Relatório de Prova de Conceito: Digitalização Segura de Registros Históricos com IA Local

Proposta do Projeto

Vastos acervos da história humana — registros de nascimentos, casamentos, óbitos, censos e outros documentos civis — estão atualmente "presos" em formato físico ou em imagens digitalizadas não pesquisáveis. Esses documentos são a matéria-prima para pesquisas em genealogia, história social, demografia e muitas outras áreas. No entanto, o acesso a eles é extremamente limitado pelo maior gargalo existente: a transcrição manual. Este processo é lento, caro, propenso a erros e inviável em grande escala.

O objetivo deste projeto é desenvolver e treinar um modelo de Inteligência Artificial para a tarefa de Reconhecimento de Texto Manuscrito (HTR - Handwritten Text Recognition) como prova de conceito da possibilidade. A finalidade é criar uma ferramenta capaz de ler e transcrever automaticamente o conteúdo de documentos históricos, como registros de batismo, casamento e óbito. A automação desse processo visa acelerar a digitalização e a pesquisa de vastos arquivos genealógicos, tornando-os acessíveis de forma segura e eficiente.

Dentre a relevância do desenvolvimento de um projeto de maior escala sobre essa POC, observa-se:

- **Democratização do Acesso:** Ao automatizar a transcrição, removemos a barreira física e de custo para pesquisadores, genealogistas, estudantes e o público geral. Qualquer pessoa com acesso à internet poderá pesquisar nomes, datas e locais em documentos que antes exigiriam viagens a arquivos e horas de leitura minuciosa.
- **Aceleração da Pesquisa Acadêmica:** A capacidade de converter grandes volumes de registros históricos em texto pesquisável abre portas para análises de big data do passado. Historiadores e sociólogos poderão identificar padrões de migração, tendências de saúde pública, redes sociais e mudanças demográficas em uma escala e velocidade antes inimagináveis.
- **Preservação do Patrimônio Cultural:** Documentos físicos se degradam com o tempo. A criação de transcrições digitais precisas e associadas às suas imagens originais garante a preservação permanente deste patrimônio, protegendo-o contra incêndios, inundações e a simples deterioração pelo tempo.
- **Eficiência e Escalabilidade:** Uma vez treinado, um modelo de IA pode transcrever documentos 24 horas por dia, 7 dias por semana, a uma fração do custo e do tempo

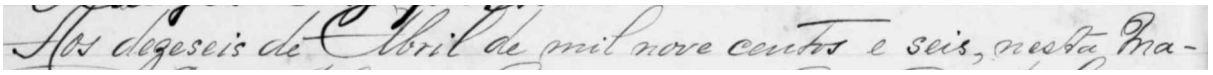
de um esforço humano. Isso torna viável a digitalização de arquivos inteiros, não apenas de coleções selecionadas.

Formato e Criação do Dataset

O conjunto de dados (dataset) utilizado para o treinamento foi criado de forma artesanal, garantindo alta fidelidade entre a imagem e o texto correspondente. O processo consistiu nas seguintes etapas, de forma generalizada:

1. **Obtenção de dados não processados:** Foram utilizados imagens provenientes do FamilySearch, como livros de registro de batizados, certidões de nascimento e óbito
2. **Transcrição Manual:** O conteúdo do documento foi cuidadosamente transcrito à mão para garantir a máxima precisão no texto de referência. Esse tipo de conteúdo de nicho, documentos históricos em portugues, são escassos. Como a qualidade do modelo depende diretamente da entrada dos dados, esse é talvez o processo mais importante. Ele é afetado por vieses, como a impossibilidade de transcrever alguns documentos por alta dificuldade de leitura e, consequentemente, não processamento pelo modelo final.
3. **Segmentação:** A imagem original foi segmentada, sendo "quebrada" em imagens menores, cada uma correspondendo a uma única linha de texto. Esse processo é altamente paralelizável, ainda que feito manualmente.
4. **Associação:** Foi criado um arquivo de metadados (`.csv`) que associa o nome de cada arquivo de imagem de linha (ex: `pagina150_rosalina_linha01.jpg`) com a sua respectiva transcrição textual.

Rozalina



Aos dezesseis de Abril de mil nove centos e seis, nesta ma-

Considerações de Segurança: Processamento Local

Um requisito fundamental deste projeto foi garantir a segurança e a privacidade dos dados históricos, que podem conter informações sensíveis. Para atender a essa questão de segurança, todo o ciclo de vida do modelo foi executado em um ambiente local (*on-premises*).

Isso significa que nenhuma informação — nem as imagens dos documentos, nem as transcrições — foi enviada para serviços de processamento em nuvem ou APIs de terceiros. Todo o treinamento e a inferência do modelo ocorreram na máquina local, o que garante a integridade total e a confidencialidade dos dados, mitigando riscos de vazamento ou acesso não autorizado durante o tráfego de rede.

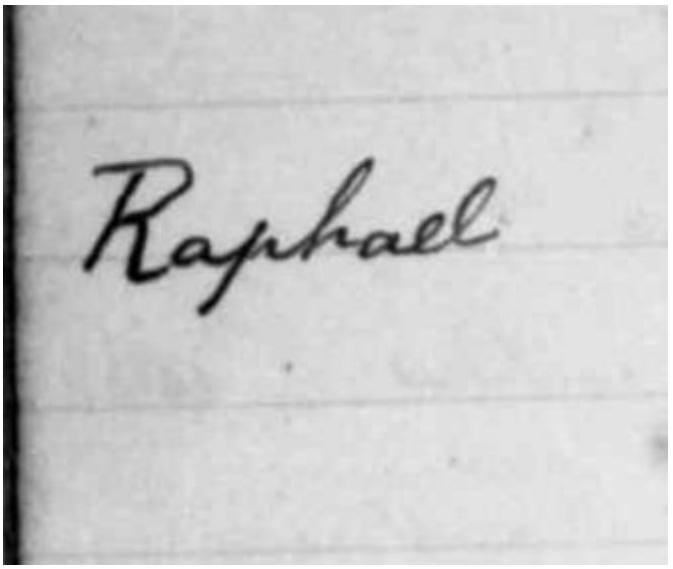
Resultados e Prova de Conceito

O modelo de visão computacional, baseado na arquitetura `microsoft/trocr-base-handwritten`, foi treinado localmente utilizando o dataset descrito.

Os resultados iniciais indicam que, devido ao baixo volume de dados de treinamento, a precisão geral do modelo em documentos completamente novos ainda é limitada. No entanto, o projeto serve como uma **prova de conceito bem-sucedida**, demonstrando que a abordagem é viável e segura.

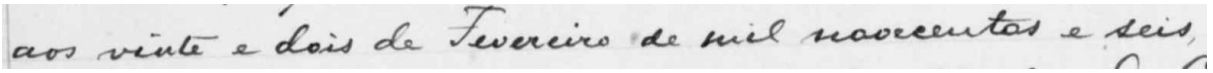
Para ilustrar o potencial do modelo treinado, apresentamos um exemplo favorável de inferência:

- **Imagem de Entrada:**



- **Transcrição Gerada pelo Modelo:**
  “Ail mil”

- **Imagem de Entrada:**



- **Ada debr debr debr debr de no cent ein.asc**

  Relevantes -> Ada … de no cent

Conclusão e Melhorias Futuras

Este projeto demonstrou com sucesso a viabilidade de utilizar um modelo de Inteligência Artificial para o Reconhecimento de Texto Manuscrito (HTR) em documentos históricos. A abordagem de treinar e executar o modelo em um ambiente totalmente local foi um pilar fundamental, garantindo a integridade, a privacidade e a segurança dos dados, um requisito crucial para a manipulação de registros que podem conter informações sensíveis.

Apesar do conjunto de dados limitado, o modelo treinado provou ser uma **prova de conceito eficaz**, conseguindo transcrever corretamente alguns segmentos de texto para os quais foi treinado. Isso valida que a arquitetura TrOCR é adequada para a tarefa e que, com mais dados, a precisão pode ser significativamente aprimorada. O projeto não apenas atingiu seu objetivo técnico, mas também reforçou a relevância de se criar soluções de IA que devolvam o controle dos dados ao usuário, mitigando os riscos associados ao processamento em nuvem.

Melhorias Futuras

O sucesso desta fase inicial abre um vasto leque de oportunidades para aprimoramento e expansão. As seguintes melhorias são propostas para as próximas fases do projeto:

1. **Expansão Massiva do Dataset:** A limitação mais significativa do modelo atual é o baixo volume de dados. A próxima etapa deve focar em coletar e transcrever um volume muito maior de imagens, abrangendo centenas ou milhares de páginas de diferentes livros e períodos.
2. **Diversificação dos Dados:** Para criar um modelo mais robusto e generalista, é crucial incluir uma maior variedade de caligrafias, diferentes tipos de documentos (casamentos, óbitos, testamentos), e documentos com diferentes estados de conservação (manchas, rasgos, desbotamento).
3. **Técnicas de Aumento de Dados (Data Augmentation):** Aplicar transformações artificiais às imagens existentes (como pequenas rotações, variações de brilho e contraste) para simular novas amostras e aumentar a capacidade do modelo de generalizar para documentos nunca vistos.
4. **Otimização de Hiperparâmetros e Modelos:** Experimentar com diferentes arquiteturas de modelos de HTR e realizar um ajuste fino dos hiperparâmetros (como taxa de aprendizado e tamanho do lote) para maximizar a performance com o dataset expandido.