Relatório Técnico: Coleta de Dados do Mercado Livre com OpenAI e Firecrawl

# 1. Introdução

O presente relatório tem como objetivo descrever a metodologia, os principais desafios enfrentados e os resultados obtidos durante o desenvolvimento de uma aplicação de coleta automática de dados (web scraping) de produtos listados no Mercado Livre. O sistema desenvolvido tem como base as tecnologias OpenAI (GPT-4o) e Firecrawl, com foco na extração eficiente de dados relevantes para pesquisas de mercado no setor de impressão (cartuchos, toners, impressoras etc.).  
  
A solução foi estruturada em Python e automatiza três etapas fundamentais:  
1. Geração de sinônimos contextuais para uma palavra-chave base, utilizando o modelo GPT-4o.  
2. Coleta de resultados do Mercado Livre por meio do serviço Firecrawl.  
3. Extração e organização dos dados em um arquivo CSV com informações úteis como título, link e preço dos produtos.

# 2. Metodologia

A metodologia adotada para o projeto visou maximizar a simplicidade e eficiência, utilizando ferramentas modernas com APIs acessíveis e de fácil integração. Os principais pontos do processo foram:

## 2.1. Geração de Sinônimos com OpenAI (GPT-4o)

A primeira etapa consiste em utilizar o modelo gpt-4o-mini para gerar sinônimos contextuais em português relacionados à palavra-chave fornecida. A escolha desse modelo se deve à sua capacidade de gerar sugestões semânticas com boa precisão e à sua resposta rápida para tarefas simples.

## 2.2. Scraping com Firecrawl

Para a captura dos dados da web, optou-se pelo uso da API do Firecrawl, que permite extrair o conteúdo de páginas web em formato estruturado como Markdown ou JSON. A escolha dessa ferramenta se baseou nos seguintes fatores:  
- Facilidade de uso  
- Compatibilidade com o Mercado Livre  
- Evita bloqueios

## 2.3. Extração e Organização em CSV

Após o salvamento dos dados em arquivos .txt, foi implementada uma rotina de parsing com expressões regulares para extrair informações estruturadas como palavra-chave, título, link e preço. Essas informações foram salvas em um arquivo CSV, facilitando a posterior análise ou importação para planilhas e dashboards.

# 3. Principais Desafios

## 3.1. Velocidade de Coleta dos Dados

Apesar da simplicidade do Firecrawl, observou-se que a velocidade de scraping era relativamente baixa ao realizar múltiplas requisições em sequência. Esse comportamento provavelmente está relacionado a medidas de proteção da API ou a limites internos de requisições por minuto.

## 3.2. Prevenção contra Detecção de Web Scraping

Outro desafio importante foi minimizar a possibilidade de o scraping ser detectado como comportamento automatizado. Optamos por capturar apenas a primeira página de resultados para cada sinônimo, o que normalmente retorna em torno de 50 produtos por termo — suficiente para análise inicial sem comprometer a integridade do serviço.

## 3.3. Limitações de Profundidade de Dados

Embora o scraping tenha retornado os principais campos, campos adicionais como avaliações, frete, loja oficial etc. não foram extraídos no momento. Essa decisão foi tomada para manter a aplicação leve, rápida e focada no essencial.

# 4. Resultados Obtidos

Apesar das limitações, os resultados foram considerados muito satisfatórios:  
- A aplicação extraiu com sucesso os principais dados de mais de 100 produtos em testes com poucas palavras-chave.  
- O conteúdo gerado apresenta boa estrutura, com campos fáceis de manipular.  
- O uso de sinônimos aumentou a cobertura semântica das buscas.

# 5. Conclusão e Próximos Passos

O sistema desenvolvido demonstrou ser eficiente, escalável e de fácil manutenção, cumprindo com o objetivo de realizar uma coleta rápida de dados do Mercado Livre para análise de mercado.  
  
Possíveis melhorias futuras incluem:  
- Paginação para capturar mais resultados por sinônimo  
- Extração de avaliações e outros atributos do anúncio  
- Uso de banco de dados  
- Dashboards automáticos para visualização dos resultados

Resumo Final: Este projeto representa uma base sólida para soluções automatizadas de coleta e análise de dados de e-commerce, com aplicação prática imediata e possibilidade de evolução contínua.