Documento Técnico Resumido: Software de Web Scraping e Análise de IA do Índice TIOBE

Introdução

Este documento apresenta um resumo conciso do software desenvolvido para web scraping do Índice TIOBE e análise de comentários de programadores, abordando os pontos essenciais para uma compreensão rápida e eficiente.

Problema & Importância

O software aborda a necessidade de acompanhar a popularidade das linguagens de programação (via Índice TIOBE) e, crucialmente, entender a percepção da comunidade sobre as linguagens mais usadas. Isso é vital para desenvolvedores, empresas e educadores se manterem atualizados com as tendências do mercado e a opinião dos usuários sobre as ferramentas de programação.

Arquitetura do Pipeline (Extração → Engenharia → Análise)

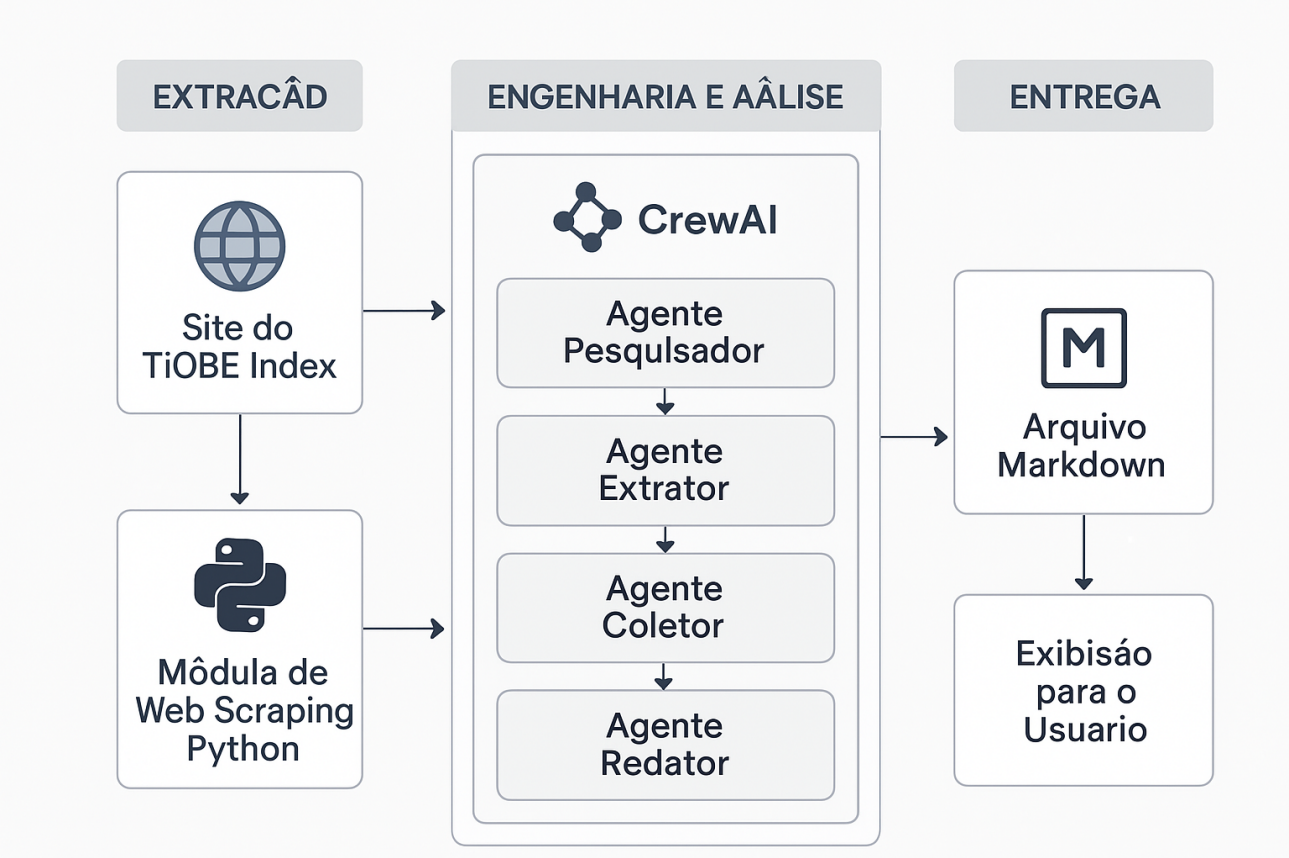
O pipeline do software é dividido em três fases principais:

1. **Extração:** Coleta o ranking atualizado das linguagens de programação do site oficial do Índice TIOBE (https://www.tiobe.com/tiobe-index/) utilizando **Python** com as bibliotecas `requests` e `BeautifulSoup`.

2. **Engenharia e Análise:** Esta é a fase inteligente, onde uma equipe de agentes de IA (**CrewAI**) processa e enriquece os dados:

- **Agente Pesquisador:** Busca URLs com comentários sobre as top 3 linguagens usando a **API Serper (Google)**.
- **Agente Extrator:** Obtém o conteúdo das URLs encontradas usando `requests` e HTML parsing.
- **Agente Coletor:** Analisa o conteúdo para identificar comentários positivos/negativos usando a **API do ChatGPT 4o mini**.
- **Agente Redator:** Resume e organiza os comentários, também utilizando a **API do ChatGPT 4o mini**.

3. **Entrega:** As informações (ranking e comentários resumidos) são formatadas em um arquivo **Markdown** e exibidas ao usuário.



Principais Descobertas (Gráficos/Insights)

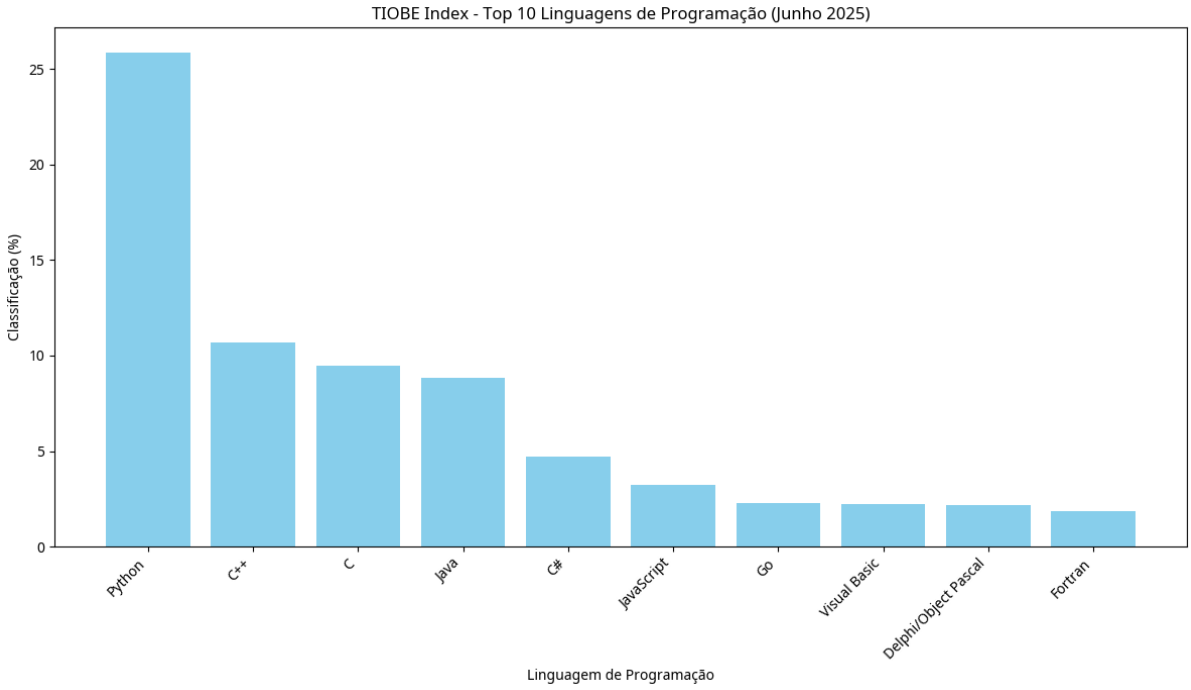
O software gera dois tipos principais de insights:

1. **Insights do Ranking de Popularidade (Índice TIOBE):** Apresenta a dominância das linguagens top (ex: Python, C++, C em Junho de 2025), tendências de crescimento/declínio e comparações entre linguagens. Visualizações incluem

gráficos de barras para o ranking atual e gráficos de linha para tendências históricas (se dados forem coletados ao longo do tempo).

Gráfico de Popularidade das Linguagens (Junho 2025)

Conforme a raspagem do site do TIOBE Index, as linguagens de programação mais populares em Junho de 2025 são:



2. **Insights dos Comentários da Comunidade (Análise Qualitativa):** Fornece uma análise aprofundada da percepção dos programadores sobre as 3 primeiras linguagens, incluindo sentimento geral (positivo/negativo), temas recorrentes em comentários (pontos fortes e fracos), e casos de uso. Visualizações sugeridas incluem nuvens de palavras e gráficos temáticos.

Reflexões Éticas e Legais

O software opera sob considerações éticas e legais importantes:

- **LGPD (Lei Geral de Proteção de Dados):** Se dados pessoais identificáveis forem coletados (mesmo que publicamente disponíveis), a LGPD se aplica. É crucial garantir uma base legal para o tratamento, finalidade clara e segurança dos dados. Recomenda-se anonimizar ou pseudonimizar comentários se vinculados a indivíduos.

- **Direitos Autorais:** O web scraping de conteúdo protegido por direitos autorais pode ser uma violação. O software mitiga isso ao transformar e resumir os comentários, agregando valor em vez de apenas reproduzir. A citação da fonte é uma boa prática.
- **`robots.txt` e Termos de Serviço (ToS):** É fundamental respeitar as diretrizes do `robots.txt` e os Termos de Serviço dos sites para evitar bloqueios e ações legais. O site do TIOBE Index deve ser verificado.
- **Uso de APIs (Serper e ChatGPT 4o mini):** O uso dessas APIs está sujeito aos seus próprios Termos de Serviço e políticas de uso. É responsabilidade do desenvolvedor cumprir esses termos e usar os modelos de IA de forma responsável, ciente de vieses e garantindo a privacidade.

Instruções Claras (README) para Rodar o Pipeline em Outro Ambiente

Um arquivo `README.md` detalhado foi criado para facilitar a configuração e execução do pipeline em novos ambientes. Ele inclui:

- **Pré-requisitos:** Python 3.x e `pip`.
- **Instalação de Dependências:** Comando `pip install requests beautifulsoup4 crewai crewai_tools openai google-search-results`.
- **Configuração de Variáveis de Ambiente:** Instruções para criar um arquivo `.env` com `SERPER_API_KEY` e `OPENAI_API_KEY`.
- **Estrutura do Projeto:** Orientação sobre a organização dos arquivos.
- **Como Executar:** Comandos para navegar, ativar ambiente virtual e rodar o script principal.
- **Solução de Problemas Comuns:** Dicas para lidar com erros de API Key, bloqueio de IP e alterações no layout do site.

Citação das Licenças de Datasets ou APIs Utilizadas

O software utiliza diversas bibliotecas e serviços de terceiros, cada um com suas licenças e termos de uso:

- **Python:** Python Software Foundation License (PSF)
- `requests`: Apache 2.0 License
- `BeautifulSoup`: MIT License
- **CrewAI & CrewAI Tools:** Geralmente MIT License (verificar repositórios oficiais).
- **Serper API (Google Search API):** Sujeito aos seus Termos de Serviço.
- **ChatGPT 4o mini API (OpenAI):** Sujeito aos Termos de Uso da OpenAI e Política de Uso.
- **Matplotlib:** Matplotlib License

É fundamental que o usuário esteja ciente e cumpra os termos de licença e uso de todas as ferramentas e APIs empregadas.