

Aprendizado de Máquina e Redes Neurais para Reconhecimento de Entidades Nomeadas em Portarias Jurídicas via Processamento de Linguagem Natural

Davi Esmeraldo da Silva Albuquerque

Orientador: Eduardo Monteiro de Castro

24 de janeiro de 2025

1 Introdução

2 Objetivo Geral

- Objetivos Específicos
- Impactos Esperados

3 Metodologia

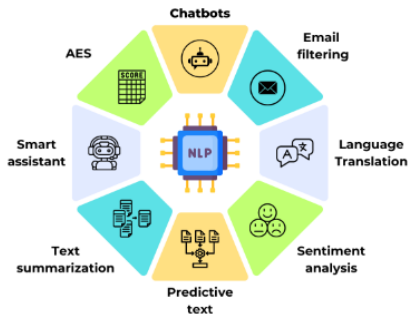
- Coleta de Dados
- Word Embeddings
 - Word2Vec
- Aprendizado Não Supervisionado
- Reconhecimento de Entidades Nomeadas (REN)
 - Redes Neurais Recorrentes Bidirecionais (BiLSTM)

4 Cronograma

5 Referências

Definição

Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial (IA) que busca desenvolver métodos e sistemas capazes de processar (interpretar, compreender e gerar) linguagem humana de forma computacional. (CASELI; NUNES, 2024)



Definição

Entidades Nomeadas são palavras ou frases que representam elementos específicos e bem definidos em um texto.

- Exemplos incluem:
 - **Pessoas:** Albert Einstein, Ada Lovelace
 - **Locais:** Brasília, Monte Everest
 - **Organizações:** ONU, Google
 - **Datas e Horários:** 22 de janeiro de 2025, 15h30

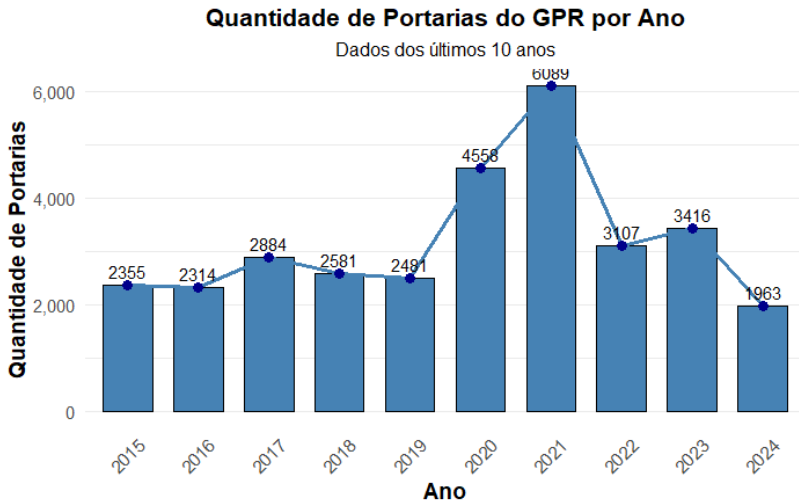
Definição

Reconhecimento de Entidades Nomeadas (NER - Named Entity Recognition) é uma tarefa de PLN que visa identificar e classificar automaticamente Entidades Nomeadas.

- Aplicações incluem:
 - Extração de informações de notícias e documentos.
 - Melhoria em sistemas de busca.
 - Mineração de opiniões.
 - Análise de currículos.

Introdução

- Instituições jurídicas geram uma quantidade significativa de documentos oficiais.



Fonte: Gabinete da Presidência do TJDF

Objetivo Geral

O objetivo deste trabalho é comparar modelos de aprendizado de máquina e redes neurais para realizar as tarefas de agrupamento (clusterização) e Reconhecimento de Entidades Nomeadas (REN) em portarias de 2024 do Gabinete da Presidência do Tribunal de Justiça do Distrito Federal e Territórios (TJDFT).

Objetivos Específicos

- Aplicar e estudar diferentes formas de representar dados textuais em formato vetorial numérico.
- Agrupar portarias jurídicas com base em temáticas similares de modo a organizá-las de forma lógica e eficiente.
- Extrair categorias-chave dos agrupamentos de portarias, como nomes, datas e legislações mencionadas e resolução.

- Organizar e analisar portarias jurídicas com maior eficiência.
- Promover transparência do fluxo de trabalho realizado .
- Promover inovação, impulsionando o uso de IA no setor público.


■ Webscraping no website do TJDF

Raspagem de dados automatizada para extrair os conteúdos das portarias disponíveis no site do Tribunal de Justiça do Distrito Federal e Territórios.

■ Expressões Regulares (Regex)

Conjunto de padrões usados para localizar, extrair ou manipular a seleção de texto de forma eficiente.

Portaria GPR 2355 de 30/12/2015



Poder Judiciário da União
Tribunal de Justiça do Distrito Federal e dos Territórios
Gabinete da Presidência

PORTARIA GPR 2355 DE 30 DE DEZEMBRO DE 2015

O PRESIDENTE DO TRIBUNAL DE JUSTIÇA DO DISTRITO FEDERAL E DOS TERRITÓRIOS, no uso de sua competência legal e tendo em vista o disposto P.A. nº 23.609/2015,

RESOLVE:

Conceder aposentadoria voluntária integral, com fundamento no art. 3º da [Emenda Constitucional 47, de 5 de julho de 2005](#), a servidora MARIA DE FATIMA DE CASTRO, matrícula 307.616, ocupante do cargo de Técnico Judiciário, Área Administrativa, Classe "C", Padrão 13, Nível Intermediário, do Quadro de Pessoal deste Tribunal, com as vantagens previstas no art. 67 da [Lei 8.112/1990](#), c/c o art. 6º da [Lei 9.624/1998](#) e com o inciso II do art. 15 da [Medida Provisória 2.225-45/2001](#); e, no art. 3º da [Lei 8.911/1994](#), c/c a Resolução 19/1994-TJDF e com o art. 15 da [Lei 9.527/1997](#).

Desembargador GETÚLIO DE MORAES OLIVEIRA



Importância

O pré-processamento de dados textuais é essencial para o sucesso de técnicas de inteligência artificial ao viabilizar a comunicação entre humano e máquina. (OLIVEIRA; NASCIMENTO, 2021).

- Objetivos do Pré-processamento
 - Transformar dados brutos em um formato limpo e estruturado.
 - Reduzir a dimensionalidade dos dados para otimizar o desempenho computacional.
 - Facilitar análises computacionais e estatísticas.

Definição

Representação de dados textuais em formas numéricas que algoritmos podem processar como vetores ou matrizes de números. (FREITAS, 2023)

- *Word Embeddings* preservam relações semânticas e sintáticas entre palavras, capturando suas similaridades de maneira eficiente.

Word2Vec, desenvolvido por (MIKOLOV et al., 2013), oferece duas arquiteturas principais para a modelagem de palavras: o **Continuous Bag of Words (CBOW)** e o **Skip-gram**.

O CBOW prevê uma palavra com base no contexto de palavras vizinhas, utilizando a seguinte fórmula:

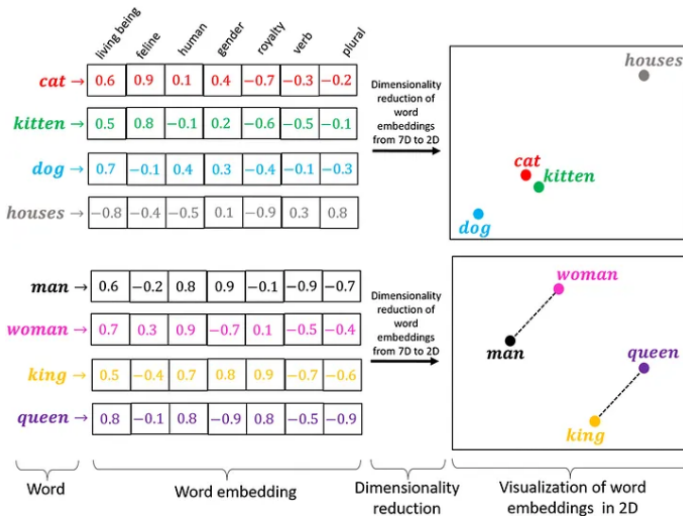
$$P(w_t \mid w_{t-c}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+c}) \quad (1)$$

Por outro lado, o *Skip-gram* prevê o contexto de uma palavra dada a palavra central, representado pela equação:

$$P(w_{t-c}, \dots, w_{t+c} \mid w_t) \quad (2)$$



Metodologia - Word Embeddings



Definição

O aprendizado de máquina não supervisionado é uma técnica em que o modelo trabalha com dados sem rótulos (labels), buscando identificar padrões ou estruturas ocultas. Os principais métodos incluem agrupamento e redução de dimensionalidade.

Metodologia - Aprendizado Não Supervisionado

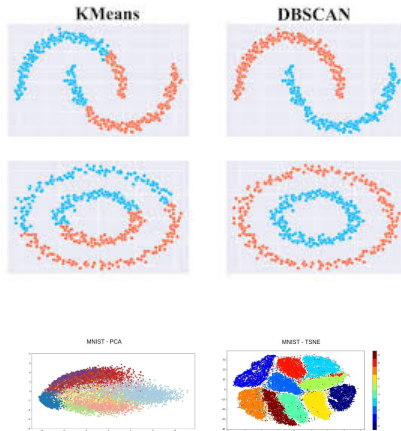
- **Kmeans**

Clusters definidos pelo centróide.

- **DBSCAN**

Clusters definidos com base em densidade.

- **t-SNE**: Técnica de redução de dimensionalidade para visualização de dados.



Metodologia - Reconhecimento de Entidades Nomeadas

PORTARIA **GPR** **MISC** 1956 DE 23 DE DEZEMBRO DE 2024

O PRESIDENTE DO **TRIBUNAL DE JUSTIÇA** **ORG** DO **DISTRITO FEDERAL** **LOC** E DOS **TERRITÓRIOS** **ORG**, no uso de sua competência legal e tendo em vista o disposto no **Processo** **MISC** SEI 0039781/2024,

RESOLVE:

Conceder **PER** aposentadoria voluntária à servidora **Alexandra de Oliveira** **PER**, matrícula 308.479, ocupante do cargo de **Técnico Judiciário** **MISC**, **Área Administrativa** **LOC**, **Classe C** **MISC**, **Padrão** **MISC** 13, do **Quadro de Pessoal** **MISC** deste **Tribunal de Justiça** **ORG**, com fundamento no art. 20, caput, **§§ 2º** **LOC**, **inciso II** **PER**, e 3º, inciso II, da **Emenda Constitucional** **MISC** 103/2019, observado o disposto no art. 40, § 16, da **Constituição Federal** **LOC**, incluído pela **Emenda Constitucional** **MISC** 20/1998, **c/** **MISC** o art. 3º, inciso II e **§** **ORG** § 1º, 2º, 3º, 5º e 6º, da **Lei 12.618/2012** **MISC**, alterada pela **Lei** **MISC** 14.463/2022, com proventos calculados e reajustados na forma do art. 26, caput, **§§ 1º** **LOC**, 3º, inciso I, e 7º, da **Emenda Constitucional** **MISC** 103/2019.

Desembargador WALDIR LEÔNICIO JÚNIOR **MISC** Presidente



Metodologia - Reconhecimento de Entidades Nomeadas

PORTARIA **GPR** **MISC** 1956 DE 23 DE DEZEMBRO DE 2024

O PRESIDENTE DO **TRIBUNAL DE JUSTIÇA** **ORG** DO **DISTRITO FEDERAL** **LOC** E DOS **TERRITÓRIOS** **ORG**, no uso de sua competência legal e tendo em vista o disposto no **Processo** **MISC** SEI 0039781/2024,

RESOLVE:

Conceder **PER** aposentadoria voluntária à servidora **Alexandra de Oliveira** **PER**, matrícula 308.479, ocupante do cargo de **Técnico Judiciário** **MISC**, **Área Administrativa** **LOC**, **Classe C** **MISC**, **Padrão** **MISC** 13, do **Quadro de Pessoal** **MISC** deste **Tribunal de Justiça** **ORG**, com fundamento no art. 20, caput, **§§ 2º** **LOC**, **inciso II** **PER**, e 3º, inciso II, da **Emenda Constitucional** **MISC** 103/2019, observado o disposto no art. 40, § 16, da **Constituição Federal** **LOC**, incluído pela **Emenda Constitucional** **MISC** 20/1998, **c/** **MISC** o art. 3º, inciso II e **§** **ORG** § 1º, 2º, 3º, 5º e 6º, da **Lei 12.618/2012** **MISC**, alterada pela **Lei** **MISC** 14.463/2022, com proventos calculados e reajustados na forma do art. 26, caput, **§§ 1º** **LOC**, 3º, inciso I, e 7º, da **Emenda Constitucional** **MISC** 103/2019.

Desembargador WALDIR LEÔNICIO JÚNIOR **MISC** Presidente

■ Desafios:

- Heterogeneidade da linguagem jurídica
- A predominância de corpora voltados à língua inglesa



Definição

As Redes Neurais Recorrentes Bidirecionais (*BiLSTM*) são projetadas para processar sequências de dados em ambas as direções, capturando informações contextuais tanto do passado quanto do futuro (COSTA, 2023).

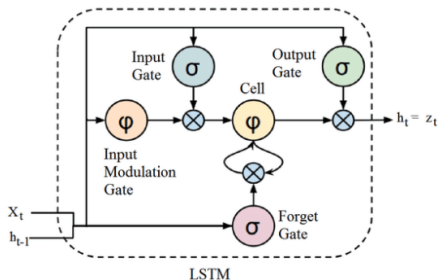
- No BiLSTM, duas redes LSTM independentes são treinadas:
 - A LSTM direta ($\overrightarrow{h_t}$), que processa a sequência do início para o fim.
 - A LSTM reversa ($\overleftarrow{h_t}$), que processa a sequência na direção oposta.
 - A saída do BiLSTM para cada posição t é dada pela concatenação dos estados ocultos das duas direções:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

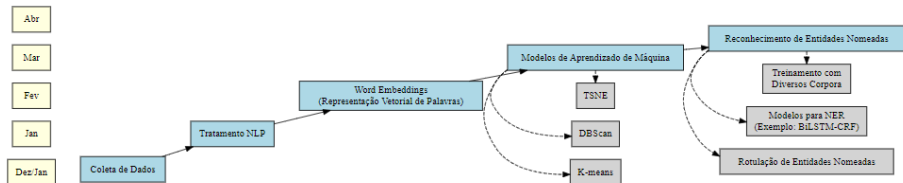


Metodologia - Long Short Term Memory (LSTM)

- LSTM é capaz de reter informações úteis por períodos prolongados.
- Permite um controle preciso sobre quais dados são armazenados ou descartados pelas portas de entrada, saída e esquecimento.



Síntese Metodologia - Cronograma



Cronograma

As atividades a serem desenvolvidas durante o Trabalho de Conclusão de Curso são:

Tabela: Cronograma do TCC 1

Atividades	2/2024				
	Out	Nov	Dez	Jan	Fev
Escolha do tema					
Levantamento de bibliografias relacionadas					
Desenvolvimento da proposta de projeto					
Entrega da proposta ao Orientador					
Revisão da proposta					
Coleta e Raspagem dos Dados					
Elaboração da apresentação da proposta					
Apresentação oral da proposta					
Pré-processamento Linguístico					
Representação Vetorial (Word Embeddings)					
Aplicação de Modelos para Agrupamento					




Cronograma


As atividades a serem desenvolvidas durante o Trabalho de Conclusão de Curso são:


Tabela: Cronograma do TCC 2


Atividades	1/2025				
	Mar	Abr	Mai	Jun	Jul
Aplicação de Modelos REN					
Avaliação de Desempenho dos Modelos					
Elaboração do relatório final					
Entrega do relatório final ao Professor Orientador					
Revisão do relatório final					
Elaboração da apresentação da proposta					
Entrega do relatório final para a banca					
Apresentação oral da proposta					
Correção do relatório final					





 BARROS, F. M. d. C. et al. Processamento de linguagem natural como ferramenta de suporte em documentos jurídicos: uma revisão sistemática. *Revista de Casos e Consultoria*, v. 15, n. 1, p. e36701, ago. 2024. Disponível em: <<https://periodicos.ufrn.br/casoseconsultoria/article/view/36701>>.

 CASELI, H. d. M.; NUNES, M. d. G. V. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – 2ª Edição*. [S.l.]: BPLN, São Carlos, 2024. Disponível em: <<https://brasileiraspln.com/livro-pln/2a-edicao/>>. ISBN 978-65-00-95750-1.

 COSTA, R. P. *Reconhecimento de entidades nomeadas em textos informais no domínio legislativo*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Goiás, Goiânia, 2023. 70 f.

 FREITAS, L. J. G. *Clusterização de textos aplicada ao tratamento de dados jurídicos desbalanceados*. Dissertação (Dissertação (Mestrado em Estatística)) — Universidade de Brasília, Departamento de Estatística, Brasília, 2023. Disponível em: <<http://repositorio.unb.br/handle/10482/48841>>.

 GARCIA, G. C. *Reconhecimento de Entidades Nomeadas na base de notificações de eventos adversos e queixas técnicas de dispositivos médicos no Brasil*. Dissertação (Dissertação (Mestrado Profissional em Computação Aplicada)) — Universidade de Brasília, Brasília, ago 2021. Data de defesa: 31 de agosto de 2021. Disponível em: <<http://repositorio.unb.br/handle/10482/42718>>.

 MIKOLOV, T. et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. Disponível em: <<https://arxiv.org/abs/1301.3781>>.





OLIVEIRA, R.; NASCIMENTO, E. G. S. Clustering by similarity of brazilian legal documents using natural language processing approaches.

In: _____. [S.l.: s.n.], 2021. ISBN 978-1-83969-887-3.

(BARROS et al., 2024) (GARCIA, 2021)