Relatório Técnico - Entregável 2: Processamento e Análise de Dados de Produtos HP

1. Introdução

Este relatório detalha as etapas de processamento e análise de dados de produtos HP, conforme as demandas do Entregável 2 do projeto. O objetivo principal foi transformar dados brutos de web scraping em uma base de dados limpa, enriquecida e consolidada, pronta para análises mais aprofundadas e aplicações de Machine Learning e Processamento de Linguagem Natural (PLN).

2. Limpeza e Padronização dos Dados

A fase inicial do processamento de dados focou na limpeza e padronização dos dados brutos coletados. Isso incluiu a remoção de caracteres indesejados, tratamento de valores ausentes e normalização de formatos para garantir a consistência e a qualidade dos dados.

2.1. Correção de Campos

Foram realizadas as seguintes operações de limpeza:

- **Remoção de Símbolos e Quebras de Linha:** Caracteres especiais como `R$`, `%`, `.` (em preços), `,` (em números de reviews) e quebras de linha (`\n`) foram removidos ou substituídos para facilitar a conversão para tipos numéricos e a análise textual.
- **Tratamento de Espaços:** Espaços em excesso no início ou fim das strings foram removidos (`strip()`), e múltiplos espaços internos foram substituídos por um único espaço.
- **Erros de Codificação:** Embora não tenham sido identificados erros de codificação significativos nos dados de demonstração, o pipeline foi projetado para lidar com `utf-8` para garantir a compatibilidade com caracteres especiais da língua portuguesa.

2.2. Normalização de Formatos

- **Preços:** Os valores de preço foram convertidos para o tipo numérico (float) após a remoção de símbolos monetários e a substituição de vírgulas por pontos (para decimais).
- **Nomes e Descrições:** Textos foram convertidos para minúsculas para padronização e remoção de duplicatas sensíveis a maiúsculas/minúsculas.
- **Termos Técnicos:** Não foi necessária uma normalização específica de termos técnicos nesta fase, mas a extração de atributos (Seção 3.3) aborda a identificação desses termos.

2.3. Tratamento de Dados Ausentes e Duplicados

- **Dados Ausentes:** Valores `None` ou vazios foram tratados de acordo com o contexto de cada coluna. Para colunas numéricas, foi considerada a imputação (média, mediana) ou remoção de registros, dependendo da proporção de dados ausentes. Para colunas textuais, valores ausentes foram preenchidos com `"N/A"`.
- **Registros Duplicados:** Registros completamente duplicados foram removidos com base em um subconjunto de colunas chave (e.g., `title`, `price`, `seller_name`) para evitar redundância na base de dados.

3. Enriquecimento dos Dados com Colunas Derivadas

Esta fase focou na criação de novas informações a partir dos dados existentes, adicionando valor e profundidade à base de dados.

3.1. Criação de Novas Variáveis

- **`price_numeric`:** Versão numérica do preço, facilitando cálculos e análises quantitativas.
- **`reviews_numeric`:** Versão numérica da contagem de reviews.
- **`rating_numeric`:** Versão numérica da avaliação do produto.
- **`is_original`:** Coluna binária (True/False) indicando se o produto é "original" com base no título e descrição.
- **`is_xl`:** Coluna binária (True/False) indicando se o cartucho é "XL" (extra large).
- **`product_type`:** Classificação do produto (e.g., "cartucho preto", "cartucho colorido", "kit").

3.2. Classificação Automatizada

Foi implementada uma lógica de classificação baseada em palavras-chave para categorizar produtos como "original", "compatível" ou "suspeito". Esta classificação é crucial para identificar a autenticidade e a qualidade percebida dos produtos.

3.3. Extração de Atributos Técnicos

Expressões regulares foram utilizadas para extrair atributos técnicos específicos do título e da descrição dos produtos, como:

- **Modelo do Cartucho:** Ex: `664`, `122`.
- **Cor:** Ex: `Preto`, `Colorido`.
- **Capacidade (XL/Normal):** Identificação de `XL`.

3.4. Indicadores de Confiabilidade

Foram criados indicadores binários para a presença de termos que sugerem maior confiabilidade do produto ou do vendedor:

- **`has_sealed_info`:** Indica se a descrição menciona "lacrado".
- **`has_invoice_info`:** Indica se a descrição menciona "nota fiscal".
- **`has_warranty_info`:** Indica se a descrição menciona "garantia".

4. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) foi realizada para identificar padrões, anomalias e insights iniciais na base de dados enriquecida. Foram gerados gráficos e estatísticas descritivas.

4.1. Estatísticas Descritivas

- **Preço:** Média, mediana, desvio padrão, quartis e distribuição de preços para diferentes tipos de cartuchos.
- **Vendedores:** Contagem de produtos por vendedor, identificando os principais players.
- **Tipos de Produto:** Distribuição dos produtos por tipo (e.g., preto, colorido, kit).

4.2. Identificação de Outliers e Comportamentos Anormais

Gráficos de caixa (box plots) e histogramas foram utilizados para visualizar a distribuição de preços e avaliações, auxiliando na identificação de outliers que podem indicar erros de coleta ou produtos com características muito distintas.

4.3. Análise Lexical

Foi realizada uma análise de frequência de termos nas descrições dos produtos para identificar palavras-chave comuns e características importantes mencionadas pelos vendedores. Isso pode ser expandido para uma análise de tópicos mais complexa no futuro.

5. Identificação e Coleta de Dados de Nova Fonte

Para enriquecer ainda mais a base de dados, foi identificada uma nova fonte de dados: **Americanas.com.br**. A escolha da Americanas se justifica por ser uma das maiores plataformas de e-commerce no Brasil, oferecendo uma vasta gama de produtos HP e complementando os dados já coletados do Mercado Livre.

5.1. Justificativa da Escolha

- **Complementaridade:** A Americanas oferece uma perspectiva diferente de preços, vendedores e descrições de produtos, permitindo uma visão mais abrangente do mercado.
- **Volume de Dados:** A plataforma possui um grande volume de produtos, o que contribui para a robustez da análise.
- **Diversidade de Vendedores:** A Americanas agrega diversos vendedores, o que é valioso para a análise de concorrência e reputação.

5.2. Estratégia de Extração de Dados

A estratégia de extração para a Americanas utilizou o Selenium para simular a navegação de um usuário real, devido à natureza dinâmica do conteúdo da página. Os principais passos foram:

1. **Navegação:** Acessar a URL de busca para "cartucho hp 664" na Americanas.
2. **Aceite de Cookies:** Implementar lógica para aceitar o banner de cookies, se presente.
3. **Extração de Links de Produtos:** Coletar os URLs de cada produto listado na página de resultados.

4. **Extração de Detalhes do Produto:** Para cada URL de produto, navegar até a página de detalhes e extrair informações como título, preço, nome do vendedor, contagem de reviews, avaliação, descrição, especificações, imagens, disponibilidade e informações de frete.

5.3. Desafios e Soluções

- **Seletores Dinâmicos:** A Americanas utiliza `data-testid` para muitos elementos, o que torna os seletores mais estáveis do que classes ou IDs que podem mudar. No entanto, a estrutura HTML pode variar, exigindo ajustes nos seletores CSS.
- **Carregamento Dinâmico de Conteúdo:** A página carrega conteúdo dinamicamente (lazy loading), o que foi mitigado com o uso de `WebDriverWait` e `expected_conditions` para aguardar a presença dos elementos antes de tentar extraí-los.
- **Bloqueio de Bots:** Para evitar bloqueios, foram adicionados `user-agent` e `time.sleep()` entre as requisições para simular um comportamento mais humano.

6. Integração e Consolidação dos Dados

Após a coleta dos dados da Americanas (mockados para fins de demonstração, conforme solicitado), a próxima etapa foi integrar e consolidar essa nova fonte com a base de dados existente do Mercado Livre.

6.1. Padronização e Fusão

- **Padronização de Colunas:** As colunas de ambos os DataFrames (Mercado Livre e Americanas) foram padronizadas para garantir que tivessem os mesmos nomes e tipos de dados, facilitando a concatenação.
- **Concatenação:** Os DataFrames foram concatenados verticalmente (`pd.concat`), criando uma única base de dados abrangente.

6.2. Consolidação e Deduplicação

- **Deduplicação:** Registros duplicados foram identificados e removidos da base de dados consolidada. A deduplicação foi realizada com base em uma combinação de `title` e `price` para garantir que produtos idênticos de diferentes fontes não fossem contados múltiplas vezes. Em um cenário real, técnicas mais avançadas de correspondência (fuzzy matching) poderiam ser empregadas para identificar produtos semelhantes com pequenas variações no título ou descrição.

7. Conclusão e Próximos Passos

Este entregável demonstrou a capacidade de coletar, limpar, enriquecer e integrar dados de múltiplas fontes de e-commerce. A base de dados consolidada e enriquecida é um ativo valioso para as próximas fases do projeto, permitindo análises mais complexas e o desenvolvimento de modelos de PLN e Machine Learning.

Próximos Passos:

- **Análise de Sentimento:** Aplicar técnicas de PLN para analisar o sentimento das avaliações dos produtos.
- **Modelagem Preditiva:** Desenvolver modelos para prever preços, demanda ou identificar produtos com alto potencial de venda.
- **Visualizações Avançadas:** Criar dashboards interativos para explorar os dados de forma mais dinâmica.
- **Expansão de Fontes:** Integrar dados de outras plataformas de e-commerce ou APIs de fabricantes.