

Coleta e Construção da Base de Dados

Integrantes do grupo:

Danilo Ramalho Silva | RM: 555183
Israel Dalcin Alves Diniz | RM: 554668
João Vitor Pires da Silva | RM: 556213
Matheus Hungaro | RM: 555677
Pablo Menezes Barreto | RM: 556389
Tiago Toshio Kumagai Gibo | 556984

link do projeto GitHub:

https://github.com/DaniloRamalhoSilva/HP_Cartucho_Pirata_Detector

1. Visão geral

Este entregável descreve um pipeline automatizado para identificar anúncios de cartuchos HP suspeitos no Mercado Livre. O fluxo compreende:

1. Coleta de URLs de anúncios;
 2. Extração de dados de produto e comentários;
 3. Armazenamento em SQLite;
 4. Rotulagem heurística (original vs. suspeito) combinando regex e LLM.
-

2. Coleta de Dados

2.1 Ferramentas e decisões

- **Selenium WebDriver (Chrome):** necessário para renderizar JavaScript e interagir com botões dinâmicos.
- **Delays e WebDriverWait:** `time.sleep (2–4s)` e espera explícita garantem carregamento e mitigam bloqueios do ML.
- **Seletores CSS:** identificados inspecionando o DOM do Mercado Livre; mantidos genéricos para resistir a pequenas mudanças de layout.

2.2 Pipeline de scraping

Módulo	Responsabilidade
scrap_list	Navega na listagem, captura e extrai links e páginas.
scrap_product	Abre anúncio, coleta: título, preço, avaliações, descrição e vendedor; salva em <code>products_data</code> .
scrap_comments	Acessa iframe de comentários, expande com JS, itera até o limite gravando cada review em <code>products_review</code> .

Decisões no processo

1. **Carregamento de 3s** antes de extrair detalhes do produto para garantir scripts e recursos externos.
 2. **Execução de scripts JS** (scrollIntoView, click com setTimeout) para acionar carregamentos adicionais de conteúdo.
 3. **Tratamento de exceções** em cada etapa para continuar fazendo scraping mesmo com falhas pontuais.
-

3. Estrutura do Banco de Dados

Utilizamos SQLite com três tabelas normalizadas:

- **products_url** (id, produto, url, scraped, data_cadastro)
- **products_data** (id, products_url_id, url, title, price, review_rating, review_amount, seller, description, positive_occurrences, negative_occurrences, label, data_cadastro)
- **products_review** (id, products_data_id, rating, review, review_date, data_cadastro)

Decisões de modelagem

- **Tabelas separadas:** isola reviews para análise granular.
 - **Colunas de evidência:** pré-calcula contagens de keywords para não sobrecarregar o LLM.
-

4. Rotulagem Heurística (Classificação Binária)

4.1 Contagem com regex (update_comment_counts)

- **Objetivo:** resumir rapidamente os comentários sem enviar todo o texto ao LLM.
- **Método:** busca padrões em products_review.review:
 - **Negativas:** falso, pirata, não é original, genérico, defeito.
 - **Positivas:** produto original, ótima qualidade, recomendo.
- **Decisão:** usar regex case-insensitive para cobertura ampla.
- **Resultado:** popula positive_occurrences e negative_occurrences em products_data.

4.2 Classificação com LLM (classify_and_update)

- **Prompt few-shot** com exemplos manuais que definem claramente cada classe.
- **Componentes do prompt:** title, price, seller, description, positive_occurrences, negative_occurrences.
- **Configurações:** temperature=0 (saída estável), modelo gpt-4o-mini.

- **Fallback:** qualquer resposta fora de original/suspeito é tratada como suspeito.
- **Racional:** combina dados quantitativos (contagens) e qualitativos (texto) para decisão mais robusta, sem necessidade de treinar modelo próprio.

5. Principais features e exemplos de amostras de dados

5.1 Features coletadas e descrição

Feature	Descrição
title	Título do anúncio
price	Preço do produto
description	Texto descritivo do produto
seller	Nome do vendedor
review_rating	Avaliação média do produto
review_amount	Número total de avaliações
url	Link direto do anúncio
positive_occurrences	Ocorrências positivas (menções positivas)
negative_occurrences	Ocorrências negativas (menções suspeitas/piratas)
label (target)	Classificação: "original" ou "suspeito"

5.1 Amostras de dados

title	price	description	seller	review_rating	review_amount	url	positive_occurrences	negative_occurrences	label
Cartucho Hp 3ed70a N°712 Preto 38ml Hp	260.91	O Cartucho HP 712 Preto 38ml 3ED70A	INK LASER INFO	4.8	28	https://www...	0	0	original
Cartucho De Tinta Hp 711 Cor Azul Do 29 MI	279.0	HP Designjet T120 e ePrinter HP Designjet série T520	OBERO INFORMATICA	4.6	28	https://www...	6	0	suspeito
Cartucho de Tinta HP 60B Preto Simples CC636WB 4ml	107.2	Cartucho de Tinta HP 60B Preto Simples	VANMASTERC OMERCIO	4.8	121	https://www...	7	0	suspeito
Kit 4 Cartuchos De Tinta Hp 712 Preto 38ml + Cores	979.0	KIT COM 4 CARTUCHOS DE TINTA 712 HP	OBERO INFORMATICA	4.9	8	https://www...	4	0	original
Cartucho Hp 3ed70a N°712 Preto 38ml Hp	260.91	O Cartucho HP 712 Preto 38ml 3ED70A ...	INK LASER INFO	4.8	28	https://www...	0	0	original

6. Conclusão

O projeto HP Cartucho Pirata Detector demonstrou com êxito a viabilidade técnica de se desenvolver uma solução inteligente e automatizada para apoio à identificação de produtos suspeitos de pirataria em marketplaces. Por meio da integração entre técnicas de Web Scraping, análise textual e modelos de linguagem avançados (LLM), foi possível construir um pipeline robusto que simula de forma eficaz a lógica de auditoria e controle de qualidade aplicada a anúncios digitais.

A estruturação do banco de dados com registros normalizados permitiu não apenas a organização eficiente dos dados coletados, mas também a criação de métricas objetivas para subsidiar as decisões do modelo classificador. O uso de expressões regulares possibilitou um primeiro filtro confiável baseado em evidências dos próprios consumidores, enquanto o uso de inteligência artificial trouxe um segundo nível de refinamento, combinando análise qualitativa e quantitativa dos anúncios.

A exportação final em CSV reforça o compromisso com a interoperabilidade dos dados e abre portas para futuras integrações com dashboards, modelos supervisionados e ferramentas de monitoramento em tempo real. A abordagem adotada também serve como base para soluções similares em outros segmentos de e-commerce.

Concluimos, portanto, que todos os objetivos propostos foram plenamente alcançados, com destaque para a qualidade técnica, organização estrutural e aplicabilidade real da solução desenvolvida.