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.

Projetos Finais - Sugestões de Temas

- **Cesta básica em Goiás × IPCA**
  - **Scraping:** relatórios mensais do PROCON Goiás
    - PDF/HTML – https://www.goias.gov.br/procon/
  - **API/CSV:** SIDRA/IBGE
    - IPCA mensal – https://sidra.ibge.gov.br
  - **Integração/Análises:** correlação e defasagem entre variação local e inflação oficial.

Projetos Finais - Sugestões de Temas

- **Status de obras públicas**
  - **Scraping:** avisos de licitação no Diário Oficial do Estado de Goiás
    - PDF – https://www.doe.go.gov.br
  - **API/CSV:** Portal da Transparência GO
    - pagamentos/desembolsos – https://www.transparencia.go.gov.br
  - **Integração/Análises:** ex.: ligar nº do processo licitatório aos pagamentos; calcular % executado vs. prazo.

Projetos Finais - Sugestões de Temas

- **Concursos públicos × mercado de trabalho**
  - **Scraping:** listagem de editais em PCI Concursos (https://www.pciconcursos.com.br/) ou Estratégia.
  - **API/CSV:** CAGED Novo (admissões/desligamentos por CBO - https://pdet.mte.gov.br).
  - **Integração/Análises:** comparar oferta de vagas públicas com variação de empregos privados na mesma área.

Projetos Finais - Sugestões de Temas

- **Streaming buzz (JustWatch × Twitter/X)**
  - **Scraping:** ranking semanal do JustWatch
    - https://www.justwatch.com/br
  - **API:** Twitter/X API v2 para contagem de menções por título
    - https://developer.x.com
  - **Integração/Análises:** medir se pico de tweets antecede (ou segue) subida no ranking de streaming, por exemplo.

Projetos Finais - Sugestões de Temas

- **Imóveis em Goiânia × renda do bairro**
  - **Scraping:** anúncios em VivaReal (https://www.vivareal.com.br), Zap Imóveis ou OLX filtrados para Goiânia e região.
  - **API/CSV:** malha + rendimento domiciliar (IBGE Censo 2022).
  - **Integração/Análises:** junção geográfica; regressão preço m^2 ~ renda + distância do centro, por exemplo.

Projetos Finais - Sugestões de Temas

- **Combustível × dólar**
  - **Scraping:** tabela semanal de preços da ANP
    - HTML – https://www.gov.br/anp
  - **API/CSV:** BACEN “Cotação diária USD/BRL”
    - https://dadosabertos.bcb.gov.br
  - **Integração/Análises:** elasticidade do preço-bomba ao câmbio; defasagem temporal.

Projetos Finais - Sugestões de Temas

- **Agenda cultural × previsão do tempo**
  - **Scraping:** eventos Sympla filtrados para Goiânia (https://www.sympla.com.br/eventos).
  - **API/CSV:** previsão de 5 dias INMET/BDMEP (https://bdmep.inmet.gov.br) ou Open-Meteo.
  - **Integração/Análises:** classificar eventos indoor/outdoor; alertas sobre chuva/temperatura.

Projetos Finais - Sugestões de Temas

- **Vagas remotas & skills emergentes**
  - **Scraping:** RemoteOK (https://remoteok.io)
  - **API/CSV:** GitHub ou Stackoverflow Techs Survey
    - Ex.: https://survey.stackoverflow.co/
  - **Integração/Análises:** frequência de skills em vagas × popularidade em repositórios/posts; radar de demanda.

Projetos Finais - Sugestões de Temas

- **Qualidade da água tratada × doenças de veiculação hídrica**
  - **Scraping:** Relatórios anuais de qualidade da água – Saneago
  - **API/CSV:** DATASUS/SINAN – doenças de veiculação hídrica via TabNet https://datasus.saude.gov.br/informacoes-de-saude-tabnet/
  - **Integração/Análises:** Correlação entre parâmetros fora do padrão e incidência de diarreia, hepatite A, etc.

Projetos Finais - Sugestões de Temas

- **Internações respiratórias × cobertura vacinal (Saúde)**
  - **Scraping:** Boletins SRAG – Secretaria de Saúde de Goiás
  - **API/CSV:** OpenDataSUS – SI-PNI / Campanha de Vacinação
    - https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao
    - https://dados.gov.br/dados/conjuntos-dados/covid-19-vacinacao1
  - **Integração/Análises:** Analisar correlação entre avanço da vacina e queda de internações.

INSTITUTO DE INFORMÁTICA
UFG