Extração Automática de Dados

Otávio Calaça Xavier

otaviocx

otaviocx@ufg.br

INSTITUTO DE INFORMÁTICA
UFG

Projetos Finais

INSTITUTO DE
INFORMÁTICA
UFG

Projetos Finais - Visão Geral

- Objetivo central: **construir um pipeline completo de extração automatizada de dados**, a partir de fonte(s) públicas e dados abertos, demonstrando desde a coleta automatizada até a comunicação dos achados em formato acadêmico.
- Os temas sugeridos são apenas ponto de partida; propostas autorais são bem-vindas, desde que incluem extração automática de dados.

Projetos Finais - Entregas

- **Repositório de Código:**
  - Código da coleta (scrapers, crawlers, scripts de coleta em APIs e dados estruturados, etc.) e notebooks/ETL.
  - **README** com: descrição do projeto, instruções de execução, diagrama resumido do pipeline, dependências, etc.
  - Licença de uso e menção às fontes.
  - Adicionar meu usuário (**otaviocx**) ao repositório se for privado.

Projetos Finais - Entregas

- **Conjunto de Dados (Dataset):**
  - Dataset limpo em formato estruturado (CSV, Parquet, JSONL, etc.)
  - Entregue via link no **README**.
- **Relatório Técnico:**
  - Usar linguagem formal/científica.
  - Preferencialmente usar Latex
  - Template da SBC, ACM ou IEEE, por exemplo.

Projetos Finais - Entregas

- **Artigo científico curto (6–10 páginas).**
- **Relatório Técnico:**
  - Introdução & motivação.
  - Fundamentação teórica (fontes, trabalhos correlatos).
  - Método (detalhamento das fontes, arquitetura de coleta e integração).
  - Resultados & discussões.
  - Reflexões éticas e limitações.
  - Conclusão & possíveis trabalhos futuros.

Projetos Finais - Entregas

- **Apresentação Oral**
  - Tempo: 7 a 10 minutos, totalizando no máximo **15 minutos com perguntas**.
  - Conteúdo mínimo:
    - Problema & importância.
    - Arquitetura do pipeline (extração ➜ engenharia ➜ análise).
    - Principais descobertas (gráficos/insights).
    - Reflexões éticas e legais (LGPD, direitos autorais, robots.txt).

Projetos Finais - Requisitos Técnicos

- **Extração automática**: ao menos um componente de *scraping/crawling* (páginas HTML dinâmicas, RSS, XML, PDFs, etc.). Pode combinar-se com APIs formais.
- **Engenharia de dados**:
  - Armazenar dados brutos e dados tratados (camadas "raw" e "clean").
  - Scripts/notebooks de transformação reproduzíveis.
  - Documentar formato de saída.
- **Reprodutibilidade**: instruções claras (README) para rodar o pipeline em outro ambiente.

Projetos Finais - Ética & Conformidade Legal

- Respeitar **robots.txt**, limites de requisição e termos de serviço.
- Enfatizar anonimização quando dados pessoais forem coletados (LGPD).
- Citar licenças de datasets ou APIs utilizadas.

Projetos Finais - Sugestões de Temas

- **Preço da passagem × lotação de voos**
  - **Scraping**: preços de voo nos sites da LATAM, GOL ou Azul
    - ex.: https://www.latamairlines.com/br/pt
  - **API/CSV**: dados “Demanda e Oferta” da ANAC
    - CSV mensal – https://www.gov.br/anac/pt-br/dadosabertos
  - **Integração/Análises**: cruzar por rota + mês;
    - ex.: verificar se promoções coincidem com voos historicamente vazios.