**UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO**

AUGUSTO AMARAL PEREIRA

ORIENTADOR: MEUS ANCESTRAIS

USO DE ALGORITMOS DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA CRIAÇÃO DE ÍNDICES INTELIGENTES

Campos dos Goytacazes – Rio de Janeiro

2023

## CAPÍTULO I: INTRODUÇÃO

O avanço contínuo da tecnologia digital nas últimas décadas trouxe consigo uma mudança notável na maneira como as informações são armazenadas e recuperadas. A transição de documentos físicos para arquivos digitais resultou em enormes volumes de dados, levando à necessidade de soluções mais avançadas para busca e recuperação de informações. No meio acadêmico, essa necessidade é particularmente acentuada, pois a produção e o acesso a artigos científicos, teses e outros materiais acadêmicos são cruciais para o avanço da pesquisa e do aprendizado.

O Processamento de Linguagem Natural (PLN) emergiu como uma solução promissora para lidar com esses desafios de informação. \*\*Originário da interseção de várias disciplinas, incluindo ciência da computação, inteligência artificial e linguística\*\*, o PLN busca permitir que as máquinas entendam e respondam à linguagem humana de forma tão eficaz quanto os humanos. Os primeiros avanços no campo do PLN \*\*surgiram na década de 1950\*\*, \*\*com programas de computador simples que podiam imitar conversas humanas\*\*. Desde então, o campo do PLN evoluiu rapidamente, permitindo avanços significativos na \*\*tradução automática de idiomas\*\*, \*\*reconhecimento de fala\*\*, \*\*análise de sentimentos\*\* e \*\*outras aplicações\*\*.

Os índices inteligentes, uma aplicação das aplicações do PLN, \*\*transformaram a busca e a recuperação de informações, permitindo uma maneira mais eficiente de localizar e acessar informações em grandes volumes de dados\*\*. Um índice inteligente não apenas lista palavras e frases em um documento, mas também entende o contexto e a relevância do conteúdo indexado. Essa compreensão é possível graças ao PLN, que permite que o computador entenda e interprete a linguagem humana.

A aplicação dessas tecnologias em um servidor de arquivos, como o proposto para a UENF, tem o potencial de revolucionar a maneira como estudantes, professores e pesquisadores acessam e interagem com materiais acadêmicos. Com um índice inteligente gerado por PLN, um servidor de arquivos pode oferecer buscas mais eficientes e precisas, reduzindo o tempo gasto na localização de informações e melhorando a acessibilidade dos recursos acadêmicos. Esta é a motivação por trás do trabalho proposto - usar a tecnologia para melhorar o acesso e a eficiência da pesquisa acadêmica.

### Problemática

A digitalização crescente de documentos acadêmicos apresenta desafios significativos para as instituições de ensino superior, especialmente no que diz respeito à busca e recuperação eficientes de informações. Atualmente, a UENF enfrenta dificuldades na localização de artigos científicos, trabalhos de alunos e outros arquivos em PDF usados para pesquisa e estudo, o que resulta em tempo e esforço consideráveis por parte de estudantes, professores e pesquisadores.

### Hipótese

Acreditamos que a implementação de um servidor de arquivos inteligente, utilizando técnicas de Processamento de Linguagem Natural (PLN) para gerar um índice inteligente, pode melhorar significativamente a eficiência na busca e recuperação de documentos na UENF.

### Objetivos

#### Objetivo Geral

O objetivo geral deste trabalho é desenvolver um servidor de arquivos inteligente para a UENF, que permita uma busca eficiente e precisa dentro de documentos em PDF.

#### Objetivos específicos

Especificamente, pretendemos implementar e testar algoritmos de PLN para a extração e indexação de conteúdo a partir de arquivos em PDF. Além disso, buscaremos desenvolver uma interface de usuário amigável e intuitiva para facilitar a busca de documentos. Outro objetivo específico é construir um sistema de banco de dados robusto para armazenar os índices gerados e facilitar consultas futuras.

### Justificativas

A necessidade de melhorar a acessibilidade e eficiência na busca de documentos acadêmicos justifica este trabalho. A criação de um servidor de arquivos com capacidade de indexação inteligente pode trazer benefícios significativos, tornando os recursos acadêmicos mais facilmente acessíveis para a comunidade da UENF.

### Método

Para atingir os objetivos propostos, este trabalho adotará uma abordagem de desenvolvimento de software em várias etapas. Inicialmente, a infraestrutura do servidor de arquivos será configurada. Em seguida, os algoritmos de PLN serão implementados para extração e indexação de conteúdo. O próximo passo envolverá a criação de uma interface de usuário para permitir buscas eficientes. Finalmente, o sistema será testado para verificar sua eficiência e eficácia na busca e recuperação de documentos. Ao longo do processo, serão adotadas práticas de desenvolvimento de software ágeis para permitir ajustes e melhorias contínuas na solução.

### Organização

O trabalho está organizado e divido em 6 capítulos, sendo o primeiro capitulo uma introdução ao tema, o segundo capitulo uma revisão bibliográfica sobre o tema, o terceiro capitulo uma descrição da metodologia utilizada, o quarto capitulo uma descrição das ferramentas e tecnologias utilizadas, o quinto capitulo uma descrição dos resultados obtidos e discussões sobre o resultado e o sexto capitulo uma conclusão sobre o trabalho.

## CAPÍTULO II: Referencial Teórico

A intersecção de linguística, ciência da computação e inteligência artificial deu **origem ao Processamento de Linguagem Natural (PLN)**, uma disciplina fundamental para o desenvolvimento do nosso servidor de arquivos inteligente. Este capítulo apresenta um panorama da evolução histórica do PLN, demonstrando como ele se expandiu de suas origens modestas até se tornar uma parte integral da tecnologia da informação contemporânea.

A seção inicial, dedicada ao contexto histórico do PLN, traça sua origem e evolução ao longo do tempo, mostrando como o campo tem avançado em paralelo ao desenvolvimento da computação e da inteligência artificial. Ao entendermos o caminho trilhado até aqui, poderemos apreciar melhor as possibilidades que o PLN nos oferece hoje, bem como suas aplicações práticas, que serão discutidas ao longo da seção.

Em seguida, apresentaremos as principais técnicas de PLN que estão sendo utilizadas no projeto. Discutiremos a normalização, a remoção de numerais, a remoção de stopwords e o stemming, que são etapas fundamentais na preparação dos textos para a indexação. Cada uma dessas técnicas desempenha um papel fundamental na análise de textos, que é a base do nosso servidor de arquivos inteligente.

Ao analisarmos o conceito de índice inteligente e sua importância para a recuperação eficiente de informações, estaremos construindo a ponte para a próxima seção, que explora a arquitetura do sistema proposto. Aqui, discutiremos o servidor de arquivos, a interface do usuário e o armazenamento do índice. Por fim, destacaremos as considerações que levaram à escolha das ferramentas utilizadas neste projeto.

Ao final deste capítulo, espera-se que o leitor tenha uma compreensão clara das técnicas de PLN, da estrutura do sistema e das motivações que orientaram nossas escolhas ao longo deste projeto.

### Contexto Histórico do PLN (Processamento de Linguagem Natural)

Antes de aprofundarmos nas técnicas específicas de PLN aplicadas no presente projeto, é importante entender o contexto histórico e a evolução da própria disciplina de Processamento de Linguagem Natural. O PLN surgiu e cresceu a partir de uma convergência de campos, como ciência da computação, inteligência artificial e linguística. Esta seção fornecerá uma visão geral da origem do PLN, de suas primeiras implementações a seus usos atuais mais sofisticados.

#### Origem e evolução do PLN

#### Aplicações do PLN

### Técnicas de PLN utilizadas no projeto

#### Normalização

##### O que é normalização

##### Metodologia de normalização e algoritmos utilizados

##### Exemplos de código e scripts utilizados

#### Remoção de Numerais

##### O que é remoção de numerais

##### Metodologia de remoção de numerais e algoritmos utilizados

##### Exemplos de código e scripts utilizados

#### Remoção de Stopwords

##### O que é remoção de stopwords

##### Metodologia de remoção de stopwords e algoritmos utilizados

##### Exemplos de código e scripts utilizados

#### Stemming

##### O que é stemming

##### Metodologia de stemming e algoritmos utilizados

##### Exemplos de código e scripts utilizados

### Índice inteligente

#### TF-IDF

##### Porque TF-IDF

##### Como funciona o TF-IDF

##### Algoritmo TF-IDF

#### Naive Bayes Multinomial

##### Porque Naive Bayes Multinomial

##### Como funciona o Naive Bayes Multinomial

##### Algoritmo Naive Bayes Multinomial

### Arquitetura do Sistema

#### Servidor de arquivos

#### Interface do usuário

#### Armazenamento do índice

### Considerações sobre a escolha das ferramentas

## Capítulo 3 - Metodologia

### Introdução: Uma breve introdução sobre o que será abordado neste capítulo.

### Abordagem geral: Uma visão geral da abordagem adotada, como eu dividi o processo de desenvolvimento do algoritmo de processamento e indexação, talvez mencionar do resto?.

### Implementação de algoritmos de PLN: Detalhes sobre como foi implementado e textado os algoritmos de PLN para extração e indexação de conteúdo. Detalhar cada técnica de PLN mencionada no Capítulo 2, mencionando os pacotes e bibliotecas utilizados.