

Notebook de Download de Sentenças (TJMA) — Relato Técnico

Objetivo: Descrever a estratégia, arquitetura e execução do notebook responsável por coletar sentenças do Tribunal de Justiça do Maranhão (TJMA) via PJe/Jurisconsult, gerando artefatos (HTML, CSV, JSON) para posterior análise de NLP/Deep Learning.

1. Stack e Dependências

- Linguagem/ambiente: Python + Jupyter Notebook.
- Bibliotecas: Selenium/WebDriver, webdriver-manager, BeautifulSoup (bs4), pandas, requests, lxml, tqdm/logging.
- Navegador/driver: Google Chrome + ChromeDriver (instalação automática pelo webdriver-manager).
- URLs alvo (consulta pública): PJe TJMA — listView.seam e páginas de detalhe (popup documento HTML).

2. Arquitetura e Fluxo de Coleta

- Configuração do WebDriver com opções de desempenho, agente de usuário e diretório de download (headless opcional).
- Busca por número do processo (CNJ) na página de consulta; envio do formulário e captura do link de detalhes (popup).
- Extração de documentos na aba de detalhes: identificação da tabela 'processoDocumentoGridTab' e do tbody ':tb'.
- Identificação de links para documentos pelo 'onclick' contendo 'documentoSemLoginHTML'; normalização de texto e data/horário.
- Classificação básica por tipo (sentença/decisão/outros) e filtro opcional para baixar somente sentenças.
- Persistência: salvamento do HTML do documento, metadados em CSV e JSON consolidando resultados por processo.

3. Funções-Chave Implementadas

- configurar_driver(headless, download_dir): inicializa Chrome com opções (desabilita extensões de automação, define user-agent, timeouts).
- buscar_processo(driver, numero_processo): preenche o campo de busca, aciona 'Pesquisar' e extrai URL de detalhes via expressão regular em 'onclick'.
- extrair_documentos(driver, url_detalhes, numero_processo): encontra a tabela 'processoDocumentoGridTab', varre linhas e extrai URLs e metadados de documentos.
- baixar_documento(driver, documento, numero_processo): acessa a URL do documento e salva o HTML no diretório de saída com convenção de nomes.
- processar_processo_completo(numero_processo, headless, apenas_sentencas): pipeline fim-a-fim (busca, extração, download, salvamento de CSV/JSON).

4. Convenções de Saída e Estrutura de Pastas

- Estrutura: sentenças_tjma/pdfs, sentenças_tjma/html, sentenças_tjma/metadados.
- Nome dos arquivos HTML baixados: _____.html (ex.: 08001517120218100056_000_22-09-2025_16-18-06_Sentença.html).
- Metadados por processo: CSV (_metadados.csv) e JSON (_completo.json).
- Relatório agregado: sentenças_tjma/metadados/relatorio_.csv com sucesso/erro e contagem de baixados.

5. Execução (Exemplo) e KPIs

Processos executados	9
----------------------	---

Sucessos	9
Falhas	0
Documentos baixados (sentenças)	10

6. Robustez, Boas Práticas e Limitações

- Paginação e intervalos (sleep) entre operações para evitar throttling/rate limit.
- Tratamento de erros (TimeoutException, NoSuchElementException) com logs e fallback seguro.
- Validação de elementos por IDs específicos e uso de regex no 'onclick' para capturar popups de documentos.
- Limitações: apenas processos públicos; dependência de disponibilidade do PJe/Jurisconsult; documentos em HTML podem demandar conversão adicional.

7. Conformidade e Referências Institucionais

- Coleta restrita a dados públicos e respeito aos termos de uso do portal TJMA/Jurisconsult.
- Alinhamento com princípios de publicidade processual (ex.: diretrizes do CNJ).
- Sugestão de citação institucional e registro da data de acesso no relatório/paper.

8. Integração com o Pipeline Analítico

- O CSV gerado alimenta a rotulagem heurística (procedente/improcedente/neutro).
- Os HTMLs baixados fornecem o dispositivo para validação de rótulos e evidências.
- Os datasets resultantes são utilizados no fine-tuning (BERTimbau) e na auditoria de fairness por gênero do julgador.

Observações Finais

Este relato técnico sintetiza a implementação e execução do notebook de coleta do TJMA (Usucapião). Para publicação acadêmica, recomenda-se anexar capturas de tela das páginas alvo, exemplos de logs, parâmetros de execução e o relatório de volume/tempo de coleta.