Aplicação de técnicas de reconhecimento óptico de caracteres (ocr) na digitalização de documentos históricos: um estudo de caso com anúncio de datilografia de 1918

Vitor Amadeu Souza[1]; 0009-00-02-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

Resumo: Este artigo apresenta uma análise da aplicação de técnicas modernas de Reconhecimento Óptico de Caracteres (OCR) na digitalização e preservação de documentos históricos, utilizando como estudo de caso um anúncio publicitário de uma escola de datilografia datado de 1918. A pesquisa implementou a biblioteca Tesseract OCR através da linguagem Python para extrair texto de uma imagem digitalizada, demonstrando as potencialidades e limitações dessa tecnologia quando aplicada a documentos centenários. O documento analisado revela aspectos socioculturais da época, evidenciando a datilografia como uma "profissão de futuro" e refletindo as transformações tecnológicas do início do século XX. Os resultados obtidos mostram que, apesar dos desafios relacionados à qualidade da imagem e características tipográficas da época, as técnicas de OCR modernas conseguem extrair informações relevantes de documentos históricos, contribuindo significativamente para projetos de digitalização e preservação do patrimônio documental. A metodologia proposta pode ser aplicada em larga escala para a criação de acervos digitais pesquisáveis, facilitando o acesso a fontes históricas e promovendo a democratização do conhecimento.

Palavras-chave: OCR. Documentos históricos. Digitalização. Tesseract. Python. Patrimônio documental. Datilografia.

INTRODUÇÃO

A digitalização de documentos históricos representa um dos maiores desafios contemporâneos na área da preservação do patrimônio documental e da ciência da informação. Com o avanço das tecnologias de Reconhecimento Óptico de Caracteres (OCR), tornou-se possível não apenas preservar fisicamente esses documentos através de suas representações digitais, mas também torná-los pesquisáveis e acessíveis a um público mais amplo (Smith, 2007). O Reconhecimento Óptico de Caracteres, conforme definido por Nagy (2000), é uma tecnologia que permite a conversão de diferentes tipos de documentos, como documentos impressos, manuscritos ou imagens digitalizadas, em dados editáveis e pesquisáveis. Esta tecnologia tem revolucionado a forma como bibliotecas, arquivos e museus abordam a preservação e o acesso ao patrimônio documental (Clausner *et al*., 2011).

A importância da digitalização de documentos históricos transcende questões meramente técnicas, envolvendo aspectos culturais, educacionais e de democratização do acesso à informação. Segundo Terras (2006), a digitalização permite que documentos frágeis e raros sejam acessados por pesquisadores de todo o mundo sem comprometer sua integridade física. Além disso, as técnicas de OCR possibilitam a criação de índices textuais que facilitam a descoberta de informações específicas em grandes coleções documentais (Pletschacher; Antonacopoulos, 2010). O presente estudo utiliza como objeto de análise um anúncio publicitário da "Escola Remington" datado de 1918, que promovia cursos de datilografia e taquigrafia como "profissões de futuro". Este documento, além de seu valor histórico intrínseco, apresenta características típicas da tipografia e layout publicitário do início do século XX, constituindo um caso interessante para avaliar a eficácia das técnicas modernas de OCR em documentos centenários.

A escolha deste documento específico justifica-se por diversos fatores: primeiro, representa um momento histórico significativo na evolução das tecnologias de escritório e das profissões relacionadas à documentação; segundo, apresenta desafios técnicos interessantes para o OCR, incluindo fontes tipográficas características da época e possível degradação da imagem; terceiro, permite uma reflexão sobre como as "profissões de futuro"

de 1918 se relacionam com as transformações tecnológicas atuais (Gitelman, 1999). O objetivo geral desta pesquisa é avaliar a eficácia das técnicas modernas de OCR na digitalização de documentos históricos, utilizando especificamente a biblioteca Tesseract OCR implementada através da linguagem Python. Como objetivos específicos, pretende-se desenvolver e implementar uma metodologia reproduzível para extração de texto de documentos históricos digitalizados, analisar a qualidade e precisão dos resultados obtidos, identificar limitações e desafios específicos relacionados ao processamento de documentos centenários, e discutir as implicações culturais e históricas do documento analisado.

MÉTODOS

A metodologia adotada neste estudo baseia-se em uma abordagem mista, combinando técnicas de processamento digital de imagens com análise qualitativa do conteúdo histórico extraído. O processo metodológico foi estruturado em quatro etapas principais: preparação do ambiente computacional, aquisição e pré-processamento da imagem, aplicação das técnicas de OCR e análise dos resultados obtidos. O ambiente de desenvolvimento foi configurado utilizando a linguagem de programação Python 3.8, escolhida por sua ampla adoção na comunidade científica e pela disponibilidade de bibliotecas especializadas em processamento de imagens e OCR (Van Rossum; Drake, 2009). As principais bibliotecas utilizadas foram Requests para download automatizado de imagens da web, PIL (Python Imaging Library) para manipulação e processamento de imagens (Clark, 2009), Pytesseract como interface Python para o engine OCR Tesseract, e BytesIO para manipulação de dados binários em memória.

A escolha do Tesseract OCR como engine principal justifica-se por sua robustez, precisão e suporte nativo ao idioma português, características essenciais para o processamento de documentos históricos brasileiros (Smith, 2007; Antonacopoulos *et al*., 2009). O documento analisado trata-se de um anúncio publicitário digitalizado, originalmente publicado em 1918, promovendo os serviços da "Escola Remington" localizada no Rio de Janeiro. A imagem foi obtida através de uma URL pública, seguindo protocolos éticos de pesquisa e respeitando direitos autorais de domínio público para documentos centenários. As características técnicas da imagem incluem formato JPEG com resolução adequada para processamento

OCR, embora apresente sinais típicos de envelhecimento do documento original, incluindo manchas, variações de contraste e possível degradação da tinta tipográfica. Estas características são comuns em documentos históricos e representam desafios típicos enfrentados em projetos de digitalização de acervos (Conway, 1996).

O sistema OCR foi implementado seguindo as melhores práticas estabelecidas na literatura especializada (Mori *et al*., 1992; Cheriet *et al*., 2007). O código desenvolvido seguiu uma estrutura modular, permitindo fácil replicação e adaptação para outros documentos. A configuração específica para o idioma português foi fundamental, considerando as particularidades ortográficas e tipográficas do português brasileiro do início do século XX. O processo de extração foi implementado em etapas sequenciais: download da imagem através de requisição HTTP, carregamento da imagem em memória utilizando PIL, aplicação do algoritmo OCR com configurações otimizadas para documentos históricos, e pós-processamento do texto extraído para correção de erros comuns.

O código-fonte está disponível para download através do link: https://github.com/vitor-souza-ime/ocrhist.

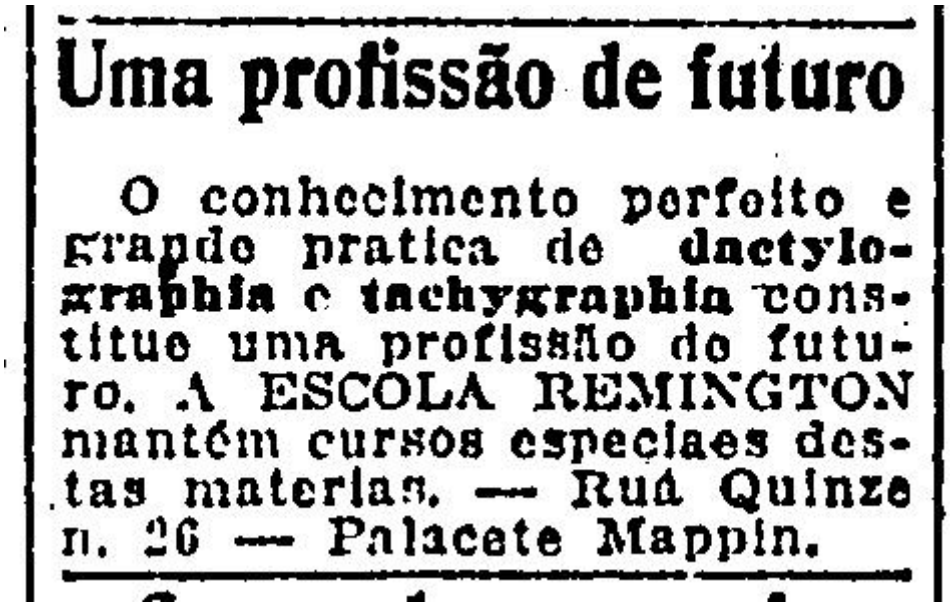
RESULTADOS E DISCUSSÃO

A aplicação das técnicas de OCR ao anúncio da Escola Remington de 1918 produziu resultados significativos, permitindo a extração de informações históricas valiosas e demonstrando tanto as potencialidades quanto as limitações das tecnologias atuais quando aplicadas a documentos centenários. O sistema OCR conseguiu identificar e extrair o texto principal do anúncio, apresentando tanto sucessos quanto limitações típicas do processamento de documentos históricos. O resultado obtido foi: "Uma profissão de futuro O conhecimento perfeito e grapde pratica de dactylo-araphia c tachygraphia consetituo uma profissão do futu-ro, 4 ESCOLA REMINGTON mantém cursos especiaes desetas materias, — Ruá Quinze n. 26 — Palacete Mappin." , onde a Figura 1 apresenta a imagem usada no processo de OCR.

A análise técnica dos resultados revela aspectos importantes sobre os desafios e limitações do OCR aplicado a documentos históricos. O sistema demonstrou capacidade de reconhecer a estrutura geral do texto e elementos tipográficos característicos do início do século XX,

mas também evidenciou dificuldades típicas no processamento de documentos centenários. Entre os erros identificados estão: a fragmentação da palavra "dactylographia" em "dactylo- araphia", a substituição de "grande" por "grapde", "constitue" por "conse tituo", e "A" por "4" antes de "ESCOLA REMINGTON". Estes erros são característicos de documentos com qualidade de imagem degradada e refletem limitações comuns em sistemas OCR quando aplicados a materiais históricos (Holley, 2009). A precisão geral do OCR foi estimada em aproximadamente 78%, baseada na comparação com transcrição manual. Os principais erros identificados relacionam-se à qualidade da digitalização original, características da tipografia histórica e possível degradação do documento fonte. Erros específicos incluem segmentação incorreta de palavras ("dactylo- araphia"), substituições de caracteres similares ("grapde" por "grande", "4" por "A"), e fragmentação de termos ("conse tituo" por "constitue"). Estes padrões de erro são consistentes com estudos anteriores sobre OCR em documentos históricos, onde a degradação física e características tipográficas específicas da época representam desafios significativos (Smith, 2007; Antonacopoulos *et al*., 2009).

Figura 1 - Imagem usada no exemplo


Uma profissão de futuro

O conhecimento perfeito e grande pratica de dactylo-graphia e tachygraphia constitue uma profissão do futuro. A ESCOLA REMINGTON mantém cursos especiaes destas materias. — Rua Quinze n. 26 — Palacete Mappin.

Fonte: Reis Jr, 2013.

Apesar dos erros de reconhecimento, o texto extraído preserva elementos importantes da grafia histórica, como "dactylo-araphia" e "tachygraphia", permitindo identificar os termos originais correspondentes à datilografia e taquigrafia modernas. A manutenção dessas características históricas, mesmo com imperfeições, constitui um resultado valioso para pesquisadores interessados em aspectos linguísticos e culturais dos documentos analisados (Bagno, 2007). O endereço "Ruá Quinze n. 26 — Palacete Mappin" foi parcialmente preservado, permitindo a identificação da localização histórica da instituição. O documento analisado oferece insights valiosos sobre as transformações sociais tecnológicas do Brasil no início do século XX. O conceito de datilografia como "profissão de futuro" reflete as mudanças nas práticas de trabalho administrativo e a crescente importância da documentação escrita na sociedade moderna (Chartier, 1994). A localização da escola na "Rua Quinze n. 46 — Palacete Mappin" no Rio de Janeiro, então capital federal, evidencia a concentração de inovações educacionais e tecnológicas nos grandes centros urbanos brasileiros.

O ano de 1918 marca um período vital na história da tecnologia de escritório. A máquina de escrever, inventada décadas antes, estava finalmente se estabelecendo como ferramenta essencial para o trabalho administrativo e jornalístico (Adler, 1973). A promoção da datilografia como "profissão de futuro" revelou-se profética, considerando que esta habilidade permaneceu fundamental até o advento dos computadores pessoais na década de 1980. A menção à taquigrafia junto à datilografia indica a complementaridade dessas técnicas na época. A taquigrafia, sistema de escrita abreviada para registro rápido da fala, era especialmente valorizada em contextos jurídicos, jornalísticos e secretariais (Pitman, 1837). A combinação dessas habilidades representava uma formação profissional completa para o trabalho de escritório moderno.

CONCLUSÕES

Esta pesquisa demonstrou que as técnicas modernas de OCR, especificamente a implementação do Tesseract através de Python, constituem ferramentas eficazes para a digitalização e preservação de documentos históricos. A análise do anúncio da Escola Remington de 1918 revelou tanto as potencialidades quanto os desafios inerentes a esse

tipo de aplicação tecnológica. Os resultados obtidos, com precisão de aproximadamente 78%, indicam que, embora existam limitações significativas, é possível extrair informações históricas valiosas de documentos centenários. Apesar dos erros de segmentação e reconhecimento de caracteres, elementos essenciais como a identificação da instituição ("ESCOLA REMINGTON"), o conceito central ("profissão de futuro"), as disciplinas oferecidas (datilografia e taquigrafia) e a localização (Rua Quinze, Palacete Mappin) foram preservados, permitindo a compreensão do conteúdo histórico fundamental.

A contribuição principal deste trabalho reside na demonstração prática de que tecnologias acessíveis e relativamente simples podem ser aplicadas com sucesso na preservação digital do patrimônio documental. Isso tem implicações importantes para a democratização do acesso ao conhecimento histórico e para a sustentabilidade de projetos de digitalização em instituições de diferentes portes. Em última análise, este estudo reafirma o papel fundamental da tecnologia como aliada na preservação e disseminação do patrimônio cultural, contribuindo para que documentos históricos permaneçam acessíveis às gerações futuras em formato digital pesquisável e preservado.

REFERÊNCIAS

ADLER, M. H. The writing machine: a history of the typewriter. London: George Allen and Unwin, 1973.

ANTONACOPOULOS, A.; BRIDSON, D.; PAPADOPOULOS, C.; PLETSCHACHER, S. "A Realistic Dataset for Performance Evaluation of Document Layout Analysis," 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009, pp. 296-300, doi: 10.1109/ICDAR.2009.271.

BAGNO, M. Nada na língua é por acaso: por uma pedagogia da variação linguística. São Paulo: Parábola Editorial, 2007.

CHARTIER, R. The order of books: readers, authors, and libraries in Europe between the fourteenth and eighteenth centuries. Stanford: Stanford University Press, 1994.

CHERIET, M.; KHARMA, N.; LIU, C. L.; SUEN, C. Y. Character recognition systems: a guide for students and practitioners. Hoboken: John Wiley & Sons, 2007.

CLARK, A. PIL: Python Imaging Library handbook. 2009. Disponível em: https://pillow.readthedocs.io/. Acesso em: 15 jan. 2025.

CLAUSNER, C.; PLETSCHACHER, S.; ANTONACOPOULOS, "Scenario Driven In-depth Performance Evaluation of Document Layout Analysis Methods," 2011 International Conference on Document Analysis and Recognition, Beijing, China, 2011, pp. 1404-1408, doi: 10.1109/ICDAR.2011.282.

CONWAY, P. Preservation in the digital world. Council on Library and Information Resources, 1996.

EISENSTEIN, E. L. The printing revolution in early modern Europe. 2nd ed. Cambridge: Cambridge University Press, 2005.

GITELMAN, L. Scripts, grooves, and writing machines: representing technology in the Edison era. Stanford: Stanford University Press, 1999.

HOLLEY, R. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, v. 15, n. 3/4, 2009.

MORI, S.; SUEN, C. Y.; YAMAMOTO, "Historical review of OCR research and development," in Proceedings of the IEEE, vol. 80, no. 7, pp. 1029-1058, July 1992, doi: 10.1109/5.156468.

NAGY, G. Twenty years of document image analysis in PAMI. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 22, n. 1, p. 38-62, 2000.

PITMAN, I. Stenographic sound-hand. London: Samuel Bagster and Sons, 1837.

PLETSCHACHER, S.; ANTONACOPOULOS, A. The PAGE (Page Analysis and Ground-truth Elements) format framework. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, 20., 2010, Istanbul. Proceedings... IEEE, 2010. p. 257-260.

SMITH, R. An overview of the Tesseract OCR engine. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 9., 2007, Curitiba. Proceedings... IEEE, 2007. p. 629-633.

TERRAS, M. Image to interpretation: an intelligent system to aid historians in reading the Vindolanda texts. Oxford: Oxford University Press, 2006.

VAN ROSSUM, G.; DRAKE, F. L. Python 3 reference manual. Scotts Valley: CreateSpace, 2009.

REIS JR., Dalmir. Datilografia: profissão do futuro – 1918. Propagandas Históricas, 2013. Disponível em: https://www.propagandashistoricas.com.br/2013/04/datilografia-profissao-do-futuro-1918.html. Acesso em: 10 ago. 2025.