

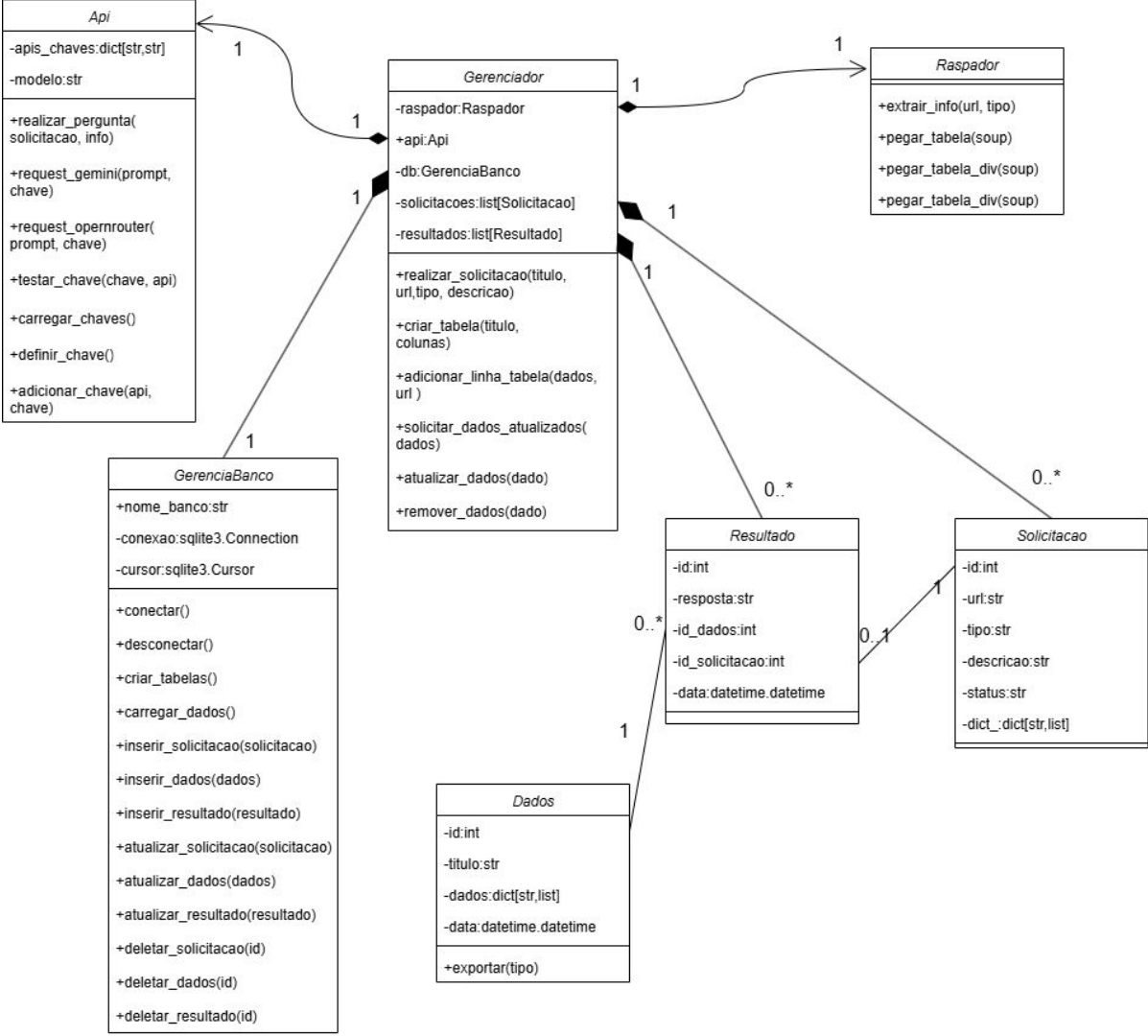
Sistema de localização de informações em páginas web

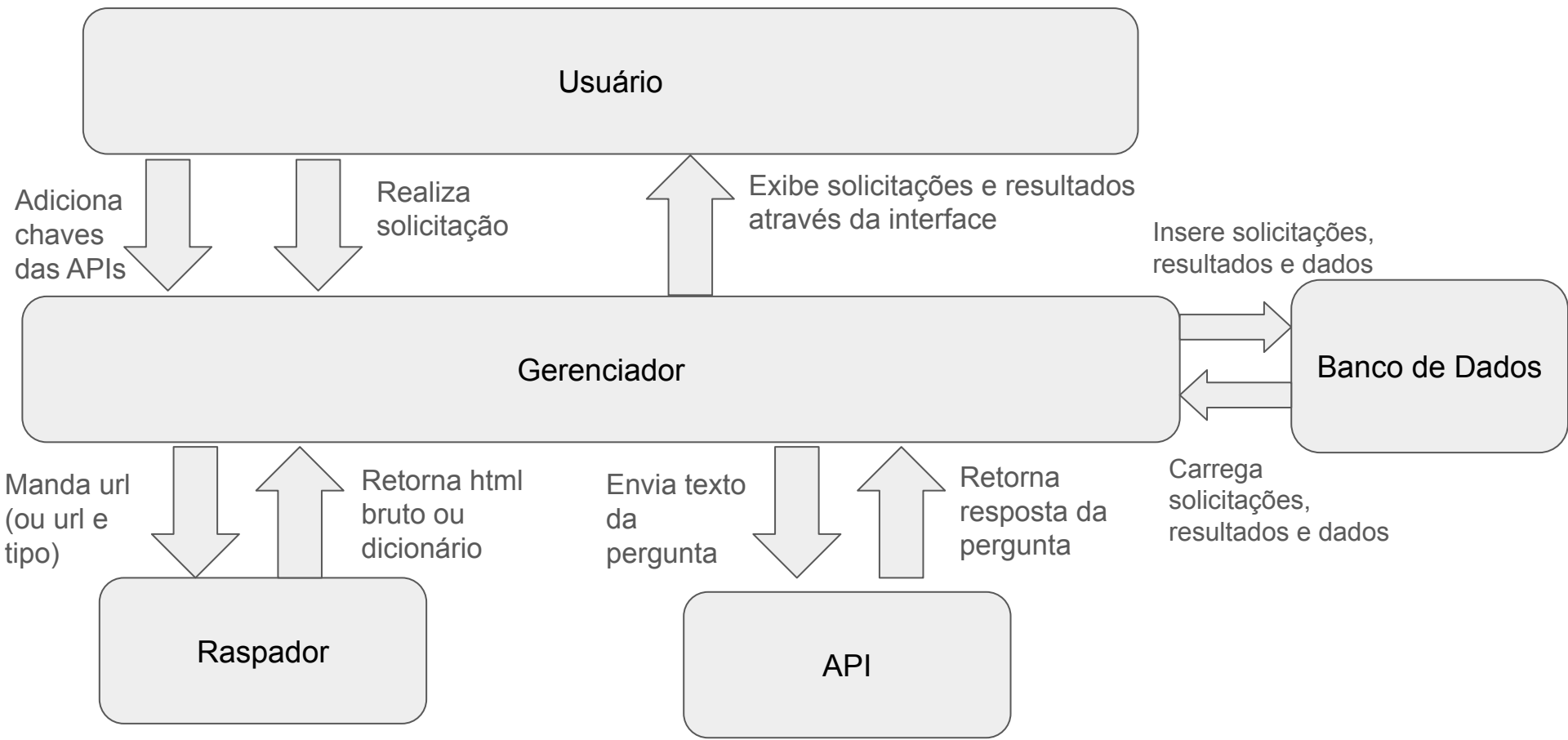
Aluno: João Pedro Darabas Cardoso
Professor: Alex Sandro Roschildt Pinto

Objetivo

Desenvolver um sistema que, utilizando técnicas de webscraping e uma API de um modelo LLM, permita ao usuário solicitar uma informação de algum site e armazene tal informação, para que possa ser visualizada e também atualizada no futuro.

Diagrama UML





Descrição do Funcionamento

- O usuário deve cadastrar pelo menos uma chave API e adicioná-la ao programa (openrouter.ai, Google Gemini);
- Então, pode realizar uma solicitação, inserindo seu tipo (tabela, card, outro), título, url da página e descrição da informação desejada ou criar uma tabela em branco;
- O Gerenciador solicita ao raspador para realizar a raspagem do site, que retorna um dicionário caso a informação foi encontrada ou o html bruto da página caso contrário;
- O Gerenciador então solicita a Api que realize uma pergunta a um modelo de linguagem para que processe tais informações, corrigindo em caso de erro de formatação ou encontrando a informação caso o raspador não tenha a encontrado;

Descrição do Funcionamento

- Em caso de erro, o Gerenciador insere somente a solicitação com um status de erro e informa o mesmo ao usuário;
- Em caso de sucesso, o Gerenciador insere as informações da solicitação e do resultado no banco de dados e também na interface gráfica;
- Após a criação de uma tabela, é possível visualizá-la, adicionar uma nova linha, atualizar as linhas existentes e exportar como arquivo (csv, json, excel, pdf).

Ferramentas utilizadas

PyQt6: Interface gráfica

OpenRouter: API para consulta em LLM

Gemini: API para consulta em LLM

JSON: Interação com a LLM, interação com o banco de dados e exportação de dados

SQLite: Persistência de dados

requests: Biblioteca para realizar chamadas nas APIs

Selenium: Raspagem de dados (para páginas com carregamento dinâmico)

BeautifulSoup: Raspagem de dados

PyFPDF: Exportação de dados

Pandas: Transformação e exportação de dados