Análise Semântica e Geração de Resumos em Bases de Conhecimento Acadêmico

1. Introdução

Proponho o desenvolvimento de um **agente inteligente** projetado para auxiliar no desafio de identificar, ler, analisar e sintetizar o conhecimento em uma revisão bibliográfica, fase em que me encontro atualmente no mestrado. O agente atuará como um "assistente de pesquisa pessoal", conectando-se diretamente à biblioteca Zotero do usuário. Ele será capaz de processar o texto completo dos artigos (em formato PDF), extrair informações chave, construir uma base de conhecimento e permitir que o pesquisador interaja com sua própria biblioteca de artigos através de uma interface de linguagem natural.

A ideia é auxiliar em tarefas cognitivas complexas, como a síntese de informações, a identificação de metodologias comuns e a descoberta de lacunas na literatura, acelerando significativamente o processo de revisão da literatura.

2. Componentes do Projeto

1. **Integração de Dados:** Desenvolver um módulo para se conectar à API do Zotero, permitindo o acesso aos metadados e aos arquivos (PDFs) da biblioteca do usuário. Ou acessar os arquivos localmente no computador do usuário
2. **Extração e Pré-processamento de Texto:** Implementar um pipeline para extrair o texto bruto de documentos PDF de forma robusta, realizando a limpeza e a estruturação necessárias para a análise subsequente.
3. **Análise e Extração de Informação:** Utilizar Modelos de Linguagem de Larga Escala (LLMs) para:
    - Gerar resumos abstrativos e extrativos dos artigos.
    - Identificar e extrair entidades nomeadas (NER - Named Entity Recognition) relevantes, como: **metodologias, datasets, métricas de avaliação, ferramentas e problemas de pesquisa**.
4. **Interface Conversacional:** Desenvolver uma interface de usuário (ex: chatbot) que permita ao pesquisador fazer perguntas em linguagem natural sobre sua base de artigos. Exemplos de perguntas:
    - *"Quais artigos utilizam a metodologia X para resolver o problema Y?"*
    - *"Faça um resumo comparativo das abordagens dos artigos de autor A e autor B."*
    - *"Liste os principais datasets mencionados para avaliação de algoritmos de classificação."*