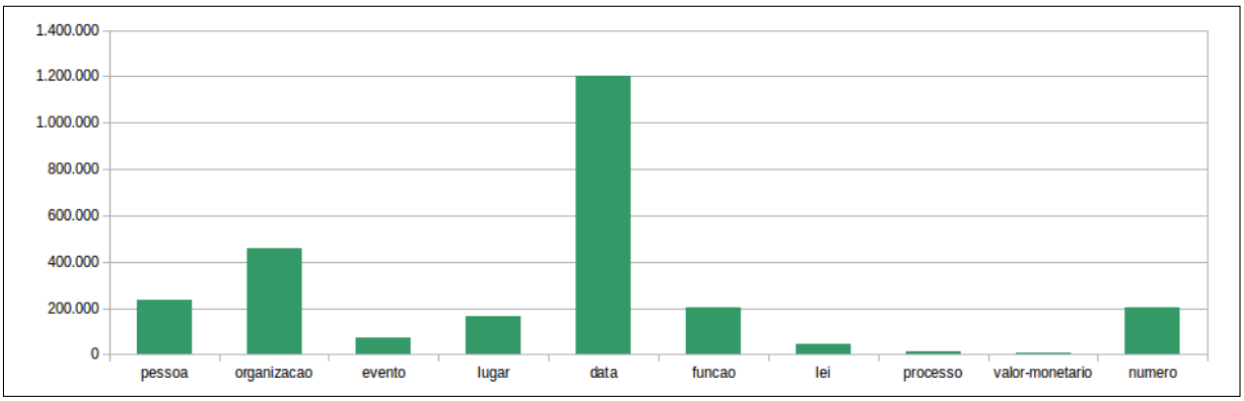


Figura 4.4: Quantidade total de *tokens* de todos os diários

Para compreender melhor a análise dos resultados, as avaliações foram divididas nas 3 seções de extras, que compreendem os diários das três seções, que eventualmente são publicados como urgência. Assim, a seguir os resultados serão analisados por cada seção, mostrando o volume de informações que foram extraídas.

4.3.3 Avaliação da Seção 1

A seção 1, descreve os diários oficiais publicados que continham informações de atos da administração pública conforme visto no item 3.1.1. O resultado quantitativo é destacado na Figura 4.5 onde podemos evidenciar que a extração de entidades como pessoa, organização, data, lugar e função tiveram resultados mais expressivos

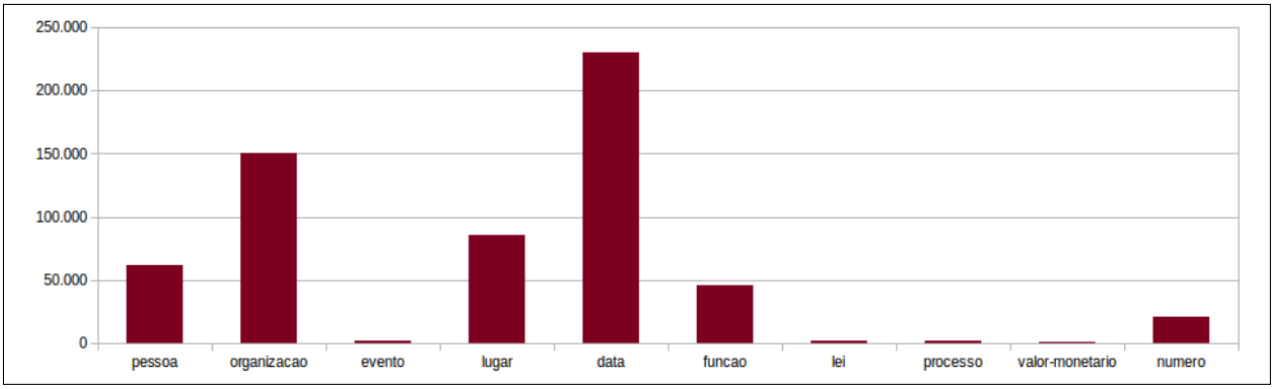


Figura 4.5: Quantidade total de sentenças de *tokens* da Seção 1

Do gráfico da Figura 4.5, conclui-se que os resultados são mais numerosos para algumas categorias, tais como pessoa com 61.635 extrações, organização com 150.557, lugar com 85.134, data com 230.351 e função com 45.110.

4.3.4 Avaliação da Seção 2

A seção 2 trata dos diários oficiais publicados que continham informações de servidores da administração pública conforme visto no item 3.1.1. Os resultados da extração na Seção 2 são mostrados na Figura 4.6

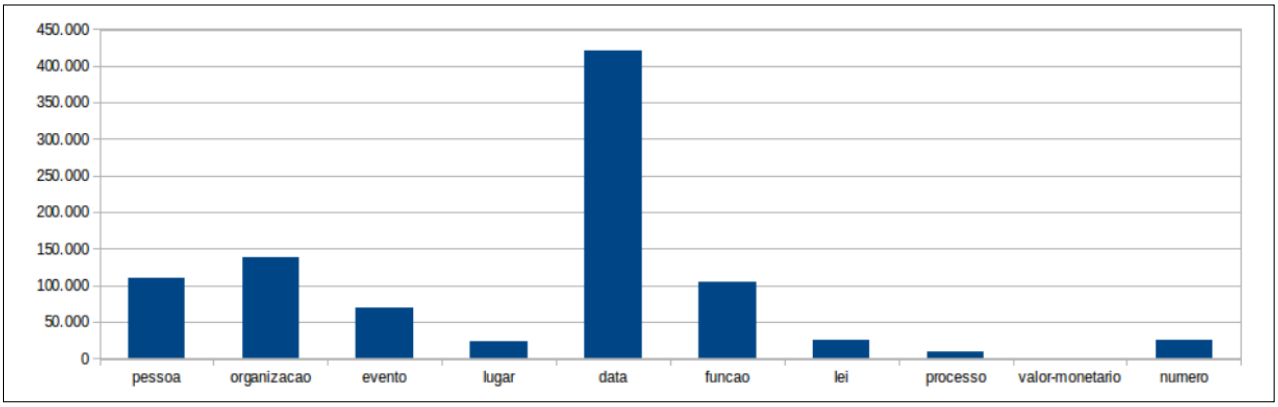


Figura 4.6: Quantidade total de sentenças de *tokens* da Seção 2

O destaque da extração são pessoa com 109.728 extrações, organizacao com 138.758 extrações, data com 420.870 extrações e funcao com 104.727. Mas neste caso pode-se destacar ainda evento com 69.925 extrações.

4.3.5 Avaliação da Seção 3

A avaliação da seção 3, que são diários publicados contendo informação ações de licitações e tomada de preços da administração pública conforme o item 3.1.1, apresentou os resultados mostrados na Figura 4.7.

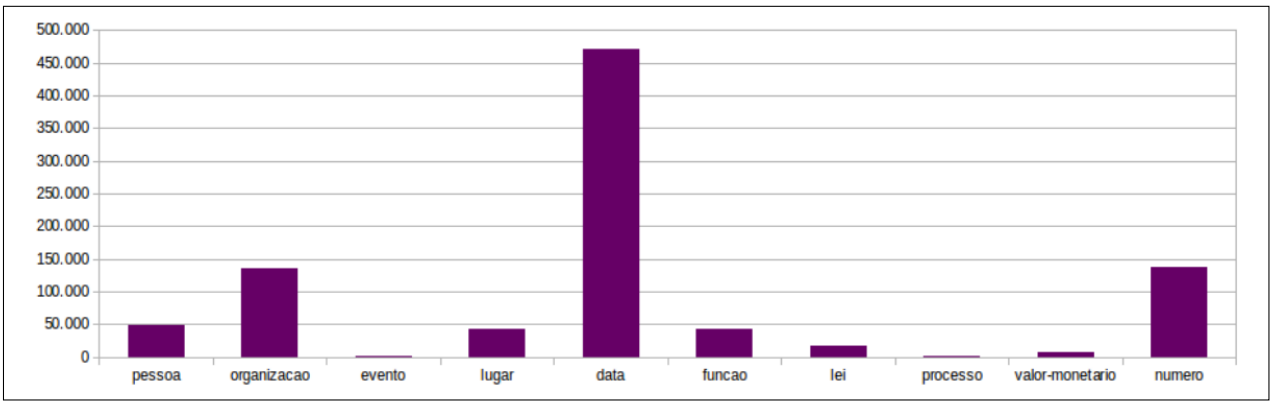


Figura 4.7: Quantidade total de sentenças de *tokens* da Seção 3

As informações coletadas e que merecem destaque são pessoa com 49.254 extrações, organização com 135.672 extrações, data com 471.041 extrações e nessa seção numero se destacou com 137.327 extrações.

4.3.6 Avaliação das Seções Extras

Para concluir os experimentos de extração utilizou-se alguns diários que são publicados como extras, obtendo-se os resultados mostrados na Figura 4.8

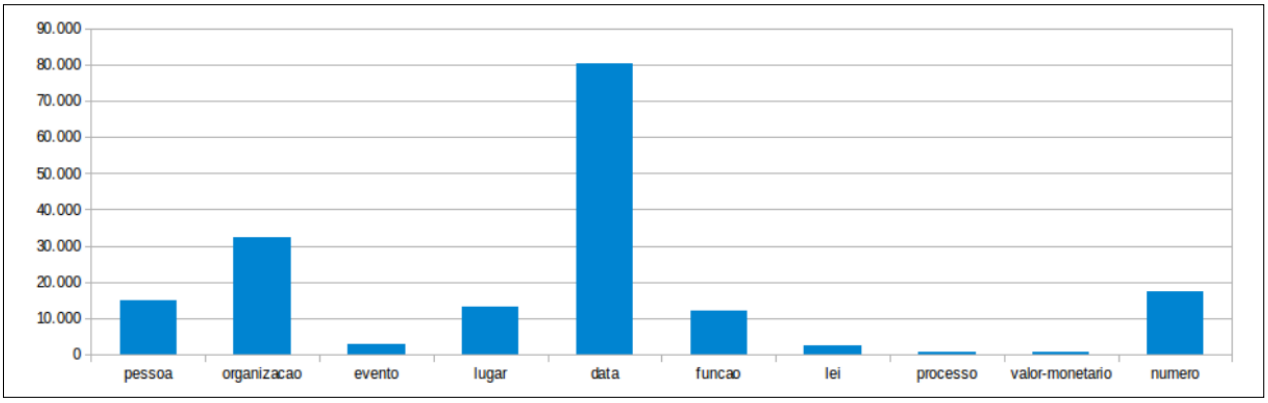


Figura 4.8: Quantidade total de sentenças de *tokens* de Seções Extra

Os gráficos das quatro seções da extração são destaques na quantidade de entidades que são classificadas como datas, seguidas pela classificação de organizacao e de pessoa. O DOU-*Corpus* é um modelo construído para atender todas as seções do DOU. As características das seções já foram mencionadas no Capítulo 3, na Sub-seção 3.1.1, em que foram elencados os seus diferentes conteúdos. Todavia, destaca-se que todas as seções têm datas, organizações e pessoas mencionadas nos diários.

Outra interface construída neste trabalho, que ajudou na análise dos resultados, mostra a totalidade *tokens* extraídos sobre uma consulta de um único Diário do seu quantitativo de *tokens*, como por exemplo no arquivo do DOU, *2017 11 14 assinado do2*. A saída é um resultado de uma lista da quantidade de *tokens* extraídos conforme mostrado na Tabela 4.4

Tabela 4.4: Quantidade de Entidades-chaves do arquivo ***"2017 11 14 assinado do2"***

Categoria	**Quantidade**
DATA	2.462
EVENTO	418
FUNCAO	416
LUGAR	184
LEI	161
PROCESSO	54
NUMERO	172
ORGANIZACAO	952
PESSOA	639

4.4 Resultados Qualitativos

Vale destacar que a avaliação, da qualidade extraída, é um dos aspectos de análise da precisão que consegue-se extrair. E, neste trabalho, a qualidade das informações extraídas

passa pela conversão da informação e na avaliação da extração no processo de treino de um *corpus*.

O processo de converter *pdf* para *JSON* pode ter resultados inconsistentes e a informação não ser corretamente extraída do *pdf*, como por exemplo, extração do nome: "Carlos Oliveira Lima"é extraído como "CarlosOliveiraLima", assim um modelo treinado não conseguiria reconhecer a pessoa. Isso pode comprometer a qualidade da saída das informações da extração. Então a inconsistências ainda precisam ser melhoradas no conversor.

Já a qualidade do *corpus* é dada pelos dados, etiquetagem e quantidade de dados etiquetados. O resultado da avaliação é composto por três indicadores: precisão, cobertura e medida F. O processo de validação e avaliação da qualidade do DOU-*Corpus* apresentou os resultados mostrados na Tabela 4.5.

Tabela 4.5: Métrica Média do *DOU-Corpus*

Corpus	*Precisão*	*Cobertura*	*Frenquência*
DOU-Corpus	**95.30%**	***60.70%***	**44.50%**

Para validar a avaliação da qualidade da extração a OpenNLP dispõe na sua biblioteca de um recurso que processa a avaliação da extração. A classe *TokenNameFinderEvaluator* espera um modelo treinado e uma sentença com classificadores anotados. A saída é mensurada com um cálculo da acurácia da qualidade das informações extraídas.

Para realizar esse cálculo foi construído um arquivo treino para testes. Assim a avaliação, da precisão, da qualidade e da frequência, é calculada de acordo com o arquivo de treino construído. A precisão mede quanto um *token* consegue ser preciso em relação ao *corpus*. Já a Qualidade indica a qualidade geral da extração de *tokens* em relação ao DOU-*Corpus* com o *corpus* do teste. E a Frequência por sua vez mede a assiduidade, i.e., a constância do *token* no texto.

4.5 Protótipo do Sistema

O foco desta seção é proporcionar uma visão sobre as informações extraídas dos diários, utilizando PLN, por meio de um protótipo que contém algumas interfaces que facilitam o entendimento dos resultados de extração.

4.5.1 Construção de Micro Serviços

Para facilitar uma futura escalabilidade do sistema, foram projetados serviços que pudessem ser autônomos, com poucas responsabilidades, comumente chamados de *microservices*. A biblioteca do *Spring Boot* disponibiliza um conjunto de técnicas para desenvolver esse tipo de solução.

Para este trabalho alguns serviços foram desenvolvidos e utilizados,tais como:

- **Treinar *Corpus***: serviço responsável por treinar um *corpus*, tendo como informação um arquivo de Treino
- **Extrair Entidades**:serviço que recebe ou um conjunto de arquivo ou um unico arquivo para extrair Entidades Nomeadas.
- **Buscar diários**: serviço que recebe informações como intervalo de dado, seção.
- **Buscar um Único Diário**: serviço que recebe exatamente o nome chamado de identificador.

A estratégia de adoção de uma solução que utilize conceito de micro serviços torna o sistema mais flexível e escalável, por ter baixo acoplamento entre os serviços. E foi pensando em evoluções futuras que esse tipo de arquitetura foi escolhida.

4.5.2 Apresentação do Protótipo

A interface de aplicação permitem o uso de filtros nas consultas que podem ser realizadas para se obter determinados resultados. Por exemplo, na Figura 4.9, é mostrada uma tela ilustrando o uso de filtros para realizar consultas restritivas, passando informação de um intervalo de data, seção e se desejar por duas categorias. Observa-se que na mesma consulta é possível recuperar um único Diário com todas as suas sentenças, passando o nome exato do Diário.

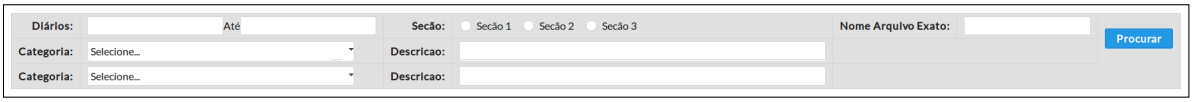


Figura 4.9: Exemplo do uso de filtros para consulta nos diários

A saída do processo de consulta é um resultado com uma lista de sentenças de cada diários, visualizando as entidades, identificando-as de acordo com a categoria e cor, e quantidade de entidades extraídas da sentença. A Figura 4.10 é uma amostra do resultado, de uma sentença com as entidades identificadas e da quantidade de entidades extraídas na sentença.

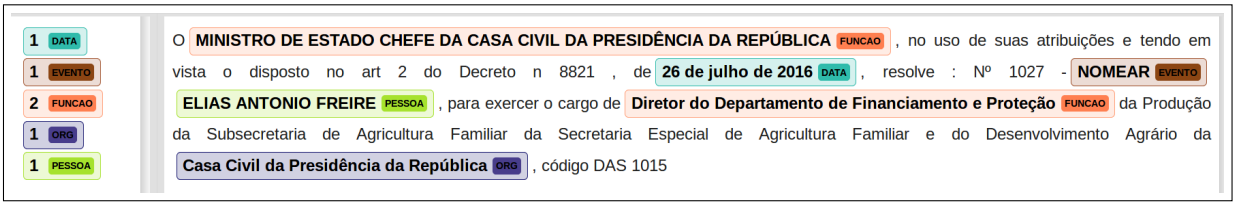


Figura 4.10: Resultado da consulta dos diários organizados por sentenças marcadas

O resultado traz todos os diários, detalhando-os em uma lista de sentenças, onde em cada sentença os *tokens* são marcados em um tom de cor de acordo com a categoria. No lado

esquerdo (Fig 4.10) pode-se observar a informação de total de *tokens* extraídos sentença. Uma outra interface de aplicação foi desenvolvida a fim de permitir a consulta da quantidade de *tokens* extraídos de um período de publicações. A Figura 4.11 ilustra uma interface onde são informados um período e a seção dos diários.

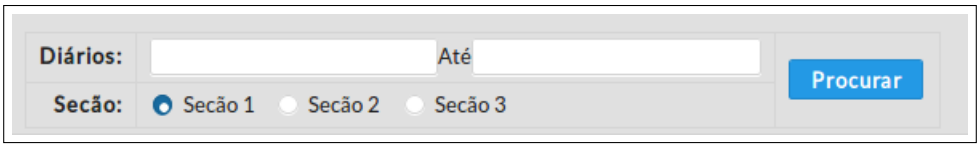


Figura 4.11: Filtro de pesquisa para trazer a quantidade de *tokens* extraídos

O resultado da pesquisa dos diários em termos de quantidade de *toknes* está ilustrado na Figura 4.12 com os dados organizados de forma ordenada dos *tokens* extraídos de um Diário.

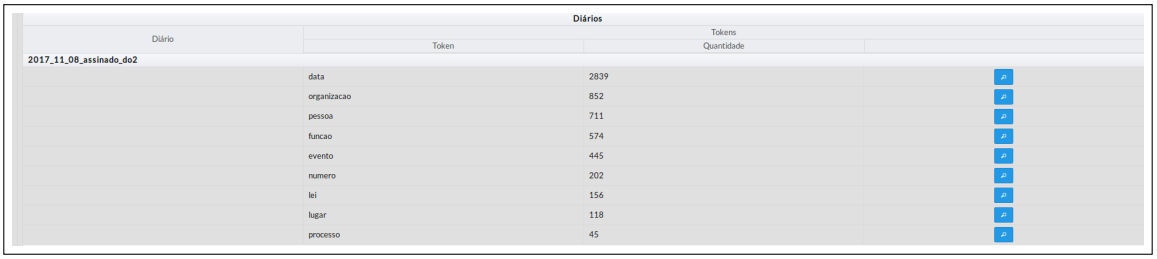


Figura 4.12: Interface com o resultado do total de *tokens* extraídos organizados por Diário

Além de consultar a totalização da informação pode-se detalhar os *tokens* extraídos. Esse resultado é ilustrado na Figura 4.13 visualizando todos os *tokens* que foram extraídos de um Diário.

Figura 4.13: Resultado de um detalhe de *tokens*

4.6 Avaliações e Discussões

Esta seção apresenta uma discussão e adianta algumas conclusões preliminares em relação ao processo de construção de uma ferramenta para extrair entidades nomeadas.

Primeiro é ressaltado que o processo de extração de informação requer um planejamento cuidadoso e um desenho das atividades de como e o que será extraído, não havendo uma solução geral para todas as aplicações.

Em segundo lugar, é preciso destacar que o ciclo de tarefas no processo de extração de informação do DOU (coleta, treino, extração e avaliação), tem um inter-relacionamento que sempre poderá ser melhorado. Nesse sentido, a construção de um corpus específico sempre tem o que melhorar.

Principalmente por que o conjunto de dados da fonte está em constante mudança, pois todos os dias são publicados diferentes nomes, organizações, lugares, etc. Pode-se citar por exemplo, a situação neste final de 2018 onde o Governo de Transição cogita uma redução do número de ministérios, com a fusão de alguns existentes e a criação de alguns outros. Nesse contexto, tendem a surgir novos nomes de organizações, novos cargos, novos nomes de pessoas, etc, resultando em dados com informações novas que o corpus atual talvez não consiga extrair.

Portanto, observa-se que o desenvolvimento de uma aplicação que resolva todos problemas, não é uma tarefa trivial. Neste trabalho o começo buscou-se construir uma solução específica que atenda e consiga extrair as informações necessárias do DOU. Entretanto, entende-se que essa construção é um processo que pode ser afetado pelas constantes mudanças, por exemplo pela necessidade de mapeamento de novas categorias, adição de mais informações no corpus de treino, e assim por diante.

4.7 Síntese do Capítulo

Este capítulo aborda a construção do modelo proposto para a extração de entidades nomeadas dos diários Oficiais da União, incluindo a construção de um corpus específico, chamado de DOU-*Corpus*.

O capítulo discute o desenvolvimento da solução proposta desde a criação do modelo e o processo de extração das entidades nomeadas do DOU. Além disso, é feita uma avaliação sobre a quantidade e a qualidade das informações extraídas, utilizando uma aplicação protótipo com algumas telas para visualização dos resultados de extração obtidos com a solução adotada neste trabalho.

Capítulo 5

Conclusão e Trabalhos Futuros

Este trabalho surgiu com o intuito de desenvolver um modelo e uma arquitetura baseada em técnicas de PLN, capaz de reconhecer entidade nomeadas extraídas do DOU, uma importante fonte de informação em língua portuguesa.

Para atingir o objetivo foram realizados diversos estudos e experimentos, envolvendo atividades de levantamento bibliográfico, de coleta e conversão nos Diários, de construção de modelos e ferramentas que realizam extração de entidades nomeadas.

Após esses estudos e experimentos preliminares, foi selecionada a ferramenta de extração OpenNLP que se que destaca por ter uma documentação clara quanto a construção de novos modelos para extrair entidades nomeadas além de ter sido utilizada anteriormente em alguns estudos envolvendo o Português-brasileiro.

Utilizando a ferramenta OpenNLP foram identificadas as principais categorias que poderiam ser classificadas de um Diário: data, evento, funcao, lei, numero, organizacao, pessoa e processo e foi desenvolvido um modelo específico de *corpus*, o DOU-*Corpus*, a fim de extrair as entidades nomeadas do DOU com resultados quantitativos e qualitativos superiores aos obtidos com outras estratégias.

Enfim, foi proposta uma arquitetura de sistema e implementado um protótipo de validação incorporando tarefas, desde a coleta de dados, de forma manual até a visualização dos resultados da extração por meio de uma interface programável, que permite a customização dos resultados a serem obtidos.

5.1 Trabalhos Futuros

Este trabalho é um primeiro passo para elaborar algo bem mais ambicioso dentro da expectativa de extração de informação. Os resultados já obtidos podem ser utilizados para embasar próximos passos no desenvolvimento de uma ferramenta de extração acrescentando novas funções, como a resolução de correferência, análise de contextos e, análise de relacio-

namentos, entre por exemplo, organização, pessoa, função e evento.

Uma outra vertente que pode aperfeiçoar este trabalho é tentar melhorar a qualidade da extração, utilizando por exemplo, modelos segregados para cada seção e otimizando a quantidade de categorias, dependendo das características do DOU.

O que ainda pode ser melhorado também é quanto ao leiaute de cada Diário. Este trabalho foi realizado baseado em um leiaute que poderá mudar ao longo de tempo.

Sugere-se também uma melhoria do modelo, adicionando novas categorias de extração.

5.2 Publicação decorrente desta pesquisa

- ALLES, Vanderlei J.; GIOZZA, William F.; DE OLIVEIRA ALBURQUERQUE, Robson. *Natural language processing to classify named entities of the Brazilian Union Official Diary*. In: 2018 13th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2018. p. 1-6

Referências Bibliográficas

[Alberti et al. 2017] Alberti, C., Andor, D., Bogatyy, I., Collins, M., Gillick, D., Kong, L., Koo, T., Ma, J., Omernick, M., Petrov, S., et al. (2017). Syntaxnet models for the conll 2017 shared task. *arXiv preprint arXiv:1703.04929*.

[Andor et al. 2016] Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

[Aprosio and Moretti 2016] Aprosio, A. P. and Moretti, G. (2016). Italy goes to stanford: a collection of corenlp modules for italian. *arXiv preprint arXiv:1609.06204*.

[Atienza 1979] Atienza, C. A. (1979). *Documentação jurídica: introdução à análise e indexação de atos legais*. Achiamé.

[Banker 2011] Banker, K. (2011). *MongoDB in action*. Manning Publications Co.

[Bird et al. 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

[Bird and Loper 2004] Bird, S. and Loper, E. (2004). Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

[Borthwick and Grishman 1999] Borthwick, A. and Grishman, R. (1999). *A maximum entropy approach to named entity recognition*. PhD thesis, Citeseer.

[Chiele et al. 2015] Chiele, G. C., Fonseca, E., and Vieira, R. (2015). Geração de modelo para reconhecimento de entidades nomeadas no opennlp. *Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Su (PUCRS)*.

[Chieu and Ng 2002] Chieu, H. L. and Ng, H. T. (2002). Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[Chiticariu et al. 2010] Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., and Vaithyanathan, S. (2010). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1002–1012. Association for Computational Linguistics.

[Chopra et al. 2016] Chopra, D., Joshi, N., and Mathur, I. (2016). *Mastering Natural Language Processing with Python*. Packt Publishing Ltd.

[Das and Behera 2017] Das, K. and Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2):1301–1309.

[de Souza 2008] de Souza, J. N. (2008). *Lógica para ciência da computação*. Elsevier Brasil.

[do Amaral and Vieira 2014] do Amaral, D. O. F. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

[Drury et al. 2017] Drury, B., Fernandes, R., and Lopes, A. d. A. (2017). Bragrinews: Um corpus temporal-causal (português-brasileiro) para a agricultura. *Linguamática*, 9(1):41–54.

[Edwards 2018] Edwards, C. (2018). Deep learning hunts for signals among the noise. `https://cacm.acm.org/magazines/2018/6/228030-deep-learning-hunts-for-signals-among-the-noise/fulltext`. [Online; acessado 15 de Outubro de 2018].

[Finatto et al. 2015] Finatto, M. J. B., Lopes, L., and Silva, A. C. (2015). Processamento de linguagem natural, linguística de corpus e estudos linguísticos: uma parceria bem-sucedida. *Domínios de lingu@ gem. Uberlândia, MG. Vol. 9, n. 5 (dez. 2015), p.[41]-59*.

[Fonseca et al. 2015] Fonseca, E. B., Chiele, G. C., and Vanin, A. A. (2015). Reconhecimento de entidades nomeadas para o portugues usando o opennlp. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2015), s. pp*.

[Gibney 2016] Gibney, E. (2016). Google ai algorithm masters ancient game of go. *Nature News*, 529(7587):445.

[Goodfellow et al. 2016] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

[Grishman and Sundheim 1996] Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

[Gutierrez 2016] Gutierrez, F. (2016). *Pro Spring Boot*. Springer.

[Hamada and Neto 2015] Hamada, L. and Neto, N. (2015). Desambiguação de homógrafos-heterófonos por aprendizado de máquina em português brasileiro (a machine learning approach for homographic heterophone disambiguation in brazilian portuguese). In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 181–190.

[Jordan and Mitchell 2015] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

[Kart 2017] Kart, T. (2017). Tuorial kart. `https://www.tutorialkart.com/opennlp/apache-opennlp-tutorial/`. [Online; acessado 20 de Agosto de 2017].

[Kotsiantis et al. 2007] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24.

[Krochmalski 2014] Krochmalski, J. (2014). *IntelliJ IDEA Essentials*. Packt Publishing Ltd.

[Lustosa 2010] Lustosa, V. G. (2010). O estado da arte em inteligência artificial. *Colabor@-A Revista Digital da CVA-RICESU*, 2(8).

[Manning et al. 2014] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

[Monard and Baranauskas 2003] Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1(1):32.

[Nadeau and Sekine 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[NETO and BONINI 2010] NETO, A. B. and BONINI, C. d. S. B. (2010). Redes neurais artificiais: Apresentação e utilização do algoritmo perceptron em biossistemas/artificial neural networks: Introduction and use of perceptron algorithm in biossystems. *Revista Brasileira de Engenharia de Biossistemas*, 4(2):87–95.

[OpenNLP 2017] OpenNLP (2017). Apache opennlp developer documentation. `http://opennlp.apache.org/docs/1.8.2/manual/opennlp.html`. [Online; acessado 20 de Julho de 2017].

[Ottoni et al. 2016] Ottoni, A. L. C., Nepomuceno, E. G., de Oliveira, M. S., Cordeiro, L. T., and Lamperti, R. D. (2016). Análise da influência da taxa de aprendizado e do fator de

desconto sobre o desempenho dos algoritmos q-learning e sarsa: aplicação do aprendizado por reforço na navegação autônoma. *Revista Brasileira de Computação Aplicada*, 8(2):44–59.

[Perini 2017] Perini, M. A. (2017). *Gramática descritiva do português brasileiro*. Editora Vozes Limitada.

[Porter et al. 2017] Porter, B., Zyl, J., and Lamy, O. (2017). Maven–welcome to apache maven. *Maven. apache. org*.

[Qiu et al. 2016] Qiu, J., Wu, Q., Ding, G., Xu, Y., and Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):67.

[Ribeiro and Medeiros 2016] Ribeiro, P. B. and Medeiros, R. P. (2016). Extração de entidades nomeadas com maximização de entropia (opennlp). *Caderno de Estudos Tecnológicos*, 4(1).

[Russell and Norvig 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[Sellitto 2002] Sellitto, M. A. (2002). Inteligência artificial: uma aplicação em uma indústria de processo contínuo. *Gestão & Produção*, 9(3):363–376.

[Souza and Claro 2014] Souza, E. N. P. and Claro, D. B. (2014). Extração de relações utilizando features diferenciadas para português. *Linguamática*, 6(2):57–65.

[Spell 2015] Spell, T. B. (2015). *Pro Java 8 Programming*. Apress.

[Sun et al. 2007] Sun, C., Lin, L., Wang, X., and Guan, Y. (2007). Using maximum entropy model to extract protein-protein interaction information from biomedical literature. In *International Conference on Intelligent Computing*, pages 730–737. Springer.

[Syntaxnet 2017] Syntaxnet, E. (2017). Syntaxnet: Neural models of syntax. `https://github.com/tensorflow/models/tree/master/research/syntaxnet`. [Online; acessado 10 de Março de 2017].

[VICCARI 1990] VICCARI, R. M. (1990). Inteligência artificial: representação do conhecimento. *Porto Alegre: II/UFRGS*.

[Weber et al. 2015] Weber, C. et al. (2015). Construção de um corpus anotado para classificação de entidades nomeadas utilizando a wikipedia e a dbpedia. *Pontifícia Universidade Católica do Rio Grande do Sul*.

[Xhafa et al. 2017] Xhafa, F., Caballé, S., and Barolli, L. (2017). *Advances on P2P, Parallel, Grid, Cloud and Internet Computing: Proceedings of the 12th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2017)*, volume 13. Springer.

[Zaccara 2012] Zaccara, R. C. C. (2012). *Anotação e classificação automática de entidades nomeadas em notícias esportivas em Português Brasileiro*. PhD thesis, Universidade de São Paulo.