Documentação - Trabalho Prático - Parte 1

Contexto e Problema

O sistema desenvolvido tem como objetivo coletar, organizar e facilitar a análise comparativa de passagens aéreas domésticas no Brasil. O principal desafio é a complexidade em avaliar múltiplos critérios simultaneamente (preço, duração, escalas, horários e companhias aéreas) diante das inúmeras combinações de rotas e datas disponíveis.

Solução Proposta

Inicialmente o sistema automatiza a coleta de estruturada de dados do Google Voos organizando as informações por rotas (origem-destino) e datas. Na etapa atual, extrai detalhes como preços, companhias aéreas, horários e escalas, viabilizando análises futuras através de comparativa que permitirá filtrar e classificar os voos com base em critérios combinados como:

- Melhor custo-benefício por companhia.
- Voos diretos com partida após X horas.
- Voos mais baratos com duração máxima de 4h.

Assim depois a visualização intuitiva com apresentação comparativa dos resultados destacando:

- Comparação entre preço e conveniência.
- Padrões de preços por rota/período.
- Opções ideais para diferentes perfis de passageiros.

Ao cruzar automaticamente os parâmetros decisórios, o sistema elimina a necessidade de comparação manual, oferecendo uma visão integrada que otimiza tempo e qualidade na escolha de voos.

Justificativa

A escolha do Google Voos como fonte de dados se deve à sua abrangência e confiabilidade. A abordagem sistemática de combinações garante cobertura completa do espaço de busca, enquanto o armazenamento estruturado facilita a posterior indexação e recuperação.

Tipo do Coletor

- **Coletor Vertical:** Especializado em extrair dados específicos de voos de páginas de um domínio específico, no caso, o Google Voos.
- **Baseado em Navegador:** Utiliza Selenium WebDriver para simular interação humana.
- **Incremental:** embora não implemente uma revisitação, a estrutura permite execução periódica para atualização dos dados coletados.

Propriedades de Coleta - Qualidade

- **Acurácia:** busca exatamente os dados certos de voos domésticos entre aeroportos do Brasil.
- **Cobertura:** lista abrangente de todos os 51 aeroportos brasileiros domésticos, garantindo que todas as combinações origem/destino sejam contempladas.

Tolerâncias Implementadas

- **Tentativas de repetição:** Até 2 tentativas por coleta
- **Pausas estratégicas:** Entre 2-5 segundos entre requisições
- **Limite de erros:** Máximo de 30 erros consecutivos antes de parar
- **Reinício do navegador:** Em caso de falhas persistentes

Critério de Parada

- Completa todas as combinações de aeroportos e datas programadas
- Atinge o limite máximo de erros consecutivos (30)

Políticas de Abordadas

- **Política de Seleção:** Coletor gera combinações de origem e destino a partir de uma lista fixa (lista de aeroportos brasileiros), sendo uma seleção de documentos por critério temático, pois a navegação é simulada via Selenium.
- **Política de Boas Maneiras:** Delay aleatório entre requisições evitando sobrecarga, Coleta apenas de dados publicamente disponíveis, Limite de tentativas e tolerância de erros.

Decisões de Projeto Justificadas

- **Uso do Selenium:** Necessário para interagir com aplicação JavaScript pesada como Google Voos.
- **Combinações de Aeroportos:** Permite cobertura completa do mercado doméstico.
- **Armazenamento de HTML Limpo:** Reduz tamanho dos arquivos mantendo apenas conteúdo relevante, mantendo máximo possível da estrutura original.

- **Logs Detalhados:** Fundamental para monitorar execuções longas e diagnosticar problemas.

Escala da Coleta

- **Aeroportos:** 51 aeroportos brasileiros (códigos IATA)
- **Rotas por aeroporto:** 50 destinos × 7 dias = 350 combinações por aeroporto.
- **Combinações**: Todas as permutações (51×50 = 2.550 rotas)
- **Período**: 7 dias de coleta a partir de 01/07/2025
- **Total:** 2550 x 7 dias = 17.850 Páginas Coletadas (Registrada)

Desempenho Real do Coletor (Uma Execução)

- **Tempo total de execução:** 4.769,30 minutos (~79,5 horas) (~3,3 dias)
- **Taxa de sucesso:** 100% (zero erros registrados)
- **Volume processado:** 17.850 páginas HTML
- **Total de Dados Armazenados:** 25,4 GB arquivos organizados

Eficiência do Coletor:

- **Taxa de coleta:** ~3,75 páginas/minuto.
- **Armazenamento:** ~498MB Por Aeroporto & ~1,42MB por página (em média).
- **Confiabilidade:** 100% de sucesso nas requisições.