Projeto Semestral - Ciência de Dados
Busca e seleção inteligente de Editais para projetos inovadores

Integrantes

Nome: Ahmad Kheder Mahfoud **RA:** 20.01323-0
Nome: Caio Bartolozzi Bastos Godoy de Toledo **RA:** 20.01430-9
Nome: Davi Fernandes Simões Soares **RA:** 20.01099-0
Nome: Lucas Romanato de Oliveira **RA:** 20.00313-7
Nome: Leonardo Campos **RA:** 20.00786-8

Resumo

Neste trabalho, apresentamos um sistema desenvolvido para buscar editais da Finep com base na similaridade textual. O processo incluiu web scraping dos sites da Finep, conversão dos documentos para embeddings e a criação de um banco de dados vetorizado, estruturado para responder a consultas de similaridade. A interface permite ao usuário inserir uma descrição de projeto e receber uma lista de editais compatíveis.

Introdução e Contextualização

Neste trabalho, apresentamos o desenvolvimento de um sistema inovador para busca e recomendação de editais da Financiadora de Estudos e Projetos (Finep), focado em identificar oportunidades de financiamento com base na similaridade textual entre as descrições de projetos e os editais disponíveis. O sistema foi projetado com o intuito de facilitar o processo de pesquisa e adequação dos editais à realidade de diferentes projetos, melhorando a assertividade na escolha dos programas de fomento mais alinhados às necessidades do usuário.

O processo inicial envolveu a utilização de técnicas de web *scraping* para coleta automática de informações nos portais da Finep. Essa etapa foi essencial para garantir que todos os editais disponíveis fossem capturados. A partir dos documentos coletados, realizamos a conversão dos textos para *embeddings,* uma técnica avançada de representação vetorial de dados que captura o contexto semântico e permite o cálculo preciso de similaridade entre textos.

Com os *embeddings* gerados, estruturamos um banco de dados vetorizado para armazenar os documentos em um formato que pudesse ser consultado com eficiência. Esse banco vetorial foi projetado para responder rapidamente a consultas de similaridade, realizando comparações entre a descrição de um projeto inserido pelo usuário e os diversos editais armazenados. Por meio dessa estrutura, o sistema é capaz de identificar editais compatíveis com base em semelhanças contextuais, retornando ao usuário uma lista ordenada de oportunidades relevantes.

Metodologia

Neste estudo, uma metodologia foi desenvolvida para realizar uma análise precisa e estruturada de documentos em formato PDF, garantindo a extração, processamento e análise de dados de forma eficiente e funcional. O processo inicia com uma etapa de scraping, onde dados são coletados via requisições HTTP GET a partir de páginas web selecionadas, focando na obtenção de links específicos para documentos em PDF. Essa etapa é realizada de forma programática, onde requisições automatizadas são enviadas para acessar e extrair os dados das páginas-alvo, constituindo a base de dados inicial necessária para as etapas subsequentes.

Após a coleta dos links, uma fase de manipulação de strings é aplicada para filtrar e extrair exclusivamente os links de arquivos em formato PDF. Essa manipulação permite isolar os links de interesse de maneira eficiente, assegurando que apenas arquivos relevantes sejam baixados e analisados. Os documentos em PDF, uma vez baixados, passam por um processo de extração de texto, onde o conteúdo é convertido em formato bruto, possibilitando a manipulação textual posterior. A extração do texto dos PDFs é realizada com o objetivo de tornar o conteúdo acessível para a análise computacional, permitindo que as informações contidas nos documentos sejam processadas de maneira automatizada.

O conteúdo extraído é então submetido a uma divisão em pequenas partes, denominadas chunks. Essa divisão dos textos em chunks é uma etapa essencial para aumentar a granularidade da análise, permitindo que o conteúdo dos PDFs seja processado em segmentos menores e que as informações possam ser analisadas em detalhes. Essa abordagem garante que partes específicas dos documentos possam ser analisadas individualmente, facilitando a aplicação de modelos de aprendizado de máquina em um nível mais profundo e detalhado.

Para realizar uma análise semântica dos chunks, o modelo de embeddings all-MiniLM-L6-v2 foi utilizado. Esse modelo permite transformar o conteúdo textual em representações vetoriais de alta dimensão, conhecidas como embeddings, que capturam as relações semânticas entre diferentes trechos dos documentos. A utilização de embeddings fornece uma base poderosa para cálculos de similaridade, permitindo que partes dos textos sejam comparadas entre si de maneira precisa e eficaz. Em seguida, uma função de cálculo de similaridade média é aplicada para quantificar a semelhança entre os chunks de diferentes documentos. Essa função baseia-se na relação entre o número de chunks semelhantes de um documento e o total de chunks, fornecendo uma métrica objetiva de similaridade que facilita a análise comparativa entre documentos.

A implementação desse fluxo metodológico requer um backend robusto para coordenar e executar as diferentes operações de processamento de dados. Esse backend foi desenvolvido utilizando o framework Flask em Python, que fornece uma estrutura leve e eficaz para a criação de APIs e a manipulação de dados. O backend é responsável por receber e processar as requisições dos clientes web, executando o cálculo de similaridade com a base construída previamente, e retorno dos resultados.

Para facilitar o acesso e a utilização da ferramenta, foi desenvolvida uma interface front-end em formato de Single Page Application (SPA) utilizando o framework ReactJS. Essa SPA permite que os usuários interajam com o backend de forma intuitiva, enviando dados e visualizando os resultados de maneira dinâmica e responsiva. A aplicação em ReactJS garante uma experiência de uso contínua e otimizada, com comunicação fluida entre o cliente e o servidor.

Essa metodologia integrada permite não apenas a coleta e processamento de dados de maneira automatizada, mas também a análise semântica detalhada de documentos, oferecendo uma ferramenta poderosa para análise de similaridade entre arquivos PDF e uma descrição. A combinação de técnicas modernas de scraping, processamento de linguagem

natural e integração web resultou em uma solução eficiente para manipulação e análise de dados textuais em grande escala.

Resultados e Discussão

A ferramenta demonstrou-se altamente eficaz ao retornar editais de grande relevância para as descrições de projetos fornecidas, facilitando a identificação das oportunidades mais adequadas de fomento. Esse sistema possibilita uma triagem ágil e precisa dos editais, otimizando o processo de busca e elevando a eficiência na seleção dos programas de financiamento apropriados para projetos de inovação.

Com o uso do FAISS como base de dados vetorizada, o sistema mostrou-se preciso nas consultas e rápido no processamento das requisições, proporcionando uma experiência fluida e resultados de alta qualidade para os usuários. Essa estrutura avançada de dados vetorizados permite que o sistema identifique com precisão editais que melhor correspondam às especificidades dos projetos, contribuindo diretamente para um melhor alinhamento estratégico e maior potencial de sucesso no acesso a recursos.

Conclusão

A plataforma desenvolvida simplifica o processo de busca por editais específicos, oferecendo uma interface acessível e intuitiva para pesquisadores e profissionais. A adoção de técnicas de AI e embeddings para criar uma base vetorizada comprova o potencial dessas tecnologias na curadoria automatizada de documentos.

Referências

FINEP – Financiadora de Estudos e Projetos. Chamadas Públicas. Disponível em: http://www.finep.gov.br/chamadas-publicas/chamadaspublicas?situacao=aberta. Acesso em: 14 nov. 2024.

THE PANDAS DEVELOPMENT TEAM. pandas. Disponível em: https://pandas.pydata.org/. Acesso em: 14 nov. 2024.

PYPI. PyPDF2. Disponível em: https://pypi.org/project/PyPDF2/. Acesso em: 14 nov. 2024.

HUGGING FACE. Sentence Transformers: all-MiniLM-L6-v2. Disponível em: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Acesso em: 14 nov. 2024.

FAISS. Faiss. Disponível em: https://faiss.ai/index.html. Acesso em: 14 nov. 2024.