**INSTITUTO FEDERAL DO PARANÁ**

**ADRIAN JUAN LARA DA SILVA**

**FLÁVIO RIBEIRO DOS SANTOS**

**ENTRELAÇAMENTO DE BASES DE DADOS CIENTÍFICAS: REDUÇÃO DE DUPLICATAS E PRIORIZAÇÃO POR MÉTRICAS DE QUALIDADE**

**PONTA GROSSA**

**2024**

**ADRIAN JUAN LARA DA SILVA**

**FLÁVIO RIBEIRO DOS SANTOS**

**ENTRELAÇAMENTO DE BASES DE DADOS CIENTÍFICAS: REDUÇÃO DE DUPLICATAS E PRIORIZAÇÃO POR MÉTRICAS DE QUALIDADE**

**INTERLACING SCIENTIFIC DATABASES: REDUCING DUPLICATES AND PRIORITIZING BY QUALITY METRICS**

Pré-Projeto apresentado ao componente curricular Projeto e Desenvolvimento de Sistemas, do curso Técnico em Informática Integrado ao Ensino Médio do Instituto Federal do Paraná (IFPR), como requisito parcial para obtenção do título de Técnico em Informática.

Orientador(a): João Henrique Berssanette

**PONTA GROSSA**

**2024**

**FOLHA DE APROVAÇÃO**

**ADRIAN JUAN LARA DA SILVA**

**FLÁVIO RIBEIRO DOS SANTOS**

**ENTRELAÇAMENTO DE BASES DE DADOS CIENTÍFICAS: REDUÇÃO DE DUPLICATAS E PRIORIZAÇÃO POR MÉTRICAS DE QUALIDADE**

O presente trabalho em nível de graduação foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Nome completo e por extenso do Membro 1 (de acordo com o Currículo Lattes)

Titulação (Especialização, Mestrado, Doutorado)

Nome completo e por extenso da instituição a qual possui vínculo

Nome completo e por extenso do Membro 2 (de acordo com o Currículo Lattes)

Titulação (Especialização, Mestrado, Doutorado)

Nome completo e por extenso da instituição a qual possui vínculo

Nome completo e por extenso do Membro 3 (de acordo com o Currículo Lattes)

Titulação (Especialização, Mestrado, Doutorado)

Nome completo e por extenso da instituição a qual possui vínculo

Certificamos que esta é a versão original e final do Projeto que foi julgado adequado para a obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas obtido no curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal do Paraná (IFPR).

Nome completo e por extenso do Professor Orientador (de acordo com o Currículo Lattes)

Professor Orientador

Nome completo e por extenso do Coordenador do Curso (de acordo com o Currículo Lattes)

Coordenação do Curso

Local e Data de aprovação: / /

Agradecemos a Deus por nos permitir a elaboração dessa atividade e dedicamos este trabalho às nossas famílias.

**AGRADECIMENTOS**

Gostaríamos de expressar nossa gratidão a todas as pessoas que estiveram ao nosso lado e, de alguma forma, contribuíram com a conclusão deste trabalho.

Agradecemos às nossas mães por todo suporte que nos foi dado.

Aos nossos amigos, que sempre estiveram ao nosso lado oferecendo apoio emocional e alegria quando mais precisamos.

Aos nossos professores, especialmente ao Prof. Dr. João Henrique Berssanette, por sua orientação, paciência e partilha de conhecimento.

Por fim, gostaríamos de expressar nossa gratidão a todos que nos ajudaram, direta ou indiretamente, nessa jornada acadêmica.

“Conhecereis a verdade, e a verdade vos libertará.”

(João 8:31-32).

**RESUMO**

Objetiva-se neste trabalho desenvolver um site capaz de integrar e entrelaçar bases de dados científicas estruturadas indicadas pelo usuário, eliminando duplicatas e ranqueando publicações através de parâmetros de classificação. A necessidade surge da crescente disponibilidade de publicações científicas online, que gera uma demanda por recursos que facilitem o acesso rápido e organizado a esses materiais. O sistema procura mitigar o interesse de pesquisadores, profissionais e acadêmicos que buscam a organização de dados por meio de uma plataforma de fácil uso. Para alcançar esse objetivo, o desenvolvimento seguirá a partir de ciclos de implementação, usando uma abordagem iterativa, identificando e analisando riscos e necessidades em cada etapa. Tendo como principais recursos as linguagens de programação Python e R. Ao término deste projeto, espera-se fornecer uma ferramenta funcional que aumente a eficiência e a produtividade dos usuários na organização de publicações científicas, contribuindo para a otimização da pesquisa acadêmica.

**Palavras-chave:** bibliometria; entrelaçamento; indexação; ranqueamento; artigos.

**ABSTRACT**

The aim of this work is to develop a website capable of integrating and interweaving structured scientific databases indicated by the user, eliminating duplicates and ranking publications using classification parameters. The need arises from the growing availability of scientific publications online, which generates a demand for resources that facilitate quick and organized access to these materials. The system seeks to mitigate the interest of researchers, professionals and academics looking to organize data through an easy-to-use platform. To achieve this goal, the development will be based on implementation cycles, using an iterative approach, identifying and analyzing risks and needs at each stage. The main resources will be the Python and R programming languages. At the end of this project, it is hoped to provide a functional tool that will increase the efficiency and productivity of users in organizing scientific publications, contributing to the optimization of academic research.

**Keywords**: bibliometrics; interlacing; indexing; ranking; articles.

**SUMÁRIO**

[**1.1 Problema**](#_2s8eyo1) **14**

[**1.2 Objetivos**](#_17dp8vu) **14**

[1.2.1 Geral](#_3rdcrjn) 14

[1.2.2 Específicos](#_26in1rg) 14

[**1.3 Justificativa**](#_lnxbz9) **15**

[**1.4 Organização do Trabalho**](#_35nkun2) **15**

[**2.3 Sistemas Existentes/ Trabalhos Correlatos**](#_44sinio) **17**

[**3.1 Abordagem de Desenvolvimento**](#_z337ya) **19**

[**3.2 Ferramentas e Tecnologias**](#_3j2qqm3) **20**

[**3.3 Arquitetura do Sistema**](#_1y810tw) **20**

[**4.1 Descrição do Projeto**](#_2xcytpi) **21**

[**4.2 Análise do Sistema**](#_1ci93xb) **21**

[4.2.1 Levantamento de Requisitos](#_3whwml4) 22

[4.2.2 Modelagem de Casos de Uso](#_2bn6wsx) 26

[4.2.3 Modelagem de Atividade](#_1pxezwc) 27

[4.2.4 Modelagem de Banco de Dados](#_49x2ik5) 27

[4.2.5 Design de Interface](#_2p2csry) 28

[**4.3 Implementação das Funcionalidades**](#_147n2zr) **28**

[**4.4 Testes e Validação**](#_3o7alnk) **29**

[**5.1 Apresentação do Sistema**](#_ihv636) **30**

[**5.2 GitHub do projeto**](#_32hioqz) **30**

[**5.3 Documentação do Sistema**](#_1hmsyys) **31**

[**6.1 Dificuldade e Limitações**](#_2grqrue) **31**

[**6.2 Trabalhos Futuros**](#_vx1227) **32**

.

1. **INTRODUÇÃO**

A procura por artigos científicos na internet é uma atividade comum para acadêmicos, profissionais e estudiosos que buscam informações pertinentes de acordo com a sua necessidade. Entretanto, pensando no cenário atual, em qual há uma vasta gama de publicações científicas disponíveis online, se faz necessário e de grande valia a existência de recursos que viabilizem a organização dos materiais recuperados.

De início, plataformas que disponibilizam bases de dados estruturadas se mostram eficientes em preencher essa lacuna, já que são capazes de fornecer uma coleção abrangente de artigos e outras publicações acadêmicas de maneira organizada. No entanto, o cenário muda ao passo de que a necessidade de grandes volumes de dados pode não ser acompanhada pela disponibilidade dos mesmos: basta compreender que todas as bases de dados podem conter quantidades massivas de papers, porém, nenhuma delas é, em caráter absoluto, totalizante.

Isso pode ser analisado através de uma premissa clássica, em que “A está para B, mas nem sempre B está para A”. Assim como nem todo conhecimento está contido em um único livro, nem todas as bases de dados científicas contêm todos os artigos disponíveis para consulta. Ou seja, ainda que uma base seja vasta, a ausência de um trabalho em uma não implica na não existência em outra.

Para atenuar tal problema, deve ser tomada uma abordagem sistemática que possibilite a integração e entrelaçamento de várias bases de dados disponíveis, eliminando as duplicatas e indexando-as de acordo com critérios e métricas de qualidade, a fim de otimizar o trabalho do pesquisador interessado. Ao final, almeja-se maximizar a utilização de recursos acadêmicos ao integrar técnicas de processamento de dados e aplicações de técnicas eficazes de filtragem e classificação para identificar os trabalhos mais relevantes a partir de um determinado contexto.

## Problema

Ainda que algumas bases de dados contenham uma grande coleção de artigos sobre determinadas áreas, outras podem oferecer uma cobertura mais limitada ou especializada. Devido a diferenças nos critérios de inclusão, políticas de indexação e fontes de dados utilizadas, é possível que alguns artigos estejam disponíveis em uma base de dados, mas não em outra; ou ainda, estejam presentes em mais de uma coletânea. Como resultado, ao realizar uma pesquisa abrangente, é fundamental consultar várias fontes.

Tendo isso em vista, quando há exigência de empregar uma maior quantidade de bases, muitas vezes, terá problemas, principalmente no que diz respeito à duplicação de dados e a necessidade de fazer toda a filtragem de conteúdo manualmente. Essa falta de praticidade reduz o potencial de obter resultados precisos e bem estruturados de forma rápida e concisa, resultando em redundâncias, desperdício de recursos e falta de otimização de análise e pesquisa.

## Objetivos

### **Geral**

O objetivo principal deste projeto é desenvolver um website capaz de receber bases de dados científicas estruturadas enviadas pelo usuário em arquivos .csv, integrando-as, eliminando duplicatas e gerando uma nova base filtrada. O sistema também permitirá classificar os artigos resultantes por critérios como ano de publicação, número de citações e autor, procurando organizar os materiais recuperados pelo pesquisador.

### **Específicos**

* Identificar as necessidades e demandas dos pesquisadores, profissionais e acadêmicos em relação à busca de artigos científicos na internet;
* Projetar e implementar uma plataforma que possibilite o aproveitamento rápido de uma ampla variedade de artigos e publicações, integrando de forma eficiente bases de dados estruturadas cedidas pelo usuário;
* Implementar métodos de filtragem e classificação para identificar os trabalhos mais significativos, levando em consideração sua provável relevância no meio acadêmico;
* Garantir a eficácia do sistema por meio de testes, garantindo que os resultados sejam consistentes e úteis para os usuários;
* Fornecer uma interface simples e intuitiva que facilite a utilização do site/software, visando principalmente a praticidade do sistema.

## Justificativa

A importância do projeto está em contribuir no aumento da análise eficaz e rápida de bases de dados científicas. Principalmente, mas não apenas, facilitando o trabalho dos pesquisadores e acadêmicos, tendo também um impacto na curva de aprendizagem técnica de processamento de dados e desenvolvimento de soluções para os problemas que abrangem a aprimoração e melhora na utilização do rico repositório de conhecimento disponível.

Em suma, ao eliminar duplicatas e integrar várias fontes de informação, é possível fornecer uma coleção mais ampla de artigos e publicações que, por sua vez, serão muito úteis na tarefa de mitigar o interesse do pesquisador, possibilitando que o indivíduo extraia insights significativos de grandes volumes de dados.

## Organização do Trabalho

Este trabalho está organizado em cinco capítulos, da seguinte forma: no capítulo 1 será apresentada a introdução; no capítulo 2, é apresentada a fundamentação teórica; no capítulo 3, serão detalhados os procedimentos metodológicos adotados para o desenvolvimento do sistema; no capítulo 4, será descrito o processo de desenvolvimento do sistema e suas funcionalidades principais; no capítulo 5, serão apresentados os resultados obtidos e discutidos; e, por fim, no capítulo 6, serão apresentadas as conclusões.

1. **FUNDAMENTAÇÃO TEÓRICA**

Este capítulo apresentará os principais conceitos e teorias que fundamentaram o desenvolvimento do sistema proposto. As decisões sobre funcionalidades, implementações e design do sistema dependem de uma compreensão desses princípios.

**2.1 Integração de Bases de Dados Científicas**

Para compreender a importância da integração de bases de dados científicas, é essencial entender as dificuldades que os pesquisadores enfrentam ao procurar informações pertinentes. A necessidade de bons métodos de bibliometria e técnicas eficientes para lidar com a dispersão de dados é evidente, uma vez que, dada a vastidão de informações disponíveis, um dos principais desafios é permitir que os usuários localizem informações relevantes de forma eficiente.

Ademais, ao examinarmos o conceito de "Ciência Aberta", percebemos sua estreita relação com os objetivos propostos. Através de uma rápida análise, notamos se tratar de um movimento que, em essência, defende a liberdade de comunicação e disseminação da informação. Embasando-se nisso, Fecher e Friesike (2014) elucidam que, para os simpatizantes da ideia, ela se destaca como um método que visa tornar a pesquisa e a disseminação do conhecimento mais eficientes, promovendo a colaboração científica através das ferramentas disponíveis na Web ([VIANA; RAQUEL, 2019, p.69](http://icts.unb.br/jspui/bitstream/10482/41693/1/2019_RaquelViana.pdf)).

Em síntese, ao considerar o que foi descrito, bem como a perspectiva aberta e colaborativa de informações, se torna plausível a relevância de ferramentas como a proposta neste trabalho. Ao combinar e filtrar bases de dados científicas, não só auxilia os pesquisadores na busca por informações relevantes de maneira mais ágil e organizada, mas também agrega conhecimento de várias fontes, facilitando e disseminando o acesso a uma variedade de publicações acadêmicas.

**2.2 Disponibilização do Sistema Como Site**

Considerando o objetivo deste trabalho e a necessidade de fornecer aos pesquisadores uma solução prática e acessível, é importante destacar a fundamentação da escolha em disponibilizar o sistema como website. Com uma simples análise, percebemos que viabilizar informações por meio de um site facilita o acesso e a interação do usuário com elas, proporcionando uma experiência mais intuitiva e eficiente.

Logo, através desse modelo de disponibilização, além de possibilitar o alcance a uma ampla audiência de pesquisadores, é plausível basear-se em decisões que favoreçam a praticidade e usabilidade do sistema, fornecendo uma solução viável e de fácil acesso às necessidades dos usuários.

## 2.3 Sistemas Existentes/ Trabalhos Correlatos

A seguir, serão analisados sistemas similares já existentes no mercado ou trabalhos correlatos desenvolvidos na área de estudo do projeto. Através dessa contextualização, é pretendido que haja um esclarecimento das ideias apresentadas e uma compreensão clara e detalhada de como e onde esses métodos são usados.

De início, podemos observar as principais plataformas que disponibilizam bases de dados estruturadas, essas que serão usadas como fontes centrais de conteúdo.

* Web of Science:

Funcionalidades: Agregação de publicações científicas de diversas áreas, métricas de citação, filtros avançados.

Características: Ampla cobertura de periódicos, ferramentas de análise bibliométrica integradas.

* Scopus Preview:

Funcionalidades: Indexação de artigos científicos, métricas de citação, busca avançada.

Características: Cobertura global, integração com outras ferramentas de análise.

Através delas, o sistema proposto pretende oferecer um entrelaçamento personalizado das bases, integrando-as a partir de exportações .csv contendo os dados obtidos pela pesquisa externa do usuário. A fim de gerar resultados pertinentes e, fundamentalmente, organizados, o sistema pretende apresentar uma lista concisa e de fácil análise dos dados filtrados, visando ao máximo a simplicidade e praticidade ao utilizador.

Adiante, analisaremos dois sistemas já existentes, que se fundamentam na bibliometria de dados:

* “Uso da análise bibliométrica como ferramenta para o levantamento de estudos sobre a metabolômica aplicada na biorremediação de áreas impactadas por hidrocarbonetos” ([ANDRADE; AUGUSTO; JARDIM, 2010](https://www.scielo.br/j/qn/a/XpCZ7QCLbcvdjkhKxHvBGdd/?format=pdf&lang=pt)):

Análise: A fim de manter somente as pesquisas que mais se assemelhavam ao tema proposto, há uma seleção dos artigos e exclusão das duplicatas através da leitura do título e resumo de cada documento obtido nas bases de dados. O Microsoft Office 2016®, por meio do Microsoft Excel® versão 2018.28, foi utilizado para compilação dos resultados e posterior produção de gráficos analíticos. Há uma eficiente e organizada manipulação dos dados para o fim proposto.

* “Estudos etnográficos na ciência da informação: análise bibliométrica na base de dados Web of Science” ([BUFREM; FREITAS; ARAÚJO, 2023](https://www.scielo.br/j/tinf/a/NnkScQxXvMq5yphZP3QGWHv/?lang=en)):

Análise: Este trabalho descreve uma coleção de artigos científicos sobre etnografia que foram indexados na base de dados Web of Science. Após uma pesquisa bibliométrica, 803 artigos são examinados, cobrindo os anos de 1979 a 2020. Retorna que, com 65 artigos, o ano de 2016 é o mais concentrado, e Michael D. Myers é o autor mais produtivo e o com maior número de citações. Há manipulação dos dados, mas não um ranqueamento e apresentação dos artigos filtrados.

Da mesma forma, a proposta deste trabalho propõe o uso de ferramentas bibliométricas para análise de dados, diferenciando-se pelo objetivo apresentado. Sendo capaz de, após a filtragem dos dados, disponibilizá-los para o livre uso do pesquisador.

1. **METODOLOGIA/ MATERIAIS E MÉTODOS**

O desenvolvimento de um sistema depende diretamente da metodologia empregada, sendo gerado a partir dela todo planejamento a ser seguido. Os processos e abordagens utilizadas para implementar o sistema, tal qual as ferramentas e tecnologias usadas no processo de criação, serão apresentadas a seguir.

## Abordagem de Desenvolvimento

Combinando componentes do modelo espiral, o desenvolvimento do sistema seguirá uma abordagem iterativa e gradual. Isso significa que o projeto será dividido em ciclos de desenvolvimento, em qual, antes de cada um, serão identificados e discutidos os principais riscos, incertezas e necessidades.

Essa estratégia foi escolhida pela sua capacidade de se ajustar a possíveis mudanças nos requisitos do projeto e produzir entregas palpáveis ao longo do tempo. Como cada ciclo incluirá novas funções, cada etapa representará uma versão mais avançada e funcional do sistema, permitindo novas validações e progresso contínuo.

## Ferramentas e Tecnologias

As principais ferramentas e tecnologias utilizadas no desenvolvimento do sistema foram:

* Linguagem de programação: HTML, CSS e JavaScript para front-end e Python para back-end.
* Bibliotecas/Frameworks: Em Python, Pandas e Django. Em HTML, CSS e JavaScript, Bootstrap foi utilizado..
* Banco de dados: SQlite3 para armazenamento dos dados.
* Ambiente de desenvolvimento: Visual Studio Code como IDE principal.
* Controle de versão: Git e GitHub para colaboração entre a equipe

## Arquitetura do Sistema

O sistema será implementado usando uma arquitetura cliente-servidor. Isso significa que uma interface web será executada no navegador do usuário e se comunicará com um servidor, que processará e executará tarefas de integração e análise de dados.

Ferramentas em Python:

* Pandas: É usado em todo o processo de filtragem de dados, sendo responsável por unir as bases, remover as duplicatas e manipular os dados das bases de dados.
* Django: É usado para desenvolver aplicações Web de forma rápida e eficiente em Python

Ferramentas de HTML, CSS e JavaScript:

* Bootstrap: Foi usado para auxiliar na estilização web do sistema.

1. **DESENVOLVIMENTO DO SISTEMA**

Neste capítulo, detalhamos o processo de criação do sistema de filtragem de bases de dados estruturadas, desde a concepção inicial até a entrega final. Serão apresentadas as etapas de descrição do projeto, análise detalhada do sistema, implementação das funcionalidades e realização de testes para garantir a qualidade e eficiência do sistema desenvolvido. Este capítulo oferece uma visão abrangente do trabalho prático envolvido na construção do sistema, fornecendo insights valiosos sobre o processo de desenvolvimento de software.

## Descrição do Projeto

O sistema de filtragem de bases de dados proposto tem como finalidade principal proporcionar uma plataforma online de uso simples, capaz de receber e unir bases de dados recuperadas através de busca externa do usuário em diferentes repositórios acadêmicos. De maneira específica, o sistema deve receber distintas bases de dados estruturadas em arquivos .csv, uni-las, remover as duplicatas consequentes da união, gerar uma nova base de dados com o resultado da filtragem. Podendo também exibir os artigos da nova base através de métricas de classificação como ano de publicação, autor, citações.

## Análise do Sistema

Para a análise do sistema, o foco principal foi o planejamento e a definição das funcionalidades essenciais para atender às necessidades dos pesquisadores na gestão de bases de dados acadêmicas. Para isso, foram estabelecidos os principais requisitos, priorizadas as funcionalidades mais importantes e implementadas ferramentas que garantem o desempenho e a confiabilidade do sistema.

O sistema foi desenvolvido usando em grande maioria Python, utilizando a biblioteca Pandas para a filtragem de dados e o framework Django para integração a web, com suporte a autenticação integrada ao Google para simplificar o processo de login dos usuários. Ajustes foram feitos com Bootstrap, framework para desenvolvimento de componentes de interface e front-end através de HTML, CSS e JavaScript.

Principais funcionalidades:

* Gerenciamento de Usuários: Cadastro/login;
* Processamento de Bases de Dados: União e filtragem de arquivos .csv, removendo duplicatas e gerando uma nova base de dados com o resultado da filtragem;
* Exibição e Download: Permite o download da nova base de dados filtrada no formato .csv e visualização das publicações acadêmicas processadas diretamente na plataforma com base em critérios definidos pelo usuário, como ano, citações e autor.

Metodologia de desenvolvimento: durante o desenvolvimento, a abordagem foi centrada na simplicidade e na usabilidade, com ênfase nos seguintes aspectos:

* Interface Intuitiva: Uso do Bootstrap para criar uma interface visualmente simples.
* Manipulação de Dados: Utilização da biblioteca Pandas para garantir o processamento eficiente das bases enviadas.

As abordagens tomadas permitiram uma análise eficaz do desenvolvimento do sistema proposto, garantindo uma entrega contínua de funcionalidades.

### **Levantamento de Requisitos**

Durante o levantamento de requisitos, foram identificadas necessidades de professores e pesquisadores, coletadas através de conversas com nosso orientador Prof. Dr. João Henrique Berssanette e uma técnica de biblioteconomia da UTFPR.

Requisitos-Funcionais e Casos de Uso:

| **ID** | **Caso de Uso** | **Descrição** | **Evento** | **Prioridade** | **Dependente** |
| --- | --- | --- | --- | --- | --- |
| **RF01** | Gerenciar Usuário | Usuário gerencia (cadastrar, alterar, excluir, consultar) dados do usuário: nome, data de nascimento, e-mail, login, senha, status. | Usuário solicita gerenciar dados do usuário | Alta | Usuário deve estar cadastrado no sistema |
| **RF02** | Efetuar Login | Usuário efetua login no sistema informando login e senha. | Usuário solicita efetuar login no sistema | Alta | Usuário deve estar cadastrado no sistema |
| **RF03** | Efetuar Logout | Usuário efetua logout do sistema. | Usuário solicita efetuar logout no sistema | Alta | Usuário deve estar logado no sistema |
| **RF04** | Enviar Bases de Dados | Usuário envia ao sistema bases de dados .csv. | Usuário solicita enviar bases de dados ao sistema | Alta | Usuário deve estar logado no sistema |
| **RF05** | Processar Bases de Dados | Sistema processa as bases de dados enviadas pelo usuário: unificar bases de dados, remoção de duplicatas. | Sistema processa as bases de dados removendo duplicatas | Alta | Depende do usuário cadastrado inserir as bases de dados. |
| **RF06** | Classificar Bases de Dados Processadas | Sistema permite ranqueamento de publicações na exibição: relevância, citações, ano, autor. | Sistema permite classificar a base de dados processada de diferentes formas | Alta | Depende do usuário cadastrado inserir as bases de dados. |
| **RF07** | Download da Base de Dados Processada | O sistema possibilita ao usuário efetuar download da base de dados filtrada. | Sistema libera download do arquivo processado contendo as bases de dados filtradas | Alta | Depende do usuário cadastrado inserir as bases de dados. |
| **RF08** | Relatório de Administrador | O sistema possibilita ao administrador ter acesso a relatórios de uso e desempenho do sistema. | Sistema mostra relatório de dados ao Administrador | Baixa | Administrador deve estar conectado ao sistema |
| **RF09** | Relatório de Usuário | O sistema possibilita ao usuário ter acesso ao seu histórico de uso. | Sistema mostra relatório de dados ao Usuário | Baixa | Usuário deve estar logado no sistema |

Requisitos Não-Funcionais:

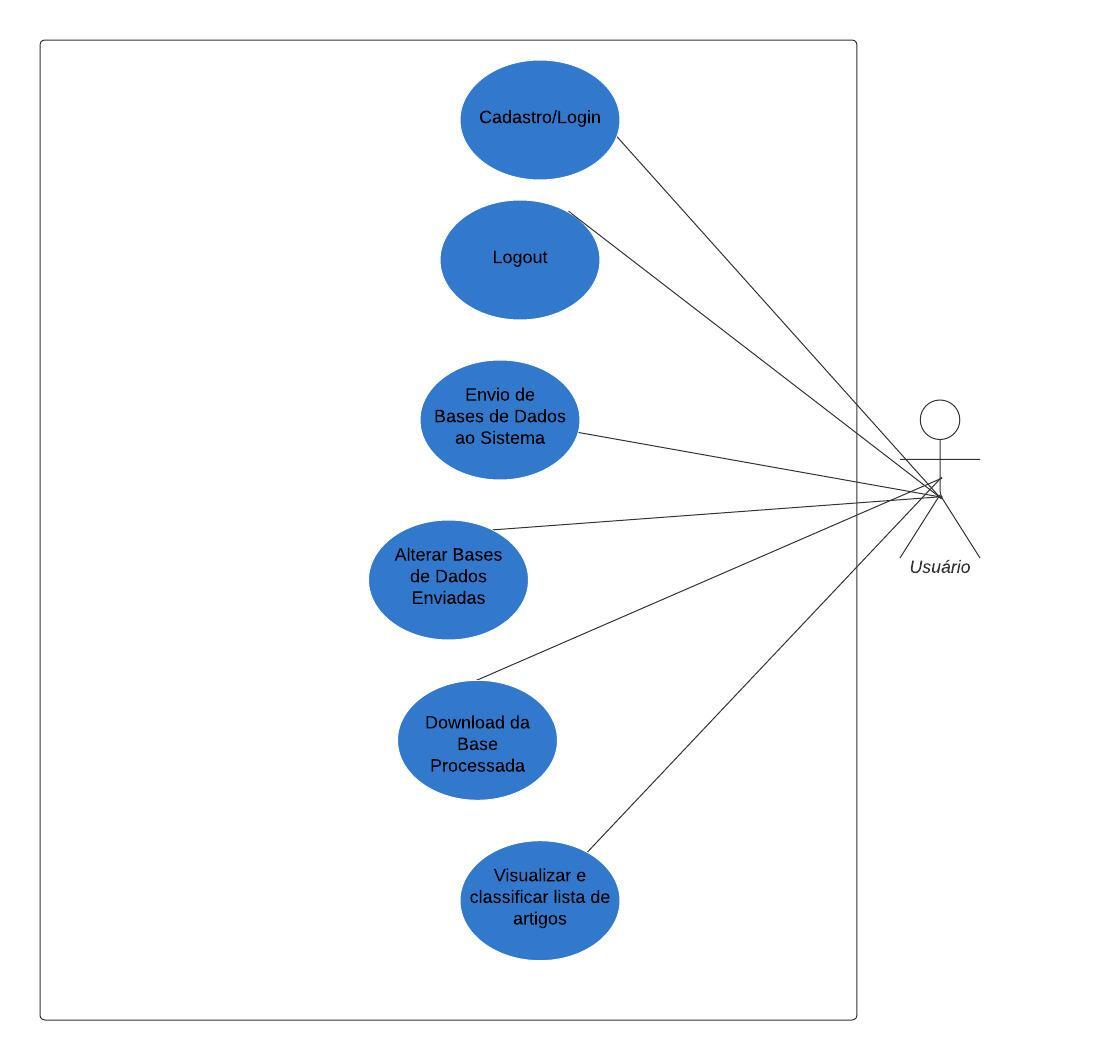
| **ID** | **Caso de Uso** | **Descrição** | **Prioridade** |
| --- | --- | --- | --- |
| **RNF01** | Usabilidade | Ser de fácil uso, com uma interface simples e prática. | Alta |
| **RNF02** | Desempenho | Ser capaz de gerenciar grandes volumes de dados de forma rápida. | Alta |
| **RNF03** | Manutenção | Ter simples padrões de desenvolvimento para que atualizações e modificações, caso necessário, sejam fáceis. | Alta |
| **RNF04** | Confiabilidade | Garantir a eficácia do sistema por meio de testes, garantindo que os resultados sejam consistentes e úteis para os usuários. | Alta |
| **RNF05** | Relevância | O objetivo é criar uma plataforma para acessar e utilizar rapidamente uma diversidade de artigos e publicações. Serão implementados métodos de filtragem e classificação para identificar os trabalhos mais relevantes academicamente. | Alta |

Lista de Atividades:

| **#** | **Descrição** | **Tipo** | **Ator** | **Dados** |
| --- | --- | --- | --- | --- |
| 1 | Cadastro | Cadastro | Usuário | Nome, Sobrenome, Data de nascimento, E-mail, Login, Senha |
| 2 | Login | Cadastro | Usuário | Login, Senha |
| 3 | Consulta de dados | Cadastro | Usuário | Login, Senha |
| 4 | Alteração de dados | Edição | Usuário | Login, Senha |
| 5 | Encerrar aplicação | Cadastro | Usuário | Login, Senha |
| 6 | Importar bases de dados | Gerenciar | Usuário | .csv |
| 7 | Alterar bases de dados enviadas | Gerenciar | Usuário | .csv |
| 8 | Unir bases de dados | Gerenciar | Sistema | .csv |
| 9 | Remover duplicatas | Gerenciar | Sistema | .csv |
| 10 | Ranquear publicações | Gerenciar | Sistema | .csv |
| 11 | Exibir lista processada | Gerenciar | Sistema | .txt |
| 12 | Permitir download | Gerenciar | Sistema | .csv |
| 13 | Mostrar relatório ao administrador |  | Sistema | .txt |
| 14 | Mostrar relatório ao Usuário |  | Sistema | .txt |

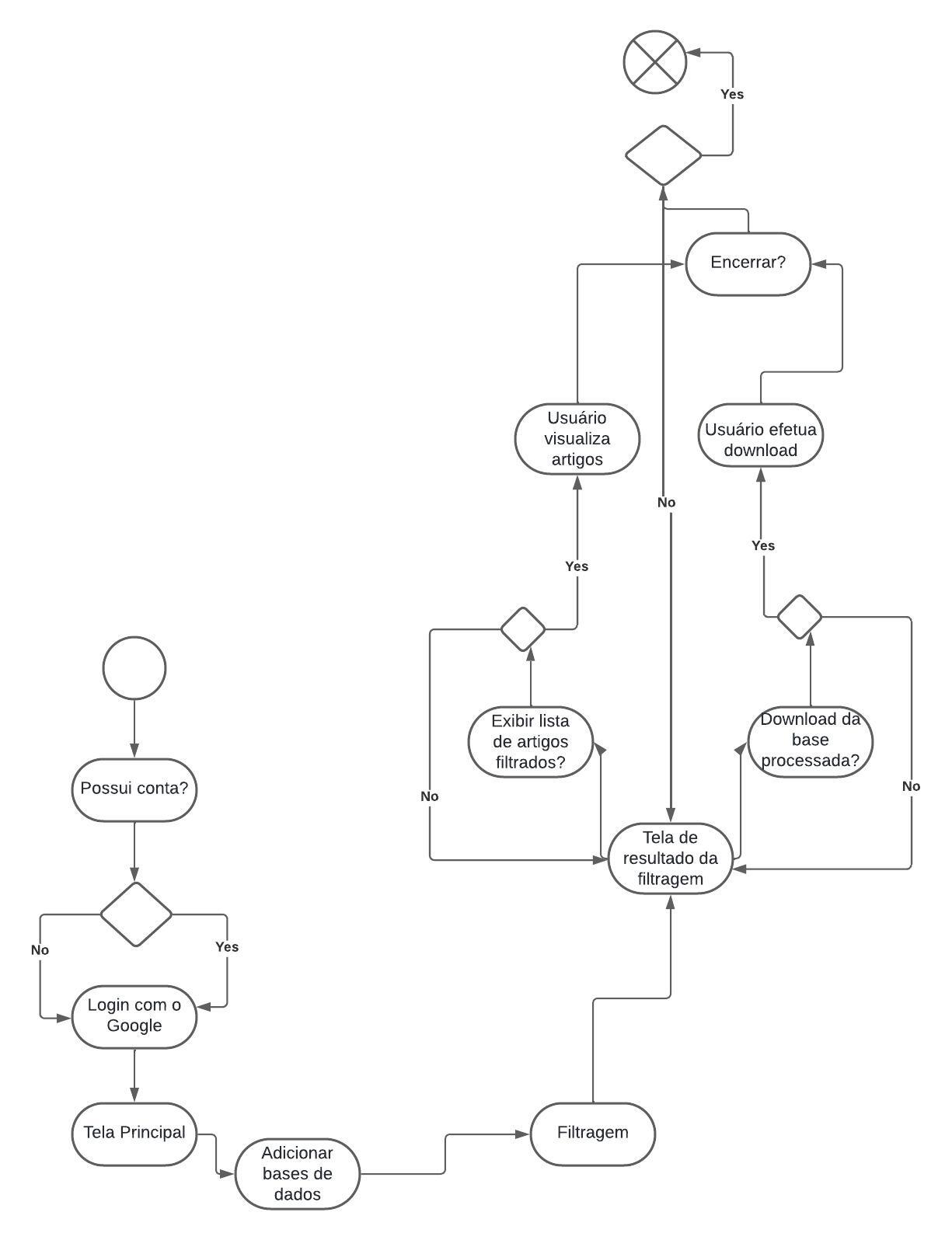
### **Modelagem de Casos de Uso**

Os casos de uso representam as interações entre o usuário e o sistema, descrevendo os principais cenários de uso. Aqui são destacadas as ações principais do usuário com relação ao sistema, como seu cadastro/login e logout, envio e alteração das bases de dados no sistema e download ou exibição do resultado da filtragem.



### **Modelagem de Atividade**

Os diagramas de atividade representam o fluxo de atividades dentro do sistema, elucidando a forma com que as ações do usuário ocorrem em relação às respostas do sistema. Aqui é destacado o processo pelo qual o usuário é condicionado, desde seu login até o encerramento da aplicação.



### **Modelagem de Banco de Dados**

Nesta etapa, foi definida a estrutura do banco de dados do sistema, considerando os requisitos e as funcionalidades planejadas. Como o único objetivo de armazenamento é gerenciar os dados relacionados ao login dos usuários através do Google, optou-se pelo uso do SQLite3, que é simples e perfeitamente integrado ao Django, o framework utilizado no desenvolvimento do sistema.

Os dados armazenados estão centralizados na autenticação de usuários, permitindo o gerenciamento eficiente de credenciais de login.

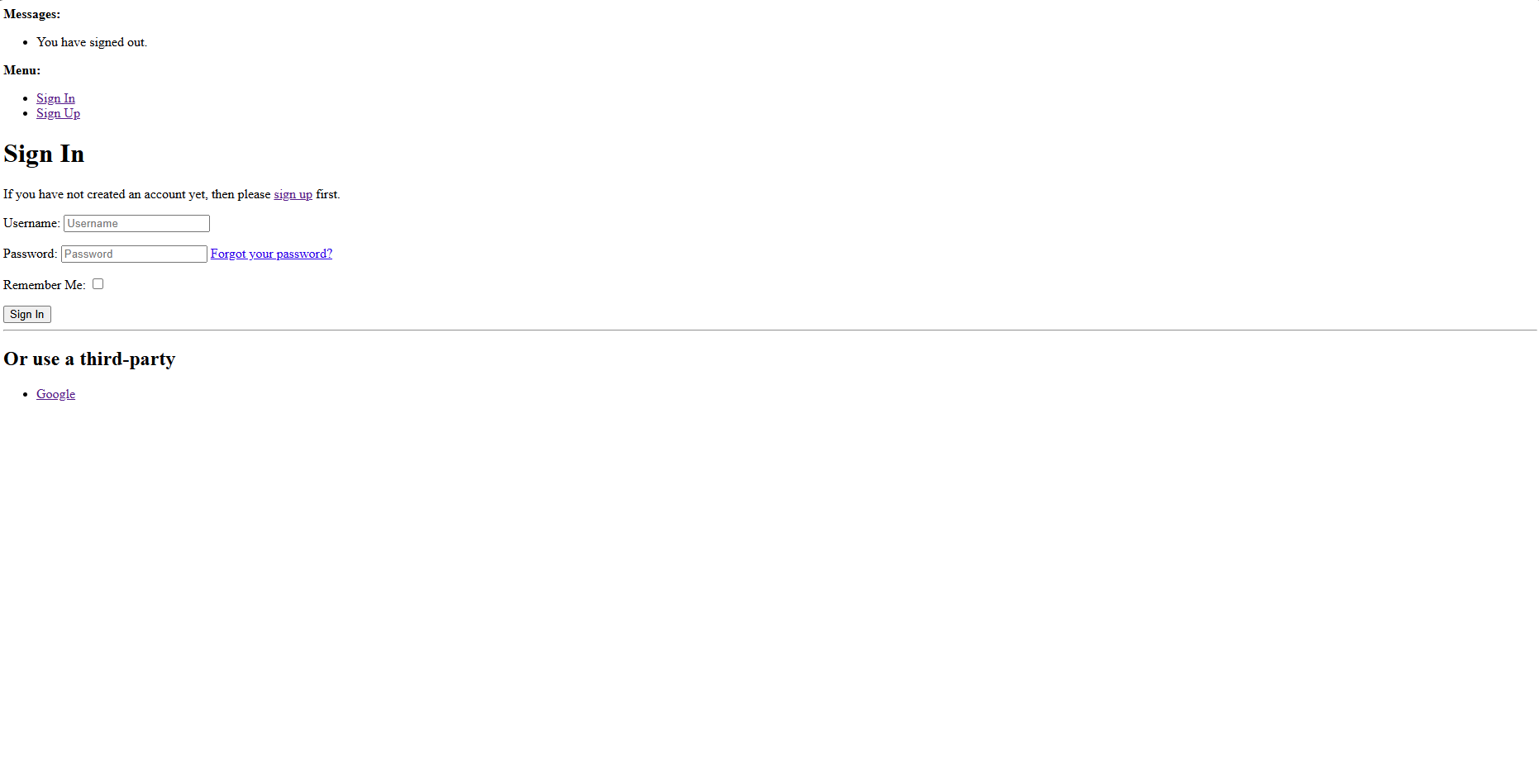
### **Design de Interface**

O design de interface do sistema foi projetado para proporcionar uma experiência de usuário simples, atendendo, em essência, apenas as funcionalidades principais do sistema. As interfaces foram organizadas em telas distintas, cada uma dedicada a uma etapa específica do sistema.

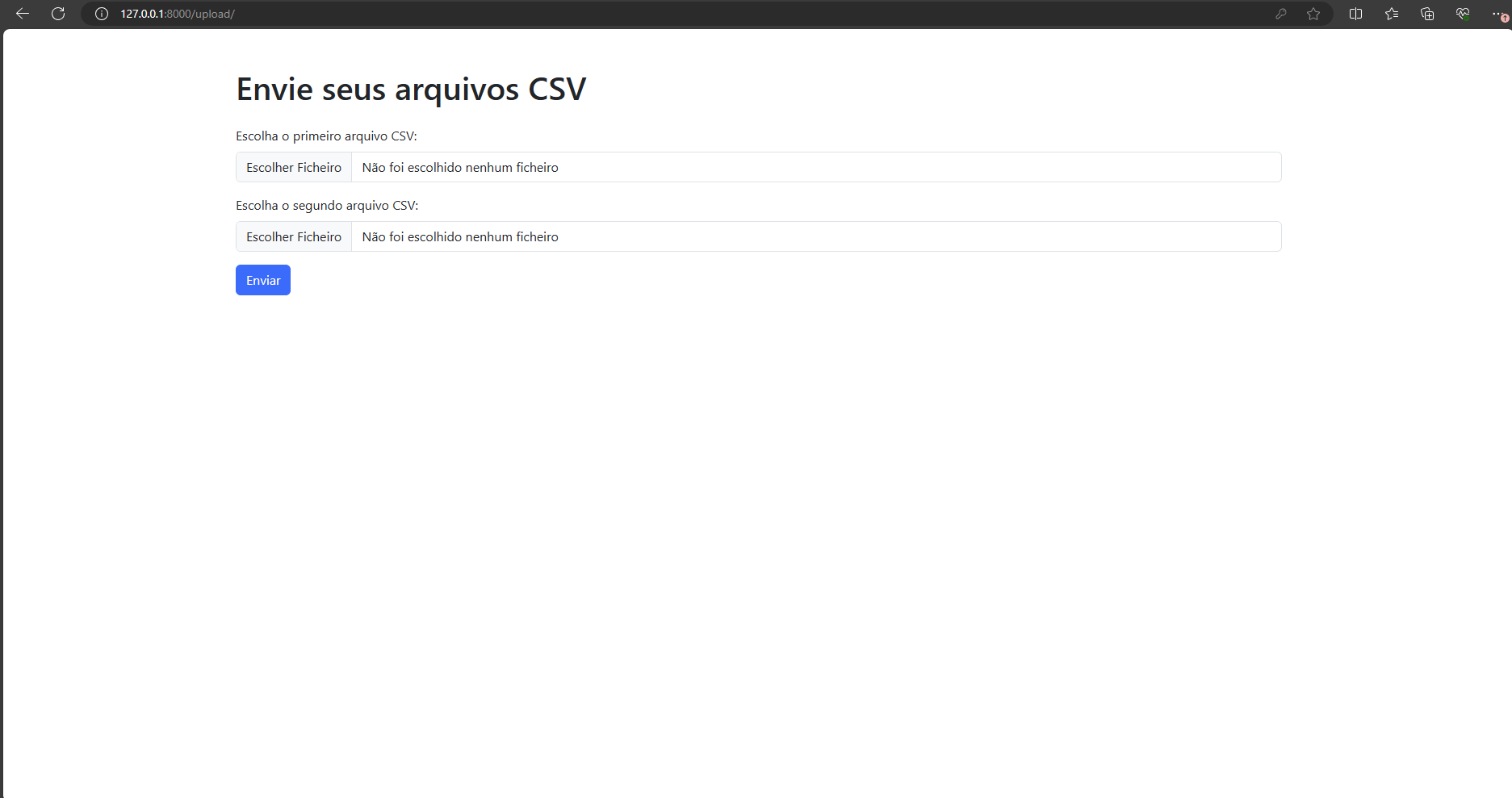
A tela inicial é a tela de login que utiliza o sistema de autenticação do Google. Após o login, o usuário é direcionado para a tela principal, onde pode realizar o upload das bases de dados no formato .csv. O layout dessa tela utiliza caixas interativas para envio de bases e botões destacados para iniciar o processo de filtragem.

Após a conclusão do processamento, o sistema proporciona ao usuário, através de botões, a possibilidade de fazer o download da nova base filtrada ou ser redirecionado para a tela de exibição dos artigos. Na tela de exibição, os artigos processados são apresentados, com funcionalidades que permitem ordená-los por critérios de classificação.

Tela de login:



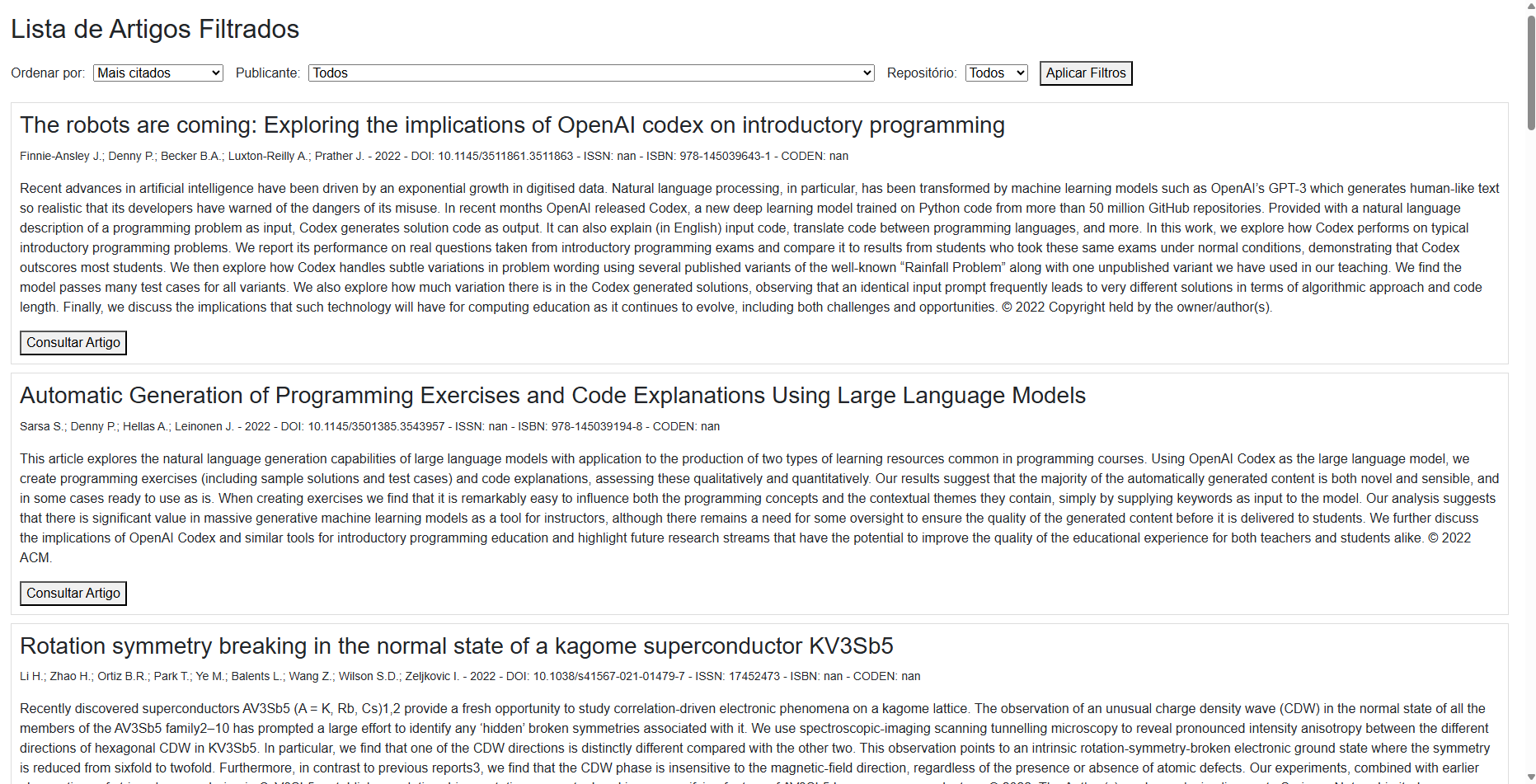
Tela inicial:



Tela pós recebimento e filtragem de dados:



Tela de exibição e classificação de artigos



## Implementação das Funcionalidades

A implementação do sistema foi conduzida de forma iterativa, integrando gradualmente as funcionalidades planejadas. O primeiro passo foi desenvolver a funcionalidade principal: unificação de bases de dados e remoção de duplicatas. Utilizando a biblioteca Pandas, foi desenvolvido um script para processar arquivos .csv, unificá-los e eliminar duplicatas, gerando uma base filtrada.

Em seguida, essa funcionalidade foi integrada a uma aplicação web utilizando o framework Django. A aplicação permitiu que usuários enviassem bases de dados ao sistema, que as processa e retorna a possibilidade do download dos arquivos filtrados diretamente pelo sistema. Posteriormente, foi desenvolvida a funcionalidade para exibir artigos classificados por critérios, também utilizando Pandas para o processamento.

Para o controle de acesso, foi implementado um sistema de login via Google, aproveitando as ferramentas nativas do Django. Após autenticados, os usuários têm acesso a tela principal do sistema, podendo então utilizá-lo.

## Testes e Validação

Foram realizados testes em cada parte implementada do sistema a fim de garantir o funcionamento pleno de tudo que o compõe. Na funcionalidade básica de filtragem e redução de duplicatas, foram realizados testes para garantir que todos os artigos duplicados foram realmente removidos, certificando de que o sistema é capaz de lidar com grandes volumes de dados recuperados de diferentes repositórios. Também foram conduzidos testes para assegurar que a exibição por métricas de classificação funcionava da maneira esperada, exibindo os artigos fielmente aos parâmetros selecionados pelo usuário. No processo de login, foi averiguado se os dados eram guardados no banco de dados, garantindo a segurança dos dados e o redirecionamento para tela inicial após o usuário logado.

1. **RESULTADOS**

Entre os resultados obtidos, destaca-se um sistema funcional que processa e filtra bases de dados em formato .csv, removendo duplicatas e gerando uma nova base melhor organizada. Permitindo também a visualização da mesma através de critérios de classificação como número de citações, ano de publicação, autor e revista.

A interface web, criada com Django e estilizada com Bootstrap, oferece uma navegação intuitiva e eficiente, com funcionalidades claras que vão desde o upload das bases até a visualização e download dos resultados processados. Também destaca-se a implementação do login via Google, que assegura que apenas usuários autenticados possam acessar as funcionalidades do sistema. Esses resultados evidenciam a utilidade do sistema proposto para a comunidade acadêmica

## Apresentação do Sistema

O sistema desenvolvido apresenta uma solução integrada e eficiente para o gerenciamento eficiente de diferentes bases de dados científicas. O sistema automatiza a união dos arquivos, identifica e remove duplicatas, resultando em uma base consolidada e organizada para análise.

Entre suas principais funcionalidades, destaca-se a filtragem de dados a partir das bases de dados do usuário e a exibição personalizada dos artigos filtrados, permitindo que os usuários visualizem os 20 principais registros classificados por critérios como número de citações, ano de publicação, autor ou repositório. Essas funcionalidades facilitam a análise inicial e otimizam a busca por publicações relevantes, proporcionando economia de tempo ao pesquisador.

A integração das funcionalidades do sistema evidencia sua capacidade de simplificar tarefas complexas relacionadas à manipulação e análise de dados acadêmicos. Assim, o sistema foi capaz de atingir seus objetivos iniciais, se estabelecendo como uma ferramenta relevante para pesquisadores que necessitam de agilidade na análise de grandes volumes de informações científicas.

## GitHub do projeto

O código-fonte do sistema de entrelaçamento de bases de dados científicas, redução de duplicatas e priorização por métricas de qualidade está hospedado em um repositório no GitHub, fornecendo acesso aberto e transparente ao código fonte e ao histórico de desenvolvimento. No repositório do projeto, os interessados podem explorar o código, revisar as alterações recentes, contribuir com sugestões e reportar problemas.

O presente projeto pode ser acessado por meio do seguinte endereço web:

## Documentação do Sistema

A presente documentação do projeto é essencial para compreender a operação e o funcionamento do sistema desenvolvido. Nas seções abordadas, foram apresentados os diferentes aspectos do sistema, que, aliadas ao readme.txt (presente no github do projeto), auxiliam na compreensão e utilização do sistema.

1. **CONCLUSÃO**

O trabalho realizado reflete a aplicação de conceitos técnicos de programação e integração de ferramentas, resultando em um sistema funcional e eficiente para a remoção de duplicatas em bases de dados científicas. A utilização do Python, sua biblioteca Pandas e o Framework Django foi indispensável para a realização eficiente do projeto proposto.

Ao longo deste projeto, foi possível atender aos requisitos iniciais, oferecendo uma interface intuitiva e um fluxo de trabalho eficiente para o envio, filtragem e download de bases de dados. A solução final reflete a união de tecnologias que simplificam o processo de gestão de dados e melhoram a eficiência dos usuários na execução de tarefas complexas, como a eliminação de duplicatas e organização de grandes quantidades de artigos científicos.

O sistema desenvolvido representa uma boa solução para os problemas antes apresentados, tendo grande potencial de aprimoramento. As direções futuras incluem a possibilidade de integrar mais funcionalidades, como a implementação de filtros avançados e expansão do sistema para lidar com um maior número de bases de dados. Em conclusão, o sistema atendeu aos objetivos propostos, demonstrando sua eficácia para a área de gestão de dados científicos.

## Dificuldade e Limitações

Durante o desenvolvimento do sistema, algumas dificuldades e limitações foram identificadas. Entre elas, destaca-se questões relacionadas à garantia de funcionalidade plena e eficiente do sistema, de maneira indiferente à estrutura e tamanho das bases enviadas para filtragem. Também é possível destacar a dificuldade de assegurar um login funcional e seguro aos usuários. Além disso, restrições de tempo também representaram desafios significativos ao longo do processo de desenvolvimento.

## Trabalhos Futuros

Ao notar que ainda existem possibilidades e melhorias a serem implementadas no sistema desenvolvido, é plausível que no futuro haja atualizações a fim de expandir e aprimorar aquilo que já foi alcançado. Entre os possíveis passos futuros, estão a otimização completa do sistema, tornando-o capaz de entrelaçar várias bases de dados de forma simultânea, fornecer uma exibição e ranqueamento de artigos mais abrangente e interativa, com uma maior quantidade de insights a respeito das obras, a melhora das interfaces gráficas e implementação de novas funcionalidades de interação entre usuário e sistema.

**REFERÊNCIAS**

BALL, Rafael; TUNGER, Dirk. **Bibliometric analysis-a new business area for information professionals in libraries? Support for scientific research by perception and trend analysis.** Scientometrics, v. 66, p. 561-577, 2006.

SCHÜTZE, Hinrich; MANNING, Christopher D.; RAGHAVAN, Prabhakar. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

PRESSMAN, Roger S.; MAXIM, Bruce R. **Engenharia de Software: Uma Abordagem Profissional**. 8. ed. Porto Alegre: AMGH, 2016.

BUFREM, L. S.; FREITAS, J. L.; ARAÚJO, P. C. **Estudos etnográficos na ciência da informação: análise bibliométrica na base de dados Web of Science.** Transinformação, v. 35, e236917, 2023.

ANDRADE, J. A.; AUGUSTO, F.; JARDIM, I. C. S. F. **Uso da análise bibliométrica como ferramenta para o levantamento de estudos sobre a metabolômica aplicada na biorremediação de áreas impactadas por hidrocarbonetos.** Ecletica Quim. J., v. 35, p. 17, 2010.

VIANA, R. **Compartilhamento de dados de pesquisa em repositórios digitais: o cenário latino-americano.** Brasília: Faculdade Ciência da Informação (FCI), 2019.

NIELSEN, J. **Usability engineering.** San Francisco: Academic Press, 1993.

SPINAK, E. **Indicadores cienciométricos.** Ciência da Informação, Brasília, v. 27, n. 2, 1998.