Notebook de Extração de Sentenças (TJMA v3) — Relato Técnico

Objetivo: Extrair informações estruturadas de sentenças judiciais do TJMA a partir de arquivos HTML previamente baixados, garantindo robustez contra variações de encoding e padrões textuais. O resultado é um arquivo CSV consolidado para análises de NLP e fairness.

1. Principais Melhorias da Versão v3

- Uso de UnicodeDammit para detecção de encoding e fallback para latin-1.
- Busca precisa de cargo/gênero no entorno da assinatura eletrônica.
- Reconhecimento ampliado de padrões decisórios (REJEITO A INICIAL, JULGO/EXTINGO SEM/COM RESOLUÇÃO DO MÉRITO, DEFIRO/INDEFIRO, HOMOLOGO).
- Captura do número do processo em diferentes formatos (PROCESSO Nº, Processo:).
- Suporte a cabeçalhos variados para vara/comarca.

2. Arquitetura e Funções-Chave

- read_html_robust(path): leitura binária + normalização de encoding.
- _find_vara_comarca(text): regex para extrair vara e comarca.
- parse_html(path): extrai processo_numero, vara, comarca, magistrado_nome, magistrado_genero, magistrado_cargo, assinatura_datahora, assinatura_id_documento, decisao.
- Pipeline em lote: itera sobre arquivos .html e salva CSV consolidado.

3. Estrutura de Saída

- Arquivo CSV consolidado: resultado_tjma_v3.csv.
- Campos: arquivo, processo_numero, vara, comarca, magistrado_nome, magistrado_genero, magistrado_cargo, assinatura_datahora, assinatura_id_documento, decisao.

4. Robustez e Boas Práticas

- Tratamento de erros por arquivo com fallback seguro.
- Regex específicas para evitar confusão com nomes de fórum.
- Normalização de encoding para evitar falhas em HTMLs heterogêneos.

5. Integração com Pipeline Analítico

- CSV gerado alimenta rotulagem heurística (procedente/improcedente/neutro).
- HTMLs fornecem dispositivo para validação de rótulos e evidências.
- Datasets resultantes usados no fine-tuning (BERTimbau) e auditoria de fairness por gênero do julgador.

Observações Finais

Este relato técnico sintetiza a implementação e execução do notebook de extração TJMA v3. Recomenda-se anexar exemplos de código, parâmetros de execução e métricas de volume/tempo para publicação acadêmica.