



Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp),
Instituto de Geociências e Ciências Exatas, Programa de Pós-
Graduação em Geografia UNESP - Campus de Rio Claro

doi.org/10.5016/geografia.v51i1.19576

ISSN 1983-8700 | E-19576

Computadores fazem secas, pesquisadores fazem teses: notas metodológicas sobre IA e Geografia Histórica

Igor Carlos Feitosa Alencar¹

¹ Universidade de São Paulo, Faculdade de Filosofia, Letras e Ciências Humanas, São Paulo, São Paulo, Brasil.

ORCID: <https://orcid.org/0000-0001-9707-3393>

E-mail correspondente: igor.carlos@usp.br

Destaques:

- Proposta metodológica para uso de IA em pesquisas de Geografia Histórica.
- Detalha o uso de Python e IA para processar fontes documentais do século XIX.
- Supera desafios de OCR e linguagem arcaica com treinamento de IA.
- Reflete sobre a contradição e os desafios éticos da IA no ambiente acadêmico.
- Demonstra como a IA pode potencializar a análise de fontes com supervisão humana.

Resumo: Este texto é uma nota metodológica sobre a utilização de Inteligência Artificial (IA) em pesquisas de Geografia Histórica, partindo da contradição entre o consumo hídrico das Big Techs por detrás dessas ferramentas e o seu uso no processamento de fontes do século XIX sobre as secas no Ceará. Relatamos os primeiros passos e desafios na elaboração de códigos Python com ChatGPT e Gemini para a transcrição e resumo do jornal “O Cearense” (1846-1891). Discutimos a evolução metodológica da proposta, desde as dificuldades iniciais até a robustez da versão mais recente, que utiliza IA's generativas treinadas com documentos e glossários personalizados para transcrições, resumos e relatórios automáticos. Demonstramos, como resultado, a necessidade de apropriação dessas ferramentas para potencializar a análise das fontes, ressaltando a imprescindível supervisão humana para garantir a qualidade dos resultados e os aspectos éticos do uso de IA no ambiente acadêmico.

Palavras-chave: Inteligência Artificial; Geografia Histórica; Ceará; Fontes Documentais; Metodologia.

Como citar este artigo:

ALENCAR, Igor Carlos Feitosa. Computadores fazem secas, pesquisadores fazem teses: notas metodológicas sobre IA e Geografia Histórica. **Geografia**, Rio Claro-SP, v. 51, e-19576, 2026. Disponível em: <https://www.periodicos.rc.biblioteca.unesp.br/index.php/ageteo/article/view/19576>. DOI: doi.org/10.5016/geografia.v51i1.19576.



Este é um artigo publicado em acesso aberto (*Open Access*) sob a licença CC - *Creative Commons Attribution* 4.0, que permite o uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

AI, droughts, and dissertations: methodological notes on artificial intelligence in historical geography

Abstract: This text is a methodological note on the use of Artificial Intelligence (AI) in Historical Geography research, beginning with the contradiction between the water consumption of the Big Techs behind these tools and their application in processing 19th-century sources on droughts in Ceará. We report the initial steps and challenges in developing Python code with ChatGPT and Gemini for the transcription and summarization of the historical newspaper “O Cearense” (1846–1891). We discuss the methodological evolution of the proposal, from the initial difficulties to the robustness of the most recent version, which employs generative AIs trained with customized documents and glossaries for automatic transcriptions, summaries, and reports. As a result, we demonstrate the necessity of adopting these tools to potentialize the analysis of sources, emphasizing the essential role of human supervision to ensure the quality of the results and the ethical aspects of using AI in the academic environment.

Keywords: Artificial Intelligence; Historical Geography; Ceará; Documentary Sources; Methodology.

IA, sequías y tesis: apuntes metodológicos sobre inteligencia artificial en geografía histórica

Resumen: Este texto es una nota metodológica sobre la utilización de la Inteligencia Artificial (IA) en investigaciones de Geografía Histórica, partiendo de la contradicción entre el consumo hídrico de las Big Techs detrás de estas herramientas y su uso en el procesamiento de fuentes del siglo XIX sobre las sequías en Ceará. Relatamos los primeros pasos y desafíos en la elaboración de códigos Python con ChatGPT y Gemini para la transcripción y el resumen del periódico histórico “O Cearense” (1846–1891). Discutimos la evolución metodológica de la propuesta, desde las dificultades iniciales hasta la solidez de la versión más reciente, que utiliza IA generativas entrenadas con documentos y glosarios personalizados para transcripciones, resúmenes e informes automáticos. Como resultado, demostramos la necesidad de apropiarse de estas herramientas para potenciar el análisis de las fuentes, resaltando la imprescindible supervisión humana para garantizar la calidad de los resultados y los aspectos éticos del uso de la IA en el ámbito académico.

Palabras clave: Inteligencia Artificial; Geografía Histórica; Ceará; Fuentes Documentales; Metodología.

INTRODUÇÃO

Se faz desnecessário reservar extensas linhas para reafirmar o quanto em várias esferas as ferramentas de inteligência artificial estão cada vez mais presentes em atividades cotidianas, não sendo diferente no meio acadêmico. Em abril de 2025, somente o *ChatGPT* registrava, segundo o *CEO* da *OpenAI's*, Sam Altman, cerca de 800 milhões de usuários, dos quais 15,5 milhões dispunham de

registro pago (Paris, 2025). Ainda segundo o *SimilarWeb*, em junho de 2025, 16,38% dos acessos ao *ChatGPT* eram nos EUA, seguido pela Índia (7,77%), Brasil (5,86%), Japão (3,52%) e Alemanha (3,45%). Os demais países representavam cerca de 63% (*SimilarWeb*, 2025).

A disponibilidade crescente e cada vez mais integrada dessas ferramentas é um fator que influencia o aumento de usuários. Basta apontar que, ao abrir o *Google Docs* para escrever esse texto, antes de grafar uma única palavra, a mensagem “Gerar documento” aparece na tela; ou, ainda, quando se pausa a escrita por alguns segundos, surge: “Quero ajuda para escrever Alt + W”.

Resistida as tentações de recorrer a essas ferramentas no *Google Docs* e “apenas” ditar quais pretensões temos com esse texto, cabe inferir que se trata de um esforço de reflexão sobre o uso de ferramentas de inteligência artificial, especialmente *ChatGPT* e *Gemini*, na pesquisa acadêmica a partir do desenvolvimento de códigos *Python* executados no *Google Colab*.

Essa temática, urgente e necessária, aparece na bibliografia recente em análises pontuais, muito em função da rapidez no desenvolvimento de novas ferramentas e funções, além das múltiplas possibilidades de aplicação. Sobreira (2025) realiza concomitantemente um resgate histórico do desenvolvimento das IA's, aponta possíveis agendas de pesquisa e critica os historiadores por ainda negligenciarem (por desconhecimento e receio) a utilização dessas ferramentas em suas pesquisas. Já Freitas (2025) analisa a produção historiográfica sobre a temática nos periódicos nos últimos cinco anos, destacando a escassez de trabalhos empíricos e as lacunas conceituais sobre o tema.

Por outro lado, se há uma ausência de trabalhos empíricos utilizando IA, especialmente nas humanidades, as preocupações com o uso dessas ferramentas na produção científica de modo geral aparece com mais frequência. Uma reportagem publicada na Revista Pesquisa FAPESP relata o esforço das universidades brasileiras em formular diretrizes sobre o uso ético da IA, destacando os riscos de erros, desinformação e vieses discriminatórios em pesquisas (Schmidt, 2024). As revistas científicas também ensaiam um movimento no mesmo sentido, seja formulando políticas para o uso de IA, como a Revista Caminhos de Geografia, ou ainda incorporando na política editorial esse

tema, exigindo que autores(as) informem quando e como usaram IA nos manuscritos - é o caso do Boletim Paulista de Geografia e da Revista Nera.

Por essas razões optou-se pela escrita desse texto combinando uma análise empírica exposta na proposta metodológica associada à crítica da atuação e condições de funcionamento das *Big Techs*. Pois, se o uso de IA atualmente na temática das secas objetiva realizar previsões e modelos com vistas a atenuar os efeitos dos eventos climáticos extremos, nós voltamos esse olhar para o passado, a partir do levantamento, sistematização e organização das fontes documentais da nossa investigação de doutorado em curso, acerca da transformação das secas em oportunidades de avanço do capital e superexploração do trabalho ao longo do século XIX no território cearense¹.

Assim, no que se refere a relação entre secas e avanço do capital, não podemos deixar de assinalar o papel das *Big Techs* nesse cenário. Visto que por detrás do funcionamento das ferramentas de inteligência artificial, há um elevado e crescente consumo hídrico para o resfriamento de *data centers* (supercomputadores). Estudos estimam que a cada *prompt* requisitado há um consumo de cerca de 16 onças d'água (O'brien; Fingerhut, 2023).

Nesse cenário de agravamento da crise climática e de disputas por reservas hídricas, a oferta de água condiciona o crescimento do potencial das inteligências artificiais. Incidindo numa movimentação das *Big Techs* em busca de territórios para exploração desses recursos, como a África Sub-saariana, além de provocar novos conflitos como o relatado por Gupta, Bosch e Vliet: “Google’s planned data centre in Uruguay, which recently suffered its worst drought in 74 years, would require 7.6 million litres per day, sparking widespread protest” (2024, p. 2)².

As contradições e ironias desse cenário são notáveis: a mesma tecnologia, que hoje consome recursos em escala exponencial, nos ajuda a compreender o impacto das secas no passado. Cientes e atentos a essas questões, apresentamos a seguir um esforço de construção de uma proposta metodológica que envolve o

¹ Pesquisa financiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP. Processo: 23/03972-5.

² “O data center planejado pelo Google no Uruguai, que recentemente sofreu sua pior seca em 74 anos, exigiria 7,6 milhões de litros diariamente, provocando protestos generalizados”. Tradução nossa.

processamento de fontes documentais do século XIX através de ferramentas de inteligência artificial.

Para tal, dividimos o texto em três partes. Na primeira explicamos os percalços nas elaborações iniciais do código principal; na segunda expomos como se deu o processo de treinamento do código para a linguagem e documentos dos Oitocentos; e na terceira apresentamos e analisamos amostragens dos resultados obtidos. Ademais, nas considerações finais discutimos os limites e desafios no uso dessas ferramentas de IA para as pesquisas.

FONTES DOCUMENTAIS, INTELIGÊNCIAS ARTIFICIAIS E PERCALÇOS HUMANO

Se por um lado são extensos os debates sobre as definições da própria área, bem como sobre o que e quando se investiga em Geografia Histórica (Morris, 1987; Harvey 1990), a ancoragem dessa na utilização das fontes documentais é, por outro lado, um consenso em meio as distintas e diversas abordagens teórico-metodológicas.

Um dos desafios posto nas investigações em Geografia Histórica está no manejo das fontes documentais (uma herança das próprias pesquisas historiográficas onde, diferente dos nossos colegas historiadores, sofremos de uma defasagem em nossa formação). Assim, pensamos que o primeiro passo para encarar tal desafio está na definição do recorte temporal - ação arbitrária e íntima ao objeto (Rodrigues, 2019) - pois esse, por consequência, aponta as possibilidades de levantamento das fontes.

Cabe ressaltar que, nesse caso, o recorte temporal prescinde o espacial, pois, a exemplo, por mais que o nosso recorte seja o território cearense, as fontes utilizadas não se restringem ao mesmo. Afinal, como se poderia pesquisar o século XIX, especialmente o Segundo Reinado, sem nos valer dos acervos e arquivos da antiga capital do império?

Dito isso, delimitado o recorte temporal e os possíveis locais de levantamento das fontes, bem como da tipologia dessas (manuscritos, relatórios, jornais, opúsculos, manuais, diários, etc.), o desafio seguinte está no registro e, o mais crucial, na capacidade de análise frente a volumosa disponibilidade de

algumas, como são os jornais do século XIX. Nesses casos, cabe ao pesquisador(a) escolher: realizar uma rigorosa pré-análise que defina os períodos ou edições fundamentais para um aprofundamento nessas fontes - incorrendo no risco de as limitar demasiadamente -; ou se debruçar na máxima quantidade possível de material, flertando com a superficialidade?

É no enfrentamento a esse dilema no qual encontramos nas ferramentas de inteligência artificial a possibilidade de um auxílio complexo e eficaz. Pois, apesar da nossa utilização recorrente do *ChatGPT* para atividades como revisão ortográfica e auxílio na sistematização de algumas informações, nosso *insight* para um uso mais efetivo e sofisticado dessas ferramentas na pesquisa documental se deu a partir do curso de extensão “AI Tá On: Inteligência Artificial para Humanidades e Educação”, promovida pelo Grupo Experimental em Semiótica Computacional (SemioCom), uma iniciativa interdisciplinar de diversos pesquisadores/as da Universidade de São Paulo.

Nesse curso, um dos módulos explorava a modelagem de tópicos - uma forma de organização automática de textos em temas ou tópicos, essencial para analisar grandes volumes de dados, “*a subfield of natural language processing (NLP), uses statistical algorithms to extract latent topics or themes from a collection of documents, making it easier to analyze and interpret complex data*” (Taghandiki; Mohammadi, 2023, p. 1). Porquanto, se a modelagem de tópicos é utilizada “*to identify the most common patterns or themes in unstructured text data, which can be used to extract relevant information*” (Taghandiki; Mohammadi, 2023, p. 1), pensamos em realizar o sentido inverso: apontar preliminarmente os termos (*key-words*) que deveriam ser identificados no nosso material.

Isso foi possível a partir do exemplo prático do curso, o qual se deu com a apresentação de um exercício com a elaboração de um código Python - uma linguagem de programação para análise de dados - aplicado para analisar comentários misóginos no *YouTube* durante as transmissões da Copa do Mundo Feminina. Esse código foi executado na plataforma *Google Colab*, um serviço de nuvem ideal para o ensino e a pesquisa em aprendizado de máquina, por oferecer

funcionalidades essenciais, como uma robusta GPU (*Graphics Processing Unit*) integrada às bibliotecas de IA previamente configuradas (Carneiro *et al.*, 2018).

Partindo dessa estrutura de análise dos comentários no *YouTube* (que ultrapassavam dez mil), voltamos nosso olhar para os jornais que vínhamos analisando a partir de palavras-chave na plataforma da Hemeroteca Digital da Fundação Biblioteca Nacional (HDBN). Em resumo, a ideia seria: como automatizar essa busca de combinação de palavras-chave nos jornais do século XIX disponíveis na HDBN para identificar notícias que nos interessam, transcrever, registrar informações como a edição, o ano, a página e, especialmente, elaborar um resumo fidedigno e coeso?

Para tal, e enquanto leigos no que se refere à linguagem de programação, partimos dos códigos *Python* disponibilizados no supracitado curso para elaborar nossos próprios *prompts* no *ChatGPT* que atendessem as nossas expectativas. Assim, escolhemos o principal periódico que trabalhamos na pesquisa, o “Jornal O Cearense”, para realizar um projeto-piloto de levantamento de todas as notícias com menções às secas, abarcando a íntegra das publicações do jornal (de 1846 a 1891).

Daí decorreram equívocos e limitações que gradualmente foram corrigidos. O primeiro equívoco ocorreu no sentido de tentar humanizar ações de natureza automatizadas. Em outros termos, se imaginou ser possível transferir a sequência de ações que realizadas manualmente para o código, ou seja, que ele acessasse o portal da HDBN, pesquisasse os termos “secca” e “seccas” (respeitando a grafia da época) e pelo uso do OCR (*Optical Character Recognition*) disponível nos documentos hospedados na própria plataforma, gerasse uma tabela com a transcrição da notícia e as demais informações.

O código, ao tentar acessar o site de forma recorrente e automatizada, levou a plataforma da HDBN a interpretar essa atividade como um comportamento suspeito ou de ataque, bloqueando o acesso. Isso impediu qualquer coleta de informação diretamente do portal, mesmo com as sucessivas tentativas de aumento no intervalo entre os acessos.

Dessa forma, tomou-se a primeira lição: era necessário que os arquivos estivessem hospedados em pastas no *Google Drive*, afinal, a plataforma do

Google Colab, onde se executam os códigos, tem vinculação direta com o serviço de nuvem. Para tal, foi realizado o *download* das edições do ano de 1846 do jornal O Cearense, depois o *upload* desses arquivos para a nuvem e a elaboração de um novo prompt para o *ChatGPT*: “um *script* em código *Python* para o *Google Colab* que acesse PDFs de jornais do século XIX no *Google Drive*, identifique ocorrências específicas dos termos ‘secca’, ‘estiagem’ e ‘seccas’ (excluindo variações de falso-positivos, como ‘Fonseca’), e compile esses dados (ano, edição, data, página, resumo) em uma tabela em formato CSV., salvando-a diretamente na pasta de origem no *Drive*”.

Se por um lado obteve-se sucesso quando finalmente o código conseguiu acessar os documentos, por outro, duas limitações surgiram: a dificuldade de captar as informações como ano, edição, página e data (o que estava relacionado a própria estrutura de formatação e da qualidade das páginas escaneadas, como rasuras e rasgos) e a segunda, e mais grave, na qualidade da transcrição que inviabilizava a geração de um resumo.

A primeira limitação técnica, relacionada à qualidade do OCR, foi superada com uma abordagem de pré-processamento de imagem mais robusta. Em vez de depender do OCR nativo dos documentos, o *ChatGPT* sugeriu - após várias execuções do código e o envio dos resultados insatisfatórios - que o *script* realizasse uma rasterização de cada página PDF em PNG, aplicando parâmetros de alta resolução e otimização da imagem (como binarização e correção de inclinação). Posteriormente, essa imagem aprimorada seria re-OCRizada por um motor mais potente e convertida novamente em PDF com uma camada de texto pesquisável. Esse processo garantiu que o motor de OCR recebesse imagens de texto de qualidade superior, melhorando significativamente a acurácia do reconhecimento de caracteres e a capacidade de captura das informações. Como é possível observar na Figura 1:

Figura 1 - *Printscreen* de uma das primeiras transcrições com pré-processamento de imagem

jornal	pagina	texto	palavras_chave_encontradas
1846_00001.pdf	1		Nenhuma
1846_00001.pdf	2		Nenhuma
1846_00001.pdf	3	★ PARÁGRAFO RELEVANTE ★ (Palavras-chave: secas) l f KSV x>f' i i - " ' : ' : J>'~: ', ' i ' , : "-'- ':-"" >.' '. O CEARENSE m "":jil.rã Vhi -. M B do paiz. Nem se supponhada haver de real na proposição, ue acabamos de emitir: propalado por todas as classes o amor do trabalho; infiltrado elle no coração de todos; todos conhecendo que só desta guisa é q se pôde, adonando-se-lhe a economia, fazer huma fortuna, e até enriquece; sendo claro que este estado traz necessariamente a força com sig; é evidente qo, quando o povo conhecer os seus verdadeiros interesses; quando assim pensar e obrar, o Ceará, o-misero Ceará entregue á tantos obstaculos, que entopcem o eu desenvolvimento; Victima dos rigores de repetidas secas, mas filhas do nosso desleixo e incuria do que de nossa posição topographica, não precisará menigar o paô, como ora o tem feito; não precisará estar dependente de algumas de suas irmaas, como até hoje tem estado: elle se collocarána verdadeira posição nee que a muito devia estar collocado, e será respeitado como deve ser. .E não valle apenas que para este fim, mesmo remoto, como parece ser e na verdade o e, façamos hum sacrificio commum, e procuremos de nossa parte convergir quanto s' possa para conseguimol-o? Cremos qo sim; ésta nossa crença nos leva a convicção: alma de que teremos franco e leal apoio f de, nossos concidadãos na espinhosa em presa-de que estamos^{AA}sta. Tal é o nosso fim l dando á luz o. Ces^{rumov} revestido de tão Caro nome procuraremos nunca desmentir- p T, nunca manchar-o; e se não conseguirmos^r l i o tructo que desejamos, nos restará á con- solação de que fiseremos o quanto podemos; 11 o quanto estava em nossas tracas forças, a l A imprensa esltre nós, mos contra a nossa imprensa; uma injuria que lhe irrogamos? Antes o fora! nós nos correríamos de vergonha por assim queremos desacreditá-la, em vez de nos cobrimos della, vendo-a neste triste estado. In da mais; quaes são os bens, que em nossa provincia, tem produzido a imprensa? Não queremos descer aos factos; não queremos particularisal-os: se o quisessemos, cremos que mostraríamos, que em lugar de bens ella nos tem causado males incalculaveis, e que nol-os continuará a causar se a não fiseremos tomar outro norte. Se por meio da imprensa podemos derramar por todas as classes úteis conhecimentos, de que venha depois gosar a sociedade: se por meio della podemos adotar os costumes do povo: mostrar-lhes seus verda- deiros interesses; guial-o para o bem; incital-o ao trabalho e a economia, sem o q jamais elle será feliz: por que o não fazemos? Por que, pelo contrario damos-lhe uma idea tão desfavorável desta grande invenção: o obrigamos a fazer de nós que scrivermos para elle hum juizo tão triste, tão mesquinho? Por que, sobre tudo, nos apomos fora de nossa provincia á irrisão le nossos irmãos, ao desprezo de não sermos lidos" pois só escrevemos decomposturas e arrieiradas; ao descrédito de nossas capacidades, de nossas intelligências e illustração? Ao menos não valle apenas seguirmos outro rumo, por amor de nós mesmos? Cremos que sim; e esperamos dos nossos collegas no jornalismo uma completa reforma na marcha, que tem segrido até agora. Uma vez que os Ceaxenses não se podem ligar em um só nensamento politico; uma vez que isto é absolutamente impossivel pela natureza do bello systema, que felizmente nos rege: procuremos unir-nos naquillo em queopolemos, inda que façamos um pequeno sacrificio, se é que nisto o fazemos: liguemos-nos ao menos naquillo que, nada tendo de commum com o espirito de partido, tem tudo com o credito de nós todos em geral, com a felicidade presente e futura de nossa provincia. Que se dirá de nós ao lere-m-se algumas de nossas folhas? duções, esperamos não nos desacreditarmos pela nossa linguagem. Eis o que desejávamos fisessem todos os que são levados a escrever para o publico; eis o que lhes pedimos com todas as veras por amor do torraô caro em que nascemos: quasi que até agora só se ha escrito para o nosso descrédito; não será possível rehabilitar-mos-nos? Escrever em uma linguagem mais pura e castigada ? Escrever cousas que interessem á nossa infeliz provincia, victima das secas pela nossa incuria e desleixo; victima do frenezil do partidos, por não conhecermos bem os nossos verdadeiros interesses? A imprensa, hum dos inventos mais sufolimeV do homem: aquelle por meio do <lúaj desaparece a desigualdade de conheimentos, tão contraria aos interesses da sociedade: que multiplicando com uma facilidade infinita o que de bom e útil se tñhi escrito sobre todas as artes, e sciências" ai derrama no meio do povo eoins- "trifle de seus direitos e obrigações; em nossa provincia, nem sempre tem guardado 'm HraUes da decência, e de uma bem Que idea tristissima não se fará de nós, <ntiendida liber lade. Multas vezes, ella se tem transviado da l: n H verdadeira senda por onde deviria camii thai e usando de humo linguagem as- querW e immitinda, tem lançado de si ver- ^ <ldadeiros pasquins indignos de quem os esM crevo, o daquelle" para qn'm são	secas

Fonte: O autor (2025).

No *printscreen* acima é possível visualizar quatro colunas: a primeira intitulada “jornal” identifica o arquivo analisado, no caso, a edição 001 de 1846; na sequência o número da página; a terceira trata-se da transcrição do trecho referente a palavra-chave encontrada, apresentada na quarta coluna, o termo “secas”. Nesse caso, foram suprimidas as colunas de data e resumo, pois o essencial seria assegurar primeiro uma boa qualidade na transcrição, para então possibilitar a geração de resumos com o mesmo padrão.

No entanto, é evidente na Figura 1 que a transcrição, apesar legível em sua maior parte, era demasiada extensa e apresentava uma série de palavras transcritas incorretamente, além de uma excessiva poluição de elementos aleatórios, como a sequência da primeira linha “i|f|KSV x>f' i i | | - " ' | | : ' | : J>'~: ', ' | i ' , : "-'-|':-""| |>.'|”. Esses elementos representavam justamente rasuras, manchas e quaisquer outras marcações que constavam nos documentos físicos quando esses foram escaneados. Assim, era fundamental então que o código realizasse uma espécie de limpeza antes da transcrição, durante a fase de pré-processamento.

Tentou-se realizar esses ajustes no *ChatGPT* sem sucesso, especialmente porque a geração de códigos *Python* requer grande capacidade de processamento, que em uma conta gratuita rapidamente se esgotava o limite de

aproximadamente dez mensagens a cada três horas disponíveis para o modelo *GPT-4o*. Assim, havia prejuízo tanto ao ter que aguardar as três horas para continuar aquela conversa ou mesmo iniciar uma nova pelo modelo *GPT-3.5*, que, além de mais lento, é limitado para atividades mais complexas como as em questão.

A alternativa nesse ponto foi optar pela utilização da plataforma chinesa *DeepSeek*. Repetindo a metodologia de replicar os códigos gerados pelo *ChatGPT*, bem como os resultados insatisfatórios para construir alternativas. Mas houve poucos avanços. Pois, mesmo que totalmente gratuito com suporte a contextos extensos de até 128K *tokens*, o *DeepSeek* sofre com limitações de infraestrutura - seus servidores apresentam lentidão ou sobrecarga em processamentos complexos, não raro gerando a mensagem “*The server is busy. Please try again later*”. De modo que se voltou ao mesmo ponto: aguardar horas para prosseguir com a conversa ou iniciar uma nova, comprometendo a linha de raciocínio até então desenvolvida com a IA.

Também tentou-se realizar os ajustes e melhorias no código pelo próprio *Google Colab*. Isso porque a plataforma dispõe de um chat integrado à IA, acessível tanto proativamente, quanto em resposta a erros, exibindo a opção “Explique meu erro” na caixa de saída das células. Contudo, esse apresentava limitações significativas, especialmente na construção de requisições mais complexas ou ainda na manutenção de um histórico duradouro das interações e soluções para consultas futuras, o que dificultava o processo iterativo de aprimoramento do código.

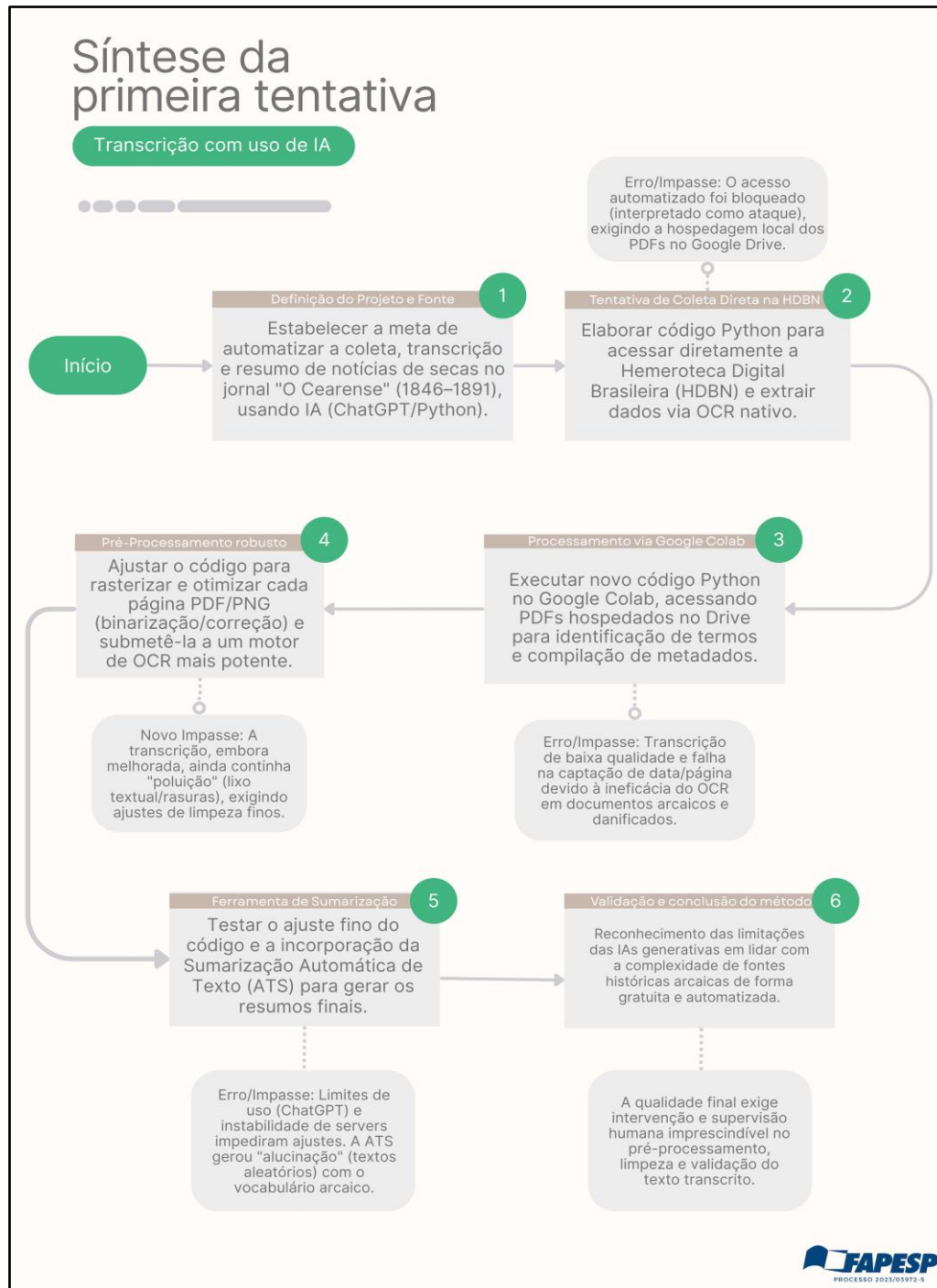
Outra opção testada foi junto a plataforma *Google AI Studio*, um ambiente gratuito de testes para os modelos de IA. Ao tentar que o código ao invés de buscar melhorar as transcrições, focasse em algum modelo de inteligência artificial capaz de “traduzir” aquelas transcrições e elaborar os resumos. A sugestão da IA foi a incorporação ao código da chamada “sumarização automática”, ou “*Automatic Text Summarization*” (ATS):

ATS systems are designed by applying one of the text summarization approaches: extractive, abstractive, or hybrid. The extractive approach selects the most important sentences

from the input text and uses them to generate the summary. The abstractive approach represents the input text in an intermediate form then generates the summary with words and sentences that differ from the original text sentences. (El-Kassas *et al.*, 2021, p. 1-2)³.

Em tese, o uso da ATS atingiria o objetivo mais pretensioso, a geração dos resumos a partir das frases mais importantes do texto. Todavia, incorreu-se em graves erros: os resumos gerados online mostravam textos aleatórios e desconexos do contexto das notícias, caracterizando “alucinação”. Já os resumos gerados *offline* eram cópias literais das transcrições. De forma que a própria plataforma, em uma mensagem, reconhecia que tal cenário “ilustra perfeitamente os desafios e as limitações atuais da sumarização automática, especialmente com textos históricos, arcaicos ou com estruturas frasais complexas”. Na Figura 2, esquematizamos o percurso até então percorrido:

³ “Sistemas de sumarização automática de texto (ATS) são elaborados a partir de três abordagens: a extrativa, a abstrativa ou a híbrida. A abordagem extrativa seleciona as frases mais importantes do texto original para criar o resumo. Já a abordagem abstrativa representa o texto de entrada de forma intermediária e, então, gera um resumo com palavras e frases diferentes do texto original”. Tradução nossa.

Figura 2 - Síntese da primeira tentativa

Fonte: O autor (2025).

Como observado, apesar de todo esse percurso demonstrado na Figura 2, naquele momento, diante do impasse que comprometia a qualidade na captação,

registro e geração de informações das fontes documentais, cogitou-se não ser possível realizar tal tarefa, especialmente em função da qualidade do material e da linguagem do século XIX. Restringindo somente para materiais mais recentes, como o exemplificado no curso, a possibilidade de uso de IA para esses fins. Pelo menos não gratuitamente, como almejado.

DA INTELIGÊNCIA ARTIFICIAL À INTELIGÊNCIA DOCUMENTAL

Em maio de 2025 o *Google* ofereceu gratuitamente para estudantes, por um período de quinze meses, o acesso ao Gemini Advanced (integrado ao seu modelo mais robusto de IA, o *Gemini 2.5 Pro*). A partir de então retomaram-se as tentativas de aprimoramento do código.

A nova estratégia se deu numa dupla abordagem, primeiro continuar compartilhando os resultados anteriormente obtidos e descrevendo os principais elementos que poluíam e prejudicavam a compreensão das transcrições. Além de, pela capacidade do *Gemini Pro*, anexar exemplos dos jornais que seriam analisados, para a própria IA compreender especificidades e propor soluções para o *script*. De modo que a primeira modificação realizada resolveu consideravelmente alguns problemas das transcrições, dado a partir da inserção da função intitulada “*preprocess_text_lightly*”, uma espécie de faxineiro digital que arruma espaços e linhas e remove caracteres estranhos ao texto.

Mas, mesmo que os textos agora eram transcritos de forma mais limpa, muitos termos apareciam incompletos, não identificados ou com graves erros ortográficos. E a razão estava no vocabulário da época, do qual o código, elaborado no e para os parâmetros contemporâneos, encontrava sérias dificuldades de compreensão. Era preciso então levar a IA ao século XIX.

Tornou-se impreterível o treinamento do código na linguagem e no contexto do século XIX. Para tal, outro aspecto fundamental para as transcrições atingirem níveis suficientes para um resumo ser gerado, foi a incorporação de glossários personalizados que seriam consultados pelo código todas às vezes que ele realizasse o OCR e a subsequente transcrição ao identificar uma das palavras-chave. A construção desse glossário foi fundamental para 1) normalização de

grafias antigas e variações ortográficas; 2) aprimoramento da Extração de Entidades Nomeadas (NER)⁴ e 3) melhora da relevância e contexto para a IA.

Destarte, para a construção do primeiro glossário utilizamos a obra “Diccionario de vocabulos brasileiros” (Beaurepaire-Rohan, 1889). Para aprimorar a transcrição interessava sobretudo a grafia das palavras e não a sua definição, por isso, o *Gemini* elaborou um código auxiliar que, acessando na nuvem o arquivo do dicionário em PDF., realizou uma limpeza, removeu todas as definições e elaborou uma lista de 1580 termos. Essa lista foi colada no próprio *Gemini* e solicitado que inserisse na frente de cada palavra do século XIX a sua respectiva grafia contemporânea. Com essa lista pronta criou-se um novo arquivo no formato txt. que foi inserido na nuvem para ser consultado pelo código principal.

Também julgou-se necessário um glossário com nomes de lugares e pessoas para que o código o consultasse. No que se refere aos lugares, para a construção do glossário solicitamos ao próprio *Gemini* que fornecesse uma lista com o nome das províncias, capitais e principais cidades do Brasil no século XIX, bem como do maior número possível de países, capitais, lugares e regiões do mundo no mesmo período.

Quanto ao levantamento do nome de pessoas, em consonância com o nosso objeto, onde interessava sobretudo os agentes que compunham a classe senhorial cearense (Mattos, 2017), utilizamos como fonte os três volumes do “Diccionario bio-bibliographico cearense” (Studart, 1915). Para facilitar a extração dos nomes, o código utilizou como fonte os índices do dicionário, publicados na Revista do Instituto do Ceará (Nery, 1971; 1972). De modo que, com a lista gerada pela IA e a outra extraída dos três volumes do dicionário, foram registrados 1504 termos em um novo documento txt.

Esses dois glossários cumprem a função de auxiliar o código principal tanto para realizar a transcrição das notícias e, quando possível, identificar em uma coluna os nomes de pessoas e, em outra, os nomes dos lugares. Todavia, o

⁴ Segundo Gonçalves e Dornelles, “O Reconhecimento de Entidades Nomeadas (NER) é uma sub-tarefa de extração de informação que busca localizar e classificar entidades nomeadas em textos em categorias pré-definidas, como nomes de pessoas, organizações, localizações, tempo, data e dinheiro” (2022, p. 427). Tradução nossa.

fundamental desse material está em fornecer um substrato de palavras, pessoas e lugares que auxiliem na análise das notícias e consequentemente na geração dos resumos - o maior foco.

No que se refere aos resumos, esses só atingiram a qualidade almejada desde o início quando o *Gemini* sugeriu a utilização de uma API *Generative*. Trocando em miúdos, primeiro, basta compreender que “as interfaces de programação de aplicativos (APIs) são conjuntos de ferramentas, definições e protocolos para a criação de aplicações de software” (Red Hat, 2023, n.p). Dito de outra maneira, essa ferramenta digital recebe um pedido do *script Python*, leva esse pedido para o serviço do *Google Gemini*, espera a resposta e a traz de volta.

Para autorizar o código a utilizar essa API *Generative*, é necessário acessar a plataforma Google Cloud Platform (GCP), criar um projeto, ativar a API *Generative Language*, gerar uma chave-credencial e fornecê-la no ambiente de execução do *Google Colab*. Somente assim a API consegue realizar uma conexão com o *software* de inteligência artificial (no caso o próprio *Gemini*). Por sua capacidade de gerar conteúdos, ao aplicá-lo no nosso código, auxiliada pelos glossários elaborados, os resumos são gerados de forma objetiva, robusta e coesa.

Após o sucesso na execução do código nos testes realizados nas edições do Jornal Cearense de 1846, surgiu um novo desafio. Pois até então foi realizado o *download* de cada edição manualmente, seguido pelo *upload* na nuvem - uma tarefa simples, todavia desafiante frente ao volume total. Então, para dar sequência ao trabalho, como otimizar o *download*⁵ das 2.385 edições do jornal O Cearense?

Retornou-se ao *Gemini* com um *prompt* solicitando um código auxiliar que acessasse o acervo da HDBN e realizasse o *download* de cada edição do jornal, salvando-o diretamente na nuvem. Para tal, a primeira tarefa foi enviar amostras de *links* de edições de anos e décadas distintas para a IA identificar um padrão. Após isso, forneceu-se manualmente o número das edições anuais

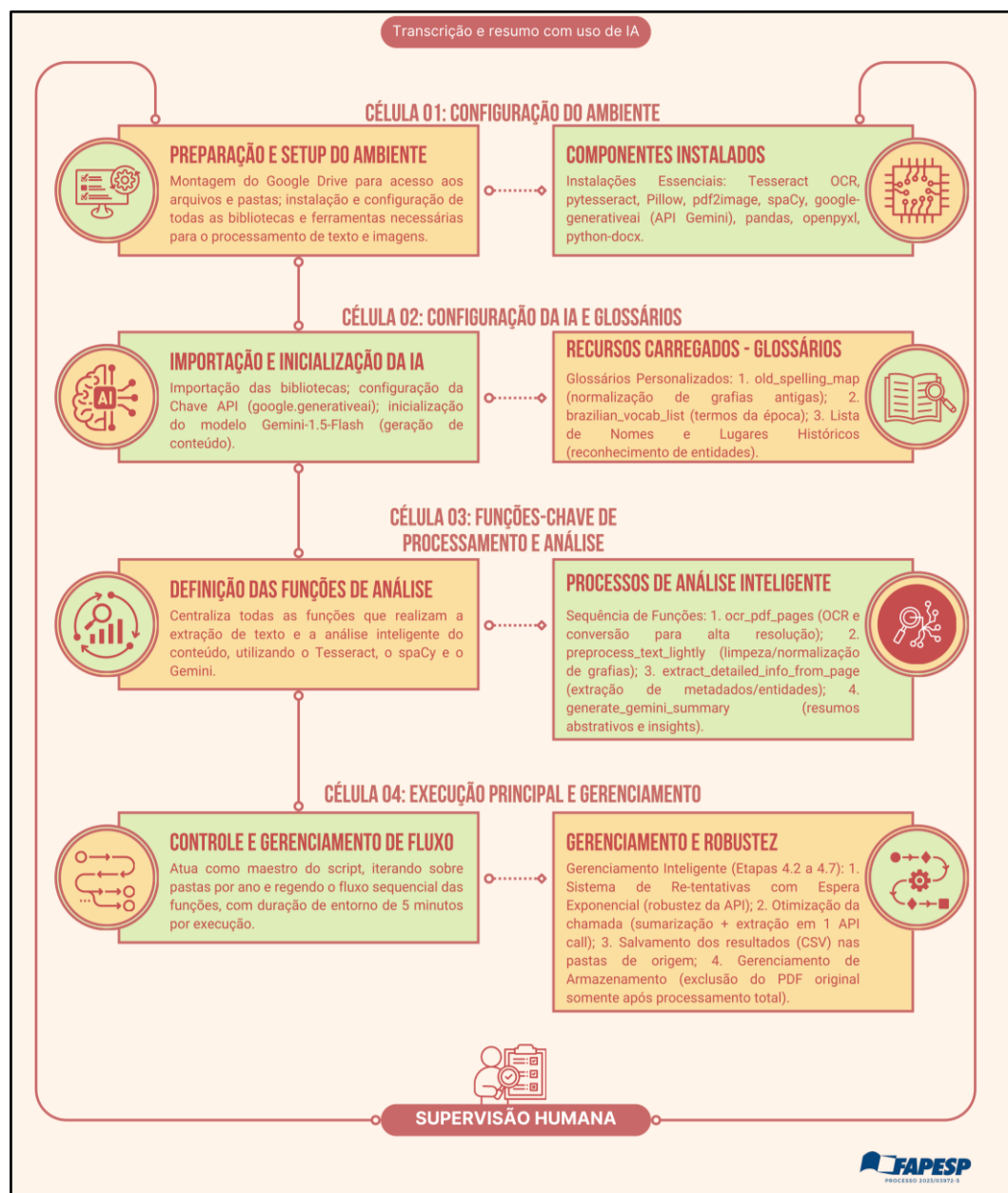
⁵ Além de disponibilizar a pesquisa do seu acervo pela plataforma interativa da Hemeroteca Digital, a Biblioteca Nacional também permite o acesso e o download dos arquivos dos jornais em formato PDF. No caso em tela, disponível em: <https://bndigital.bn.br/acervo-digital/cearense>.

(*ranges*): por exemplo, no ano de 1846 as edições vão de 00001 a 00013; em 1869 vai de 00077 a 00290. E assim sucessivamente no decurso de 1846 a 1891. Na sequência foram criadas pastas no *Drive* separadas anualmente para receber os arquivos.

De modo que o código, ao identificar o padrão de hospedagens dos arquivos, acessava o portal da HDBN, substituindo o padrão do link de hospedagem dos arquivos conforme os anos e o número de edições no intervalo fornecido para cada um, bem como os direcionando automaticamente para o caminho de cada pasta criada. A ação em cada arquivo durava cerca de dois segundos e, como estratégia para evitar erros por sobrecarga, bloqueio do site ou da própria nuvem por movimentação suspeita, o processo dos *downloads* ocorreu ano a ano, e com intervalos automáticos entre uma requisição e outra.

Com a ferramenta de geração de resumos funcionando no nosso código principal, os arquivos a serem analisados hospedados no *drive*, a chave da API *Generative* vinculada no *Google Colab* bastava somente executar o código na íntegra. Em síntese, a versão atual do código é executada no *Google Colab* em quatro células⁶ expostas na Figura 3:

⁶ Para elaboração dessa síntese fomos descrevendo os processos conforme o nosso entendimento e os verificando junto ao *Gemini*, enviando também as células originais, sendo complementado com os termos técnicos.

Figura 3 - Estrutura final do código *Python*

Fonte: O autor (2025).

Como observado no complexo e detalhista fluxograma exposto na Figura 3, o código, em sua versão atual, obtém êxito por ser estruturado de uma maneira focalizada no tipo de material a ser processado, ao tempo que dispõe de mecanismos que asseguram o seu funcionamento saudável. Essa robustez é fundamental para possibilitar uma execução longa sem uma permanente supervisão humana, mas que, por sua vez, é indispensável para essa tarefa.

ACERCA DOS RESULTADOS

Após tratar do longo percurso até a construção do código atual, bem como da explicação da sua estrutura, aventamos abaixo alguns resultados obtidos até o momento, bem como os limites e desafios que se apresentam.

O objetivo está em realizar apontamentos no que diz respeito aos avaliar os avanços obtidos na elaboração do código, bem como do seu funcionamento, da estrutura e forma dos resultados. Afinal, a intenção nesse texto está assentada em realizar um debate metodológico sobre o processamento das fontes e não propriamente uma análise e discussão dos conteúdos dessas, ou seja, nesse caso, dos processos políticos, econômicos e sociais entorno das secas.

Dito isso, para facilitar a visualização e evitar uma sequência de imagens, sintetizamos na Figura 4 o conteúdo de três tabelas cvc. geradas pelo código. Priorizamos para elaboração da Figura as colunas das respectivas tabelas que consideramos mais importantes para essa discussão. De modo que, além das expostas, constam ainda “Nome_Arquivo_PDF_Original”, “Nome_Jornal”, “Página”, “Data_Jornal_Completa”, “Termo encontrado”, “Pessoas” e “Lugares”.

Figura 4 - Printscreen de uma das primeiras transcrições com pré-processamento de imagem

Nome_Arquivo_PDF_Original	Ano_Jornal	Pagina	Termo_Encontrado	Transcricao_Snippet_OCR	Resumo_IA	Informacoes_Seca_Detalhe
1846_00003.pdf	1846		2 secas	nao é de admirar que os vegetaes e mui principalmente os bosques tenha uma immensa influencia sobre os climas, sobre as estações, sobre a fertilidade e salubridade da terra: Rey. Med. Flum, T. 4, 6, 6, p. 20. visto do que deixamos expor, ninguém haverá que nos possa contestar, que as secas em nossa provincia provem, sem menor duvida, de continna e desnecessaria derrubada de nossas or matas para a abertura e plantação de nossos roçados. Não é somente para esse fim que entre nós se fazem cair - ao despiado golpe do machado os mais bellos e frondosos matos, que possuímos: os mais lindos e sombrios arvoreos, que abrigao os nossos campos e viajantes, dos raios do nosso sol ardente, do mal que nos feam, ignorando o mal que nos fazem com um tal procedimento, pelo unico gosto de lamberem um dedo de mel, derriba com o mair or prazer uma arvore, multas vezes bem util vida, para tirarem uma abelha: donde resulta q não so perdemos uma causa principal para a chuva e fert	O texto argumenta que a derrubada de matas na provincia, para roçados e extração de mel, causa secas. A destruição de florestas afeta o clima, a fertilidade e a salubridade da terra, segundo o autor, que cita Rey. Med. Flum, T. 4, 6, 6, p. 20 como referência.	1. Proposta ou solução para a seca. Não especificado. O texto apenas aponta a derrubada de matas como causa da seca, sem propor soluções. 2. Denúncia de vítimas, causas ou impactos negativos específicos da seca. O texto menciona o seca como resultado da derrubada de matas para roçados e extração de mel. Os impactos negativos são a perda da fertilidade e salubridade da terra e o falta de chuva. Não há menção explicita a vítimas específicas. 3. Menção a obras relacionadas à seca. Não especificado. O texto não menciona obras de infraestrutura ou projetos relacionados ao combate à seca. A solução implicita seria a preservação das matas.
Cearense_1877_00027.pdf	1877		3 chuva	garantias, sece, Escrevem-nos do Assaré em 1o do corrente: Atravessamos uma secca assoladora e uma fome horrivel ameaça aniquilar a população destes sertões, destituídos de todos os recur-sos. De Boa Esperança escrevem o seguinte em 25 do passado: Até hoje ainda não tivemos a mais peque-ea chuva, é o pobre povo já começa a morrer á fome. Os generos alimentícios sobem de preço e e-scasalea de um modo espantoso; e se não ochuver nestes quatro dias as plantações estarão iremissivelmente perdidas. A única coisa que nos dava alguma esperanga - era um açude que aqui construiu o padre Ibiapina, mas esse mesmo está de todo secco, le sorte que estamos apanhando agua para beber a s cerca de uma legua de distancia. Estamos em preces e á hora em que faço esta contrange-me o coração o triste espectáculo de alguns homens que se cortão á disciplina, procurando com o oraque que orre á jorros applicar a colera celeste. O sado urra de fome e sêdde, e os retirantes atravessão a cada momen.	Cartas de Assaré e Boa Esperança relatam uma seca devastadora no sertão, causando fome e morte. Em Assaré, a população sofre com a falta de recursos, enquanto em Boa Esperança, a falta de chuva ameaça as plantações e aumenta os preços de alimentos. O açude construído pelo Padre Ibiapina secou, e a população busca água a uma légua de distância. Há relatos de pessoas se mutilando em busca de aplacar a "colera celeste", e muitos estão se tornando retirantes.	1. Sim. Embora haja menção à construção de um açude pelo-o Padre Ibiapina, não há propostas de solução para a seca além depreces. 2. Sim. A seca causou fome e morte, aumento drástico nos preços de alimentos, e fez com que pessoas se mutilassem em atos desesperados. Há também a menção de retirantes, pessoas que abandonam seus lares em busca de sobrevivência. 3. Sim. Há menção à construção de um açude pelo Padre Ibiapina em Boa Esperança, que, no entanto, está seco.
Cearense_1853	1853		4 chuva	ilhantes, Explicaremos ago-ra esse negocio tal qual se nos contou. Disse-nos que o Sr. Presidente mandara comprar porção de farinha para aquelle lado (no que me ô represensu muito bem.) quando farinha estava com effeito 14000, e talvez á maais. Disserão-nos que o Sr. presidente, logo que com a chuva começou, desvanecer-se o recio da secca, e que o prepo da farinha foi descendo, mandara contra ordem para se não comprar mais; mas que enão mandário dizer sa Sr. presidente que já tinham-se cumprido 250 alqueires 44000, tendo á dias o preço descido a ota, ou sete patacas, para o que ss referia pessoa que nos merece creito. Edital. Tendo de ser arrematado o fornecimento de vos remédios ao hospital rigimental, S. Ex. o Sr. Presidente da Provincia manda convidar a aos Srs. farmaceuticos desta capital para que no praso de oilo dias re-ametia a mesma presidência suas propostas em carta feizada acerca do dito fornecimento, tendo em vista o formulário resspectivo que fica exposto na	O Presidente Provincial ordenou a compra de farinha a 14000 (e mais) devido á seca. Após a chuva e queda de preço para 7 patacas, a compra foi suspensa, embora já tivessem sido adquiridos 250 alqueires (44000). Um edital convida farmaceuticos da capital e apresentar propostas para o fornecimento de remédios ao hospital rigimental, dentro de oilo dias, com base em formulário disponivel na Presidência.	1. Propostas/Soluções para a seca. * Compra de farinha para a população afetada pela seca. A compra foi posteriormente suspensa após a chuva e a queda no preço da farinha. 2. Vítimas, Causas e Impactos Negativos. * Menciona-se a seca como causa da necessidade de compra de farinha. * O impacto negativo é a escassez de alimentos causada pela seca, levando á intervenção governamental. * Número de vítimas não especificado, apenas a quantidade de farinha comprada como medida paliativa. 3. Obras/Infraestrutura. * Não especificado. A compra de farinha é uma medida de emergência, não uma obra de infraestrutura para prevenção de secas futuras.

Fonte: O autor (2025).

Os três resultados selecionados para compor a figura síntese são suficientes para tecer algumas considerações. De modo geral, como é possível observar na coluna da transcrição, a limpeza de caracteres estranhos funcionou satisfatoriamente, bem como a grafia da maioria das palavras está correta, possibilitando a leitura do trecho transcrito. Também é possível notar duas palavras-chave encontradas, diretamente envolvidas ao objeto.

Destarte, os principais problemas encontram-se na forma abrupta que o código recorta os trechos. Isso ocorre porque não há uma padronização no jornal para início e fim de notícias. Tentou-se, por vezes, indicar algumas linhas horizontais e os títulos em negrito como padrões que o código poderia utilizar

para interpretar o início e fim de uma notícia. Mas isso não foi bem-sucedido porque, ao longo de quase seis décadas de publicação, o *layout* do jornal foi sendo alterado. Ademais, a questão das rasuras e manchas no documento físico também são fatores que dificultam a interpretação do código em delimitar quando a transcrição deve iniciar e finalizar.

É por essa razão que o esforço se deu em garantir uma maior qualidade no resumo. Para isso, a estratégia foi exceder o trecho transcrito para a sua geração, assegurando que a *API Generative* ao elaborar o resumo mobilizasse elementos de toda página, utilizando ainda os glossários personalizados como fontes para enriquecer o vocabulário e a capacidade de coesão. Logo, os resumos expostos na Figura 02 alcançam uma objetividade necessária frente ao volume de informações a serem analisadas, de modo que rapidamente é possível realizar sua leitura e avaliar se aquele trecho interessa aos objetivos estabelecidos.

Outro ponto importante e necessário nesse processo, é que além da indicação da edição, ano e página da qual o trecho foi transcrito e resumido, se faz possível a verificação quando ao acessar o documento no repositório da HDBN. Para demonstrar essa possibilidade, no Quadro 1 confrontamos um trecho transcrito, resumido e o *printscreen* do arquivo original.

Quadro 1 - Verificação dos resultados e do arquivo original

Transcrição	Arquivo original	Resumo elaborado pela IA
<i>naô é de admirar que os vegelaes e mui principalmente os bosques tenham uma immensa. influencia sobre os climas, sobre as estações, so- bre a fertilidade e salubrida- de da terra.</i> Rey. Med. Flum, T. 4, 0.6. p. 20. vista do que deixamos expendido, ninguém haverá que nos possa contes-tar, que as seccas		O texto argumenta que a derrubada de matas na província, para roçados e extração de mel, causa secas. A destruição de florestas afeta o clima, a fertilidade e a salubridade da terra, segundo o autor, que cita Rey. Med. Flum, T. 4, 0.6, p. 20 como referência.
		Informações sobre as secas organizado por IA

em nossa provincia provem, sem menor duvida, da conti-nua e desnecessaria derribada de nos- sas rmalts para a abertura e plantaço de nossos roçados. Naôd é somente pa-ra este ffm que entre nós se fazem ca-ir ao despiedado golpe do machado os mais bellos e frondosos mattos, q possuimos; os mais lindos e sombrios ar-voredos, que abrigao os nossos campos e viajantes. dos raios do nosso sol ar-dente, Os nossos mattutos, ignorando o mal que nos fasem com um tal procedimento, pelo unico gosto de lamberem um dédo de mel, derribaô com o mai-or prazer uma arvore, muitas vezes bem util vida, para tirarem uma abelha: donde resulta q naô só perdemos uma causa principal para a chuva e fert	A Secca. <i>A' vista do que fica expen-dido não é de admirar que os vegetaes e mui principal-mente os bosques tenhaô uma immensa influencia sobre os climas, sobre as estações, so-bre a fertilidade e salubrida-de da terra.</i> Rev. Med. Flum. T. 4. n. 6. p. 20. A vista do que deixamos expellido, ninguem haverá que nos possa contestar, que as secas em nossa provincia provem, sem a menor duvida, da conti-nua e desnecessaria derribada de nos-sas mattas para a abertura e plantaço de nossos roçados. Naô é somente pa-ra este fim que entre nós se fazem ca-ir ao despiedado golpe do machado os mais bellos e frondosos mattos, q' pos-suimos; os mais lindos e sombrios ar-voredos, que abrigaô os nossos campos e viajantes dos raios do nosso sol ar-dente. Os nossos <i>mattutos</i>, ignorando o mal que nos fasem com um tal proce-dimento, pelo unico gosto de <i>lamberem um dédo de mel</i>, derribaô com o mai-or prazer uma arvore, muitas veses bem util á vida, para tirarem uma <i>abelha</i>: donde resulta q' naô só perdemos uma causa principal para a chuva e fertili-dade do nosso sollo; como um ramo de industria, q' melhor aproveitada, poderia dar á nossa provincia algum in-teresse. Além disto, elles, os nossos escravos, e todos quantos vão á lenha cortaô mul-tas vezes arvores e arbustos verdes pa-ra este fim, em lugar de lançarem uni-camente maô das madeiras, que se a-chaô no estado de servir para este ob-jecto: daqui se segue que até hoje te-mos somente trabalhado para tornar nossa provincia cada vez mais secca, cada vez mais esteril; e por conseguinte redusirmol-a a ser um dia abandonada por nós. Naô são visões o que levamos dito: é uma pura realidade; e oxalá ella, calando no animo dos que estão á testa dos nossos negocios publicos, os faça sair do lethargo, em que se a-chaô, e os impilla a tomar as medidas,	1. **Proposta ou solução para a seca:** Não especificado. O texto apenas aponta a derrubada de matas como causa da seca, sem propor soluções. 2. **Denúncia de vítimas, causas ou impactos negativos específicos da seca:** O texto menciona a seca como resultado da derrubada de matas para roçados e extração de mel. Os impactos negativos são a perda da fertilidade e salubridade da terra e a falta de chuva. Não há menção explícita a vítimas específicas. 3. **Menção a obras relacionadas à seca:** Não especificado. O texto não menciona obras de infraestrutura ou projetos relacionados ao combate à seca. A solução implícita seria a preservação das matas.
--	--	---

Fonte: HDBN (2025). Elaboração: o autor (2025).

Com base no Quadro 1 pode-se atestar a correspondência entre a transcrição realizada pelo código e o resumo gerado pela IA com o arquivo originário. Ainda, a abruptão no recorte da transcrição, não inviabiliza uma compreensão geral do conteúdo e do contexto da notícia.

Quanto ao resumo, o Quadro 1 confirma que a sua elaboração, além de estar fidedigna à notícia transcrita, não é condicionada pela qualidade e pelo recorte da transcrição. Isso também pode ser verificado pela forma de construção

das sentenças, pois também na Figura 4 nota-se que a organização do resumo não segue uma descrição sequencial da notícia, mas sim um apanhado geral. Com isso, se garante que mesmo em casos onde a transcrição esteja ilegível, pelo resumo consegue-se identificar o conteúdo da notícia, bem como se é pertinente ou não a sua consulta no arquivo original.

Destarte, os resumos da Figura 4 igualmente demonstram que a API *Generative* também se debruça sobre os aspectos sociais e políticos entorno das secas e suas consequências. Se o resumo de 1846 trata de possíveis origens do fenômeno ligado ao desmatamento, nos subsequentes de 1853 e 1877, se observa denúncias de fome, preço de alimentos, estado de açudes, etc., abarcando o máximo de aspectos possíveis sobre o nosso objeto.

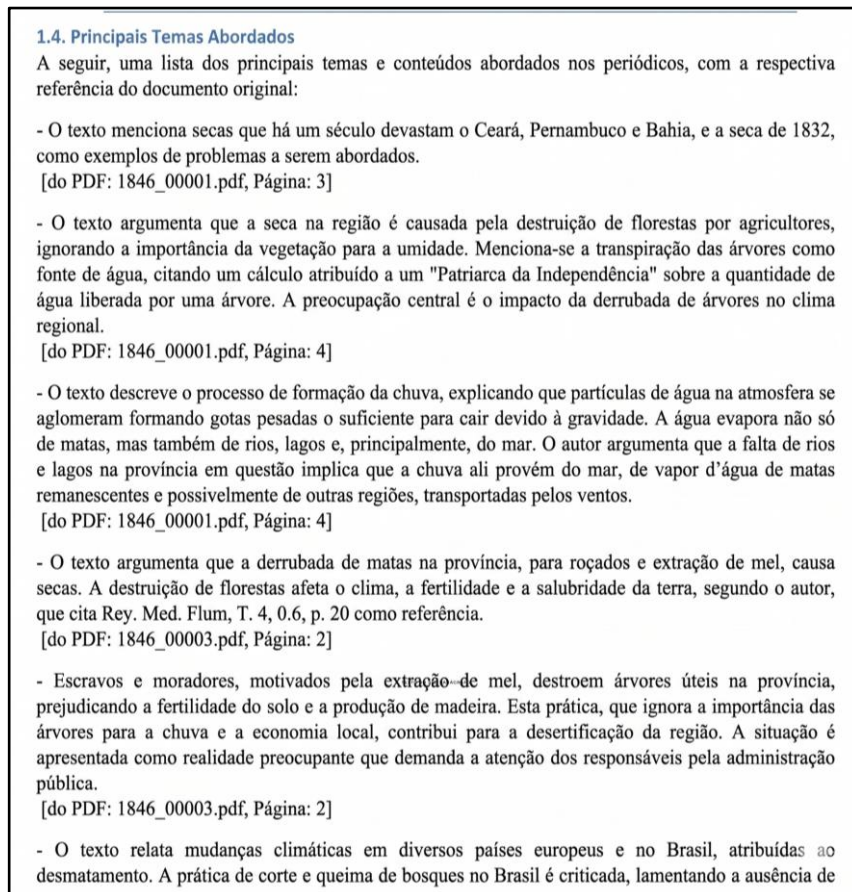
É nesse sentido que foi adicionada outra coluna de informações específicas às secas, construídas a partir de três eixos que o código verifica nas notícias: 1) propostas ou soluções; 2) denúncias de vítimas, causas ou impactos; e 3) menções a obras. Com isso se garante que o material forneça elementos para mapear e analisar essas ações com base nas medidas tomadas conforme cada contexto.

Finalmente, após executado o código em todas as edições do jornal O Cearense, gerou-se um elevado número de tabelas a serem analisadas. A título de exemplo, somente no ano de 1881 foram geradas 96 tabelas de resultados. Para isso, retornou-se ao *Gemini* para elaborar um código auxiliar que realizasse um pré-processamento das informações ao gerar um relatório anual.

Para tal, o *script* coleta os dados brutos nas tabelas (CSV.), com foco nas colunas de resumo e informações sobre as secas; extrai estatísticas e padrões sobre as palavras-chave pré-determinadas; na sequência reúne todos os detalhes da seca extraídos (propostas, denúncias, infraestrutura) e os agrupa por tema. Finalmente, é gerado um arquivo do relatório anual em documento *Word*, contendo seções como título da pesquisa, pesquisadores e instituições envolvidas; introdução; panoramas (termos, pessoas, lugares); as análises temáticas consolidadas pela IA; e uma listagem detalhada de todos os resumos por edição daquele ano. Ademais, todas as informações compiladas são referenciadas diretamente ao arquivo de origem, facilitando a verificação humana.

A título de exemplificação, observa-se o relatório de quinze páginas referente as edições de 1846. Destacou-se na Figura 5 o trecho que sintetiza as informações referentes aos trechos transcritos nos quadros supracitados.

Figura 5 - Página do relatório síntese das edições d'O Cearense de 1846 gerado por IA



Fonte: O autor (2025).

O conteúdo da Figura 5 demonstra, por si só, o potencial desses procedimentos metodológicos envolvendo o processamento de fontes documentais com ferramentas de IA. Além da qualidade na organização e disposição das informações, a referência direta aos documentos originais, assegura sua verificação.

O trecho sobre as práticas de desmatamento para extração de mel e sua relação com a ocorrência de secas aparece referenciado exatamente à mesma página da transcrição da Figura 4 e do Quadro 1, bem como mantém a mesma linha de construção do resumo. Ademais, aparecem nessa página também outros

temas transcritos que não apresentamos anteriormente, como uma notícia sobre processo de formação das chuvas e, no final da página, sobre mudanças climáticas no Brasil e países europeus.

Aí está um ponto importante, o resumo gerado, apesar de treinado com as referências, elementos e termos do contexto do jornal, ainda assim é um código do presente. O termo “mudança climática” é anacrônico ao século XIX, por isso, por mais qualidade que haja no processamento das fontes, é imprescindível o trabalho humano de revisão e verificação dos resultados.

Dito isso, esses elementos referentes aos resultados apresentados, das transcrições, resumos e do relatório, confirmam que o código construído está funcionando conforme o esperado, pois, a partir do exemplo da edição de 1846, pode-se atestar que todas as informações são verídicas, coerentes e, principalmente, passíveis de verificação humana.

CONSIDERAÇÕES FINAIS

O esforço nesse texto esteve em demonstrar como as ferramentas de inteligência artificial podem ser uma poderosa aliada nas pesquisas acadêmicas, especialmente no trato das fontes documentais - tão valiosas para as pesquisas em Geografia Histórica. Haja vista que com essas ferramentas, pode-se potencializar a quantidade e a tipologia de fontes na pesquisa. A título de exemplo, em nosso próprio caso, antes tínhamos de escolher eventos ao longo do recorte temporal onde se aprofundaria a análise das fontes (anos de secas), agora pode-se tomar todo o recorte temporal como objeto de análise a partir das fontes, graças ao processamento realizado pelo código. Fornecendo uma visão mais ampla e detalhista dos processos envolvidos ao objeto.

A aplicação dessas ferramentas permite ao pesquisador(a) trabalhar na análise do material processado ao tempo que o código continua a ser executado, gerando novos dados que serão na sequência analisados. Assim, o tempo dedicado a análise tem melhor aproveitamento e possibilita uma investigação mais profunda e documentalmente fundamentada.

Mas essa não é uma tarefa fácil, especialmente para aqueles com pouca familiaridade com a linguagem da programação e dos inúmeros termos e funções envolvidos a esse universo.

Por isso a opção pela extensa exposição de como se foi construindo o código, os percalços e dificuldades tanto da nossa limitação por desconhecimento, quanto da condição de pensar opções gratuitas para essa tarefa - um dificultante ao processo.

Os limites artificiais estão em dois aspectos: o primeiro é o financeiro, de modo que para executar gratuitamente a geração de resumos e relatórios, há uma cota de 50 chamadas à API *Generative* por projeto criado no *Google Cloud Platform*. A saída para driblar essa limitação está na criação de vários projetos (dez é o limite sem custos), permitindo uma alternância no uso das chaves de API, cuja cota é reiniciada diariamente às 4h. O segundo aspecto limitante está na qualidade das transcrições, em função da já comentada condição material dos arquivos e das dificuldades de interpretação do início e fim das notícias, inviabilizando uma transcrição direta da tabela para uma citação em algum trabalho.

Já os desafios humanos, no que se refere ao desconhecimento da linguagem da programação, as ferramentas interativas de IA, como o *Gemini*, *DeepSeek* e *ChatGPT*, resolvem em grande medida, ao possibilitarem que sejam realizadas requisições de forma mais simples. Nós, por exemplo, pensávamos algo e na sequência perguntávamos a IA se era exequível, e assim fomos e estamos aprendendo como elaborar, ajustar e melhorar os *prompts* para cada tarefa.

Ainda em relação à lida com a elaboração e execução dos códigos, outro desafio está posto na necessidade da verificação humana. É fundamental durante a execução acompanhar o fluxo, bem como quando finalizado, conferir na célula de saída a ocorrência de possíveis erros na execução que podem comprometer a qualidade do processamento.

Outro desafio humano é do ponto de vista ético. As ferramentas de inteligência artificial são auxiliares do processo de pesquisa, é mais uma opção de levantamento, organização, processamento e sistematização de dados. Em nenhum momento devem substituir o trabalho humano de análise e produção

científica, pois, ao mesmo tempo que se faz necessário romper com o receio de utilizá-las, também se deve prezar pela responsabilidade e transparência. Inclusive, esses são temas que “estão na ordem do dia, tanto em termos filosóficos como geopolíticos”, ao requererem “diversas ponderações éticas, como a apropriação indébita, sem devido crédito e/ou com distorção dessas informações publicadas na internet” (AGB, 2025, p. 35).

Finalmente, apropriar e subverter a inteligência artificial é uma necessidade. Utilizá-la como uma poderosa ferramenta para nossas investigações, especialmente para aquelas comprometidas com as transformações sociais, é um caminho fundamental. Inclusive tensionando as contradições das *Big Techs*, bem como - em nosso caso - da ironia de se valer de uma ferramenta potencialmente produtora de secas para auxiliar na realização de uma pesquisa de tema similar.

Afinal, tanto lá no século XIX quanto agora no século XXI, a seca em si não é o problema - as formas de reprodução do capital, sim.

REFERÊNCIAS

ASSOCIAÇÃO DOS GEÓGRAFOS BRASILEIROS - AGB. **Nota técnica:** Orientações para a Promoção da Ética na Produção Científica em Geografia. São Paulo, 2025.

BEAUREPAIRE-ROHAN, H. de. **Vocabulário de termos e expressões brasileiras**. Rio de Janeiro: Imprensa Nacional, 1889.

CARNEIRO, T.; MEDEIROS, R.; NEPOMUCENO, R.; BIAN, G.-B. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. **IEEE Access**, Piscataway, v. 6, p. 61677-61685, 2018. DOI: <http://doi.org/10.1109/ACCESS.2018.2874767>.

EL-KASSAS, W. S.; SALAMA, C.; RAFFA, A.; MOHAMED, H. Automatic text summarization: A comprehensive survey. **Expert Systems with Applications**, Oxford, v. 165, p. 113679, 2021. DOI: <http://doi.org/10.1016/j.eswa.2020.113679>.

FREITAS, I. Inteligência Artificial e profissionais de História (2020-2025): uma revisão sistemática. **Ponta de Lança: Revista Eletrônica de História, Memória & Cultura**, São Cristóvão, v. 19, n. 36, p. 138–156, 2025. DOI: <http://doi.org/10.61895/pl.v19i36.23633>.

GONÇALVES, R.; DORNELES, C. F. Identifying named entity from researcher curricula. *In*: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBd), 37., 2022, Búzios. **Anais** [...] Porto Alegre: Sociedade Brasileira de Computação, 2022 . p. 427-432. DOI: <https://doi.org/10.5753/sbbd.2022.226233>.

GUPTA, J.; BOSCH, H.; VAN VLIET, L. AI's excessive water consumption threatens to drown out its environmental contributions. *In*: SCIENCE-POLICY BRIEF FOR THE MULTISTAKEHOLDER FORUM ON SCIENCE, TECHNOLOGY AND INNOVATION FOR THE SDGs, 9., 2024, New York. **Anais** [...] New York: United Nations, 2024. Disponível em: <https://sdgs.un.org/sites/default/files/2024-05/Gupta%2C%20et%20al.%20AI%20excessive%20water%20consumption.pdf>.

HARVEY, D. Between Space and Time: Reflections on the Geographical Imagination. **Annals of the Association of American Geographers**, Washington, DC, v. 80, n. 3, p. 418-434, 1990. DOI: <https://doi.org/10.1111/j.1467-8306.1990.tb00305.x>.

MATTOS, I. R. de. **O Tempo Saquarema**: a formação do estado imperial. 7. ed. São Paulo: Hucitec, 2017. 314 p.

MORRIS, R. J. Historical Geographers at the Economic History Society, Cheltenham 5 April 1986. **Journal of Historical Geography**, [s.l.], v. 13, n. 1, p. 64-66, 1987. DOI: [https://doi.org/10.1016/S0305-7488\(87\)80008-7](https://doi.org/10.1016/S0305-7488(87)80008-7).

NERY, P. S. Índice Geral do Dicionário Bio-Bibliográfico Cearense de Guilherme Studart. **Revista do Instituto do Ceará**, Fortaleza, v. 85, p. 221-237, 1971. Disponível em: <https://www.institutodoceara.org.br/revista/Rev-apresentacao/RevPorAno/1971/1971-IndiceGeralDicionarioBiobibliograficoGuilhermeStudart.pdf>.

NERY, P. S. Índice Geral do Dicionário Bio-Bibliográfico Cearense de Guilherme Studart. **Revista do Instituto do Ceará**, Fortaleza, v. 86, p. 168-173, 1972.

O'BRIEN, M.; FINGERHUT, H. **Artificial intelligence technology behind ChatGPT was built in Iowa - with a lot of water**. AP News, 9 set. 2023. Disponível em: <https://apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-f551fde98083d17a7e8d904f8be822c4>. Acesso em: 5 ago. 2025.

PARIS, M. **ChatGPT Hits 1 Billion Users? 'Doubled In Just Weeks' Says OpenAI CEO**. Forbes, 12 abr. 2025. Disponível em: <https://www.forbes.com/sites/martineparis/2025/04/12/chatgpt-hits-1-billion-users-openai-ceo-says-doubled-in-weeks/>. Acesso em: 3 ago. 2025.

RED HAT. **What are Application Programming Interfaces (APIs)?** 2023. Disponível em: <https://www.redhat.com/pt-br/topics/api/what-are-application-programming-interfaces>. Acesso em: 21 set. 2025.

RODRIGUES, G. B. Geografia Histórica - notas sobre a metodologia. **Terra Brasilis**, São Paulo, p. 1-20, dez. 2019. DOI: <https://doi.org/10.4000/terrabrasilis.4578>.

SOBREIRA, V. Um panorama da História da Inteligência Artificial e suas aplicações na pesquisa histórica. **Varia Historia**, Belo Horizonte, v. 41, e25035, 2025. DOI: <https://doi.org/10.1590/0104-87752025v41e25035>.

SIMILARWEB. **chatgpt.com Traffic Statistics, Ranking & Analytics**. [s.l.], jun. 2025. Disponível em: <https://www.similarweb.com/website/chatgpt.com/#geography>. Acesso em: 5 ago. 2025.

SCHMIDT, S. Universidades brasileiras discutem regras de uso de inteligência artificial. **Revista Pesquisa FAPESP**, São Paulo, n. 342, ago. 2024. Atualizado em: 11 set. 2024. Disponível em: <https://revistapesquisa.fapesp.br/universidades-brasileiras-discutem-regras-de-uso-de-inteligencia-artificial/>. Acesso em: 2 out. 2025.

STUDART, G. **Diccionario bio-bibliographico cearense**. Fortaleza: Typo-lithographia a vapor, 1910-1915. Disponível em: https://pt.wikisource.org/wiki/Dicion%C3%A1rio_biobibliogr%C3%A1fico_cearense.

TAGHANDIKI, K.; MOHAMMADI, M. **Topic Modeling**: Exploring the Processes, Tools, Challenges and Applications. TechRxiv, 16 jun. 2023. DOI: <https://doi.org/10.36227/techrxiv.23528283.v1>.

Recebido em 06 de agosto de 2025
Aceito em 05 de dezembro de 2025

Financiamento: O autor declara que o presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processo nº 23/03972-5.

Conflito de interesses: O autor declara que não há conflito de interesses na tramitação deste manuscrito no processo editorial.

Disponibilidade de dados: Os dados utilizados estão integralmente apresentados ou descritos no próprio manuscrito.

Editores responsáveis:

Higor Lourenzoni Bonzanini 

Marcelo Costa 

Fernanda Dellaquila de Oliveira 

Murilo Henrique Rodrigues de Oliveira 

Contribuições das autorias:

Igor Carlos Feitosa Alencar: Concepção, análise formal, curadoria de dados e redação.