Estado da publicação: O preprint foi publicado em um periódico como um artigo
DOI do artigo publicado: https://doi.org/10.5007/1518-2924.2025.e101283

Gazetteer literário de Machado de Assis

Dilvan de Abreu Moreira, Davi Machado da Rocha

https://doi.org/10.1590/SciELOPreprints.9474

Submetido em: 2024-07-21
Postado em: 2024-07-22 (versão 1)
(AAAA-MM-DD)

GAZETTEER LITERÁRIO DE MACHADO DE ASSIS

MACHADO DE ASSIS'S LITERARY GAZETTEER.

AUTORIA

Dilvan de Abreu Moreira
Pós-doutor em Informática Biomédica
Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, SP, Brasil
dilvan@icmc.usp.br
https://orcid.org/0000-0002-4801-2225

Davi Machado da Rocha
Mestre em História
Secretaria da Educação do Estado de São Paulo, E.E. Prof. José Juliano Neto, São Carlos, SP, Brasil
davimachado@prof.educacao.sp.gov.br
https://orcid.org/0009-0004-3326-6881

GAZETTEER LITERÁRIO DE MACHADO DE ASSIS

MACHADO DE ASSIS'S LITERARY GAZETTEER.

Resumo

Este estudo tem o objetivo de desenvolver uma aplicação web semântica[i] que mapeia localidades geográficas nas obras de Machado de Assis, armazenando-as em uma triplestore[ii]. A partir da integração dos dados disponibilizados pela enciclopédia *MachadodeAssis.net* com as coordenadas geográficas de *Geonames.org* e *GoogleMaps*, o projeto visa oferecer uma experiência de leitura através de mapas interativos, que servirão de suporte para as menções aos espaços realizadas pelo escritor ao longo do Século XIX. Para a extração das citações, a aplicação utiliza a biblioteca python *BeautifulSoup*[iii] que realiza consultas, requisições e coleta os dados da enciclopédia estruturando-os de acordo com os parâmetros do schema.org. As citações coletadas serão submetidas aos modelos *gpt3.5-instruct*[iv] e *gpt4-turbo*[v] com o intuito de obter os nomes atuais das localidades, bem como a devida classificação destes espaços de acordo com a ontologia *Geonames.org*. Ao final, são realizadas consultas SPARQL ao portal dados.literaturabrasileira.ufsc.br[vi] com o objetivo de obter identificadores únicos para cada livro, oferecendo uma integração entre mapas, citações e textos completos, em consonância com os padrões *Linked Data*[vii].

Palavras-chave: Web Semântica; Machado de Assis; Geolocalização; Literatura Brasileira; Humanidades Digitais.

Abstract

This study aims to develop a semantic web application that maps geographic locations in Machado de Assis's works, storing them in a triplestore. By integrating data from MachadodeAssis.net encyclopedia with geographic coordinates from Geonames.org and GoogleMaps, the project offers an interactive map-based reading experience, supporting the spatial references made by the writer in the 19th century. Using the Python library BeautifulSoup, the application extracts citations, structures them according to schema.org parameters, and submits them to gpt3.5-instruct and gpt4-turbo models to identify current names and classifications of these locations as per Geonames.org ontology. Finally, SPARQL queries are made to dados.literaturabrasileira.ufsc.br to obtain unique identifiers for each book, integrating maps, citations, and full texts in line with Linked Data standards.

Keywords: Semantic Web; Machado de Assis; Geolocation; Brazilian Literature; Digital Humanities.

INTRODUÇÃO

Machado de Assis, um dos mais ilustres escritores do Brasil oitocentista, é reconhecido como uma figura central na literatura em língua portuguesa. Sua obra abrange romances, poesias, peças teatrais e crônicas, marcando profundamente o cenário literário com seu estilo inconfundível e análises perspicazes da sociedade brasileira da época. Atualmente, na área das humanidades, têm se discutido a trajetória ascendente do escritor como expressão do potencial criativo das populações afrodescendentes no Brasil[viii], cujo legado para a formação intelectual do povo brasileiro é muitas vezes negado em razão do histórico escravista que marca indelevelmente a nossa História. Cabe considerar ainda a importância do escritor para o registro dos padrões de sociabilidade e mesmo dos acontecimentos históricos ocorridos durante a sua existência, que acompanha momentos decisivos da História do Brasil[ix], tais como a transição da Monarquia para

a República, a Guerra do Paraguai e o processo de abolição do trabalho escravo. Em razão dessa centralidade, atualmente, há diversos projetos que disponibilizam sua obra em formatos digitais variados, contribuindo para ampliar o alcance de seu legado. Na área das tecnologias da informação, seus escritos têm servido ainda para o treinamento de modelos de Machine Learning[x], chatbots[xi] e outros usos, como o que propomos a seguir, por se tratar de um conjunto que oferece uma escrita apurada e condizente com os mais elevados padrões da língua portuguesa, para além dos fatores já mencionados. [xii]

Assim, na esteira dos projetos de tecnologia que encontram a obra do Bruxo do Cosme Velho, o presente trabalho consiste na construção de uma triplestore de localidades presentes em sua obra, disponibilizadas em forma de verbetes pelo portal MachadodeAssis.net[xiii]. A partir da coleta dos verbetes no referido portal – uma iniciativa assinada pela pesquisadora Marta de Senna e que conta com fontes variadas de financiamento – que se autodefine como uma "enciclopédia" especializada na obra do escritor, procedemos a identificação de coordenadas geográficas em pares de latitude e longitude dos locais referenciados e descritos no site a partir das bases do Geonames.org e do GoogleMaps, propondo outras experiências de leitura através de mapas interativos, que servem, em última análise, como suporte para excertos da obra machadiana.

Para a coleta dos dados, utilizamos a biblioteca BeautifulSoup em python, apontando as tags html da página com os nomes, os verbetes e as citações sobre um dado local em livros diversos. Além disso, armazenamos o resultado da coleta respeitando os parâmetros do schema.org, pensando em facilitar a localização desses registros pela engine do google, o que pode proporcionar, por exemplo, o melhor ranqueamento dos registros geográficos nesse buscador. Por fim, por se tratar de um tipo de dado específico (entidade geográfica) também usaremos a ontologia do geonames para a descrição dos objetos e classes, oferecendo uma organização mais completa em termos semânticos e maior perenidade dos registros, considerando eventuais mudanças de domínio, por exemplo.

CASO DE USO

Objetivo: identificar espaços (localizações) na obra de um escritor

Atores:

1. Pesquisadores nas áreas de literatura e história.
2. Usuário leigo interessado em literatura.

Interação:

1. Entrar em uma página Web
2. Selecionar a obra e o tipo de mapa
3. Interagir com os controles do mapa
4. Ao selecionar o pino no mapa, vê as informações:
    1. Descrição do espaço pela enciclopédia

2. O texto do escritor onde ele referencia o lugar
3. Frequência que o local aparece na obra (mapa de calor)
4. Visualização do local com a API do Google Street View

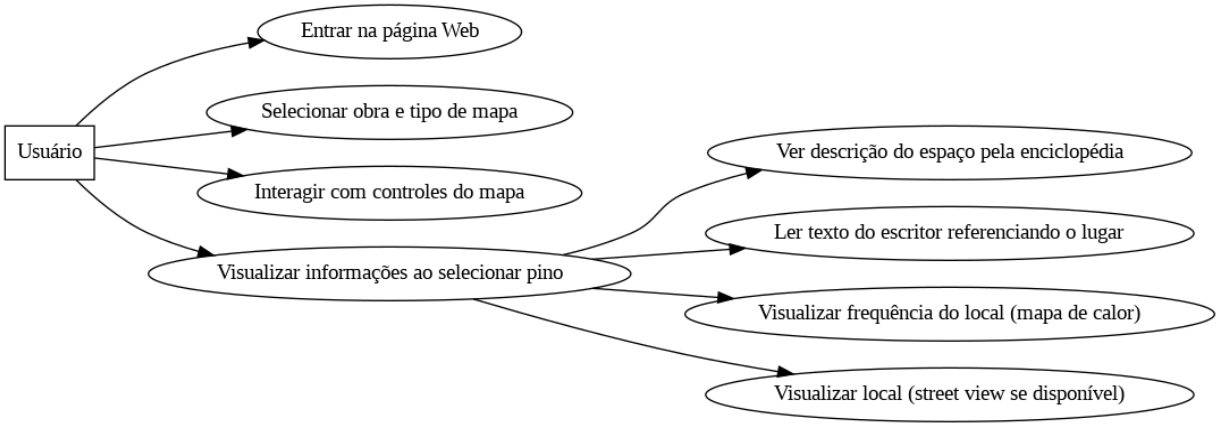

Imagem 1: Diagrama de Caso de Uso.

Em termos de casos de uso, cumpre considerar que as relações entre espaço geográfico e análise literária possuem um lugar na produção acadêmica nacional e estrangeira. Dentre esses trabalhos, destaca-se a obra do pesquisador Franco Moretti no Atlas do Romance Europeu[xiv] que, nas palavras dos editores, "sugere que o espaço pode ser um protagonista oculto na história cultural. [...] Mais que um simples estudo literário, é um novo caminho para a leitura e compreensão da literatura."[xv] Apresentando mapas e conceitos espaciais importantes, como as noções de centro e periferia, por exemplo, fundamentais para a construção de personagens ou hábitos específicos de certos grupos sociais retratados, bem como as mudanças realizadas nas cidades ao longo do tempo, pesquisadores interessados no espaço geográfico e social das obras literárias são usuários possivelmente interessados nessa solução, além dos próprios leitores do portal Machado de Assis.net que, como dito, poderão desfrutar de uma experiência de leitura usando mapas, distâncias, frequências de citações e outras métricas possíveis para a fruição da obra do escritor a partir da linguagem cartográfica, como tentaremos mostrar a seguir.

COLETA E FORMATAÇÃO DOS DADOS SEGUNDO OS PARÂMETROS SCHEMA.ORG

Schema.org é uma ontologia colaborativa, mantida pelo Google, Microsoft, Yahoo! e Yandex. Ela fornece um conjunto de vocabulários para descrever entidades, relacionamentos e ações no mundo real. Esses vocabulários podem ser usados para enriquecer o conteúdo da web com metadados estruturados, que podem ser interpretados por máquinas. Os metadados estruturados podem ser usados para uma variedade de propósitos, incluindo:

a) melhorar os resultados de pesquisa: os motores de busca podem usar metadados estruturados para entender melhor o conteúdo de uma página da web. Isso pode levar a resultados de pesquisa mais relevantes e úteis para os usuários.

b) criar experiências mais ricas para os usuários: os metadados estruturados podem ser usados para entregar resultados de pesquisa enriquecidos com informações adicionais ou conteúdo interativo.

c) facilitar a automação: os metadados também podem facilitar a automação de tarefas, como a classificação de conteúdo ou a geração de relatórios.

Com isso em mente, o pipeline a seguir foi projetado para extrair e estruturar dados de lugares mencionados no site "machadodeassis.net". O objetivo é criar um arquivo JSON-LD que contenha informações sobre os lugares e as obras que os mencionam. No início, são realizadas as importações das bibliotecas necessárias: requests (para fazer solicitações HTTP), json (para manipular dados JSON) e BeautifulSoup (uma biblioteca para análise de HTML). Também é definida uma URL específica do site, que é onde o script buscará os lugares referenciados. Assim, uma solicitação GET é feita para essa URL usando a biblioteca requests e o conteúdo da resposta é analisado pelo BeautifulSoup para facilitar a extração de informações.

O script procura todos os elementos div que têm a classe "content-card-wrap" reference. Cada uma dessas divs representa um lugar referenciado no site. Para cada div encontrada, o título (nome do lugar) e o link associado são extraídos. Em seguida, uma nova solicitação GET é feita usando o link extraído para obter detalhes adicionais sobre o lugar. Do novo conteúdo carregado, o script extrai o type-id e o info-text, que fornecem informações adicionais sobre o lugar. Além disso, o código verifica se o título (lugar) já existe em um dicionário chamado structured_data. Se não existir, uma nova entrada é criada para o título com detalhes como nome, URL e descrição. Ao final desta etapa, realiza-se a busca por informações sobre as obras (romances, contos, crônicas...) associadas a esse lugar. Para cada obra encontrada, são extraídos o título da obra, o tipo, o ano de publicação e o texto referenciado.

- Lugar
    - Nome (título)
    - Descrição
    - Lista de títulos das obras que referenciam esse lugar
- Obra
    - Nome (título)
    - Ano
    - Gênero
    - Citações do lugar no livro

Estas informações da obra são então estruturadas como CreativeWork (conforme o schema.org) e adicionadas à lista subjectOf da entrada correspondente do lugar em structured_data. A propriedade "subjectOf" é uma inversão da propriedade "about", e refere-se ao item que é o assunto da coisa (por exemplo, um Place que é o assunto de um CreativeWork). Finalmente, o script salva todos os dados estruturados no formato JSON-LD em um arquivo chamado 'structured_data.jsonl'.

Para a definição dos identificadores dos livros em que os locais são citados, faremos a coleta das URL's das obras no portal http://dados.literaturabrasileira.ufsc.br/ a partir de uma query SPARQL, pensando também em facilitar o acesso aos textos integrais pelos leitores. Em uma situação ideal, o @id seria o ISBN do livro, no entanto, por razões que envolvem a complexidade desse tipo de coleta, considerando o volume de obras e a quantidade de edições e reedições ao longo do tempo, a própria definição do ISBN adequado demandaria uma pesquisa aprofundada. Por isso, entendemos que a definição do ID a partir da URI de uma versão digital das obras hospedadas por uma instituição referenciada como a Universidade Federal de Santa Catarina é a melhor alternativa para essa atividade, por dois motivos principais: facilidade de acesso aos textos integrais em versão digital e sua estruturação consonante com as melhores práticas em Web Semântica pelo referido portal.

Assim, para coleta e atribuição dos IDs, utilizamos uma query SPARQL que obtém os campos URI, títulos e títulos alternativos, pensando em situações de obras que componham um mesmo livro, como é o caso das crônicas e contos. As obras que possuem um mesmo URI no portal da UFSC, terão a propriedade OWL "partOf" adicionadas a triplastore e a inserção do título na URI em um formato do tipo #nome+do+conto, garantindo URIs únicos para cada item na triplastore.

Em resumo, cada obra listada no site machadodeassis.net é mapeada para a respectiva URI na base de dados da UFSC, utilizando um mapeamento manual para alguns títulos que têm grafias ligeiramente diferentes. Quando uma URI é encontrada com base em um título alternativo, o título principal correspondente é adicionado como book e o título da obra é considerado partOf deste book. Entradas com 'N/A' são removidas antes de adicionar as informações em structured_data.

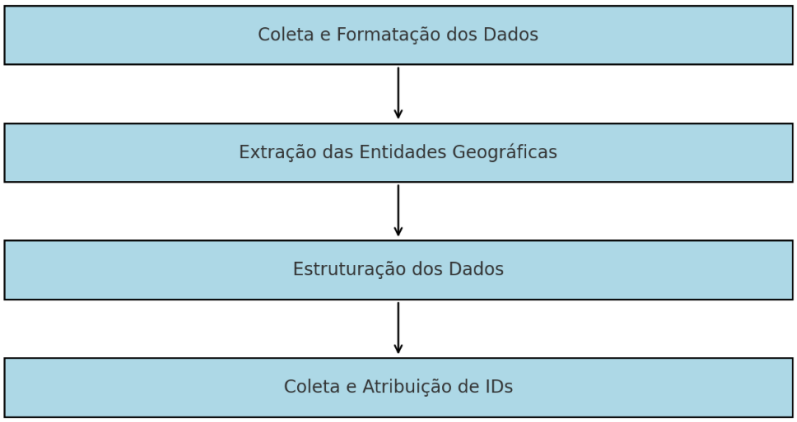

Imagem 2: Fluxograma de processamento de dados literários.

INTERPRETANDO VERBETES COM GPT-3.5-INSTRUCT

Em um primeiro momento, considerando que estamos tratando nomes de locais que podem ter sido alterados ao longo do tempo, submeteremos o verbete da enciclopédia MachadodeAssis.net para validação pelo modelo gpt-3.5-instruct. Tal escolha dentre os vários endpoints[xvi] existentes, se deu pela capacidade desse modelo em oferecer saídas sistemáticas e estruturadas para facilitar a localização dos nomes dos locais pelas APIs dos serviços de georreferenciamento. Em linhas gerais, estamos propondo que o modelo identifique o nome atual do local a partir da interpretação do verbebte. Essa lista de nomes validada por gpt-3.5-instruct será utilizada para coleta dos pares de latitude e longitude nas bases do GeoNames e do Google Maps, como forma de evitar ambiguidades e melhorar a acurácia da coleta.

O modelo InstructGPT, desenvolvido pela OpenAI, se destaca dos modelos anteriores por ser especificamente treinado para seguir instruções humanas de maneira mais alinhada e eficaz. Ele utiliza um processo de fine-tuning com feedback humano para alinhar melhor os outputs do modelo com as intenções do usuário. Isso é realizado coletando demonstrações de comportamento desejado de labelers e depois utilizando aprendizado supervisionado e aprendizado por reforço com feedback humano para ajustar o modelo. Essa abordagem resulta em melhorias significativas na capacidade do modelo de produzir respostas verdadeiras, reduzir alucinações e seguir instruções de maneira mais precisa, ainda que com menos parâmetros que modelos anteriores como o GPT-3.

No quadro a seguir, temos o exemplo de um verbete sobre o local chamado “Aterrado” na obra de Machado de Assis e descrito no referido portal conforme se lê, bem como a interpretação realizada pelo modelo a partir da instrução, que usou os parâmetros max_tokens: 50, (para focar em respostas mais diretas) e temperature: 0.5 (para equilibrar criatividade e precisão):

{"role": "user", "content": = f"Dada a seguinte descrição, responda: nome atual do lugar, cidade e país quando aplicável? Nome do lugar: Aterrado. Descrição: “No vasto alagadiço que era o mangue da Cidade Nova, desde o antigo Rossio Pequeno (atual praça Onze de Junho) foi construído um longo e estreito aterro. No tempo de D. João VI era o caminho usado pela família real para chegar a São Cristóvão. Assim foram adotados os nomes de "caminho das Lanternas" e "rua do Aterrado". A rua do Aterrado desapareceu com a construção da avenida Presidente Vargas, na década de 1940."
{"role": "assistant", "content": { "Nome atual do lugar: Praça Onze de Junho, Cidade: Rio de Janeiro, País: Brasil"

Assim, o output gerado está pronto para ser submetido às API’s de georreferenciamento do Geonames e do Google Maps, sendo este último utilizado somente para os casos não localizados na primeira base de dados.

OBTENDO COORDENADAS GEOGRÁFICAS COM GEONAMES E GOOGLE MAPS

Durante a abordagem, identificou-se que a coleta em GeoNames apresenta maior dificuldade em identificar locais específicos das cidades, como bairros, praças, ruas e outros logradouros se comparado ao GoogleMaps. Apesar de serem soluções parecidas, GeoNames e GoogleMaps descrevem de formas diferentes seus objetos, sendo GeoNames uma ontologia específica para entidades geográficas e GoogleMaps um serviço de geolocalização que não usa quaisquer ontologias de web semântica para descrever seus objetos. Assim, todos os dados coletados via Google Maps deverão ser submetidos ao GeoNames para a definição do código de característica, isto é, a definição desta ontologia para cada tipo de entidade geográfica. Ao final, o que não for localizado será identificado pela URL do Google Maps.

O pipeline representado a seguir chama a API GeoNames para enriquecer entradas JSON-LD com informações geográficas. Primeiro, são importados os módulos necessários: json para manipulação de arquivos JSON e geocoder para a interação com a API GeoNames. Em seguida, uma função chamada get_geonames_data é definida para aceitar um nome de lugar (place_name) como argumento. Se a consulta for bem-sucedida, a função constrói um dicionário que contém informações como tipo, nome, latitude, longitude, código de característica (se a entidade é uma Cidade, Estado, País e etc segundo a nomenclatura da base), além de uma URI específica do GeoNames para aquele local. A URI é criada usando o 'geonameId' da resposta da API. Em suma, a função retorna o dicionário de dados geográficos.

As informações geográficas são adicionadas ao dicionário sob a chave "geo" e, assim, a função processa recursivamente cada item da lista de modo a adequar os dados à estrtura do schema.org, que possui a propriedade GeoCoordinates como uma subclasse do tipo "geo", que armazena detalhes relativos à localização de entidades geográficas nesta ontologia.

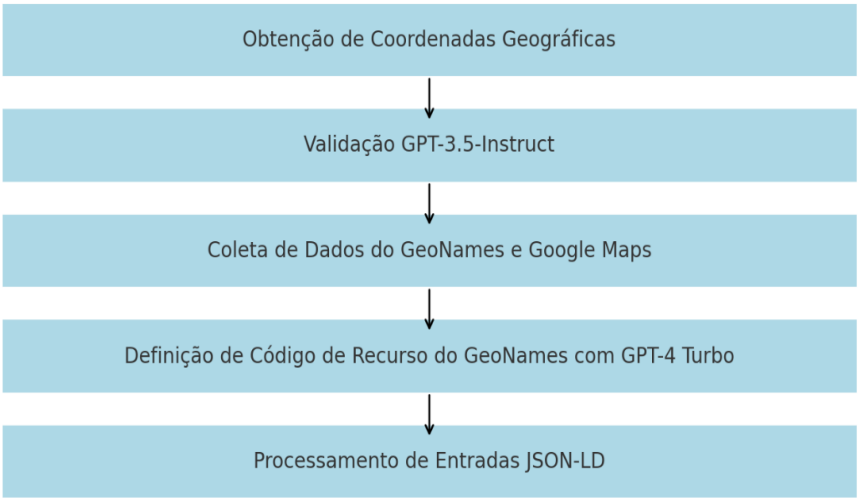

Imagem 3: Fluxograma de processamento de dados geográficos.

GPT-4 COMO CLASSIFICADOR DE DADOS GEOGRÁFICOS.

Os modelos GPT-4 representam a vanguarda da tecnologia de processamento de linguagem natural, oferecendo capacidades avançadas de compreensão e geração de texto. Eles são adequados para uma ampla gama de aplicações, desde a criação de conteúdo até o desenvolvimento de interfaces conversacionais inteligentes. No contexto da aplicação, para cada registro coletado via GoogleMaps será enviada uma requisição de classificação do local pelo referido modelo nos termos da ontologia Geonames.

Inicialmente, a classe Gpt4Turbo interage com a API e seu construtor (init) inicializa o modelo GPT-4 a ser usado (gpt-4-1106-preview), definindo um limite de tokens, e criando um cliente de chat. O método gptCall_json é responsável por enviar mensagens ao GPT-4 e receber respostas. Ele aceita parâmetros como temperatura da resposta (criatividade), se a resposta deve ser contínua (streaming) e uma lista de mensagens (perguntas e comandos para o GPT-4). Em caso de sucesso, retorna as respostas do GPT-4. Em caso de erro, imprime a exceção.

Desse modo carrega-se o arquivo JSON que contém dados de locais. Para cada local ("Place") no arquivo, ele constrói uma mensagem para perguntar ao GPT-4 sobre o código geográfico (gn:featureCode) e o nome do código geográfico (gn:featureCodeName) mais adequados para esse local. As mensagens são enviadas ao GPT-4, e as respostas são processadas e integradas de

volta aos dados do local no arquivo JSON. Após processar todos os locais, o arquivo JSON é atualizado com as novas informações obtidas do GPT-4. O prompt enviado ao modelo tem a seguinte estrutura:

{"role": "system", "content": "Você é um assistente de classificação de dados geográficos e responde no formato JSON."}
{"role": "user", "content": = f"Qual é o gn:featureCode e o gn:featureCodeName mais adequado do GeoNames para o local chamado '{place_name}' descrito como: {description}?"
{"role": "assistant", "content": { "gn:featureCode": "PCLI", "gn:featureCodeName": "independent political entity"}

RESUMO DAS ETAPAS DE CLASSIFICAÇÃO DE ENTIDADES GEOGRÁFICAS:

a) A classe Gpt4Turbo é definida no arquivo gpt4turbo.py.

b) O construtor (init) da classe Gpt4Turbo aceita um parâmetro opcional de limite de tokens. O limite de tokens é o número máximo de tokens que o GPT-4 pode gerar em uma resposta.

c) O método gptCall_json da classe Gpt4Turbo aceita parâmetros para controlar a temperatura da resposta, o streaming e a lista de mensagens.

d) O arquivo JSON é carregado usando a biblioteca json.

e) O código usa um loop para iterar sobre os locais no arquivo JSON.

f) Para cada local, o código constrói uma mensagem para perguntar ao GPT-4 sobre o código geográfico e o nome do código geográfico.

g) As mensagens são enviadas ao GPT-4 usando o método gptCall_json da classe Gpt4Turbo.

h) As respostas são processadas e integradas de volta aos dados do local no arquivo JSON que é atualizado e salvo.

CONVERSÃO DE JSON-LD PARA TURTLE COM APACHE JENA

Apache Jena é um framework[xvii] para trabalhar com dados ligados e semânticos que suporta vários formatos, incluindo JSON-LD e Turtle. JSON-LD é um formato baseado em JSON, muito útil para representar informações de maneira organizada e legível tanto para máquinas quanto para humanos. Já Turtle é um formato de serialização para dados RDF, focado em ser conciso e facilmente legível, comumente usado em aplicações de dados ligados por ser mais compacto e mais fácil de escrever e entender do que outros formatos RDF, como RDF/XML.

Para converter JSON-LD para Turtle, usamos o riot, uma ferramenta do Apache Jena que transforma a estrutura de dados de um formato para o outro, mantendo a semântica, e alterando a sintaxe. Há algumas razões para converter JSON-LD para Turtle, dentre as quais, destaca-se a facilidade de leitura e escrita: Turtle, com sua sintaxe mais concisa facilita a escrita e leitura de consultas SPARQL. Isso torna mais simples trabalhar com dados ligados, especialmente para consultas complexas. Cabe ressaltar ainda a compatibilidade e eficiência, pois alguns sistemas e ferramentas que trabalham com SPARQL podem ser otimizados para trabalhar com Turtle. Por fim, destaca-se a padronização, já que Turtle é um formato amplamente adotado para representar dados RDF, o que favorece a interoperabilidade, possibilitando a integração e o compartilhamento de dados.

CONSULTA AO ARQUIVO DE DADOS

Para realização de consultas SPARQL, inicialmente, é preciso definir os prefixos e namespaces. No contexto desta aplicação, temos o "PREFIX schema: http://schema.org/", isto é, um prefixo chamado schema para o namespace http://schema.org/. O uso de prefixos em SPARQL é uma maneira conveniente de abreviar URIs longas. No restante da consulta[xviii], qualquer vez que schema: for usado, ele se refere a http://schema.org/.

SELECT * WHERE é a parte principal da consulta SPARQL. SELECT * significa que você quer selecionar todas as variáveis disponíveis (?sujeito, ?predicado, ?objeto) nos padrões de triplas correspondentes. WHERE é a cláusula onde se define um padrão de tripla que se quer procurar no grafo RDF. Dentro da cláusula WHERE, temos dois padrões de triplas: o primeiro, ?s schema:name "África" procura por todas as triplas no grafo RDF onde o predicado é schema:name e o objeto é o literal "África". A variável ?s será qualquer sujeito que tenha "África" como um schema:name.

Já a notação ?s ?p ?o aponta um padrão mais geral e procura por todas as triplas no grafo RDF onde o sujeito (?s) é o mesmo encontrado na primeira tripla (ou seja, qualquer sujeito que tenha "África" como schema:name). ?p e ?o são variáveis que representam respectivamente qualquer predicado e objeto associados a esse sujeito. Portanto, a consulta SPARQL busca todas as triplas no grafo RDF onde o sujeito tem um nome "África" segundo o schema.org, e retorna todas as informações (predicado e objeto) associadas a esses sujeitos como localização, descrições, e as citações ao local na obra de Machado de Assis.

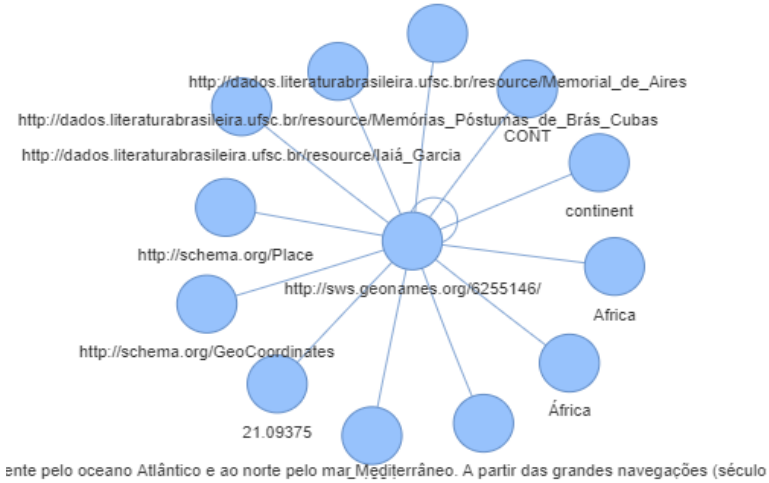

Imagem 4: Grafo com os dados da consulta ao schema:name África.

VISUALIZAÇÕES

Uma vez coletados os dados de Latitude e Longitude e estruturadas as informações dos textos e localidades, tentaremos demonstrar as visualizações possíveis a partir da estruturação dos dados. Inicialmente, serão criados pinos em um mapa, onde cada pino carrega e exibe o verbete sobre o local no portal MachadodeAssis.net. Cumpre considerar que, apesar do nosso esforço em validar as coordenadas geográficas, alguns pinos estão em locais errados, como no "Morro do Castelo", acidente geográfico que ocupava a área central do Rio de Janeiro e que foi removido em 1922 em um amplo movimento de reestruturação da então capital do Brasil. No processo de coleta, GeoNames identificou um local homônimo na Ilha de Paquetá. Há outros erros do tipo e inconsistências que foram tratadas manualmente quando identificadas, uma vez que os modelos da OpenAI usados aqui são sensíveis a alterações de temperatura e podem oscilar na apresentação dos resultados.

MAPA DE LOCAIS COLETADOS COM DESCRITORES DAS LOCALIDADES

Imagem 5: Mapa de locais citados no conjunto da obra

MAPA DE CITAÇÕES POR LOCAL

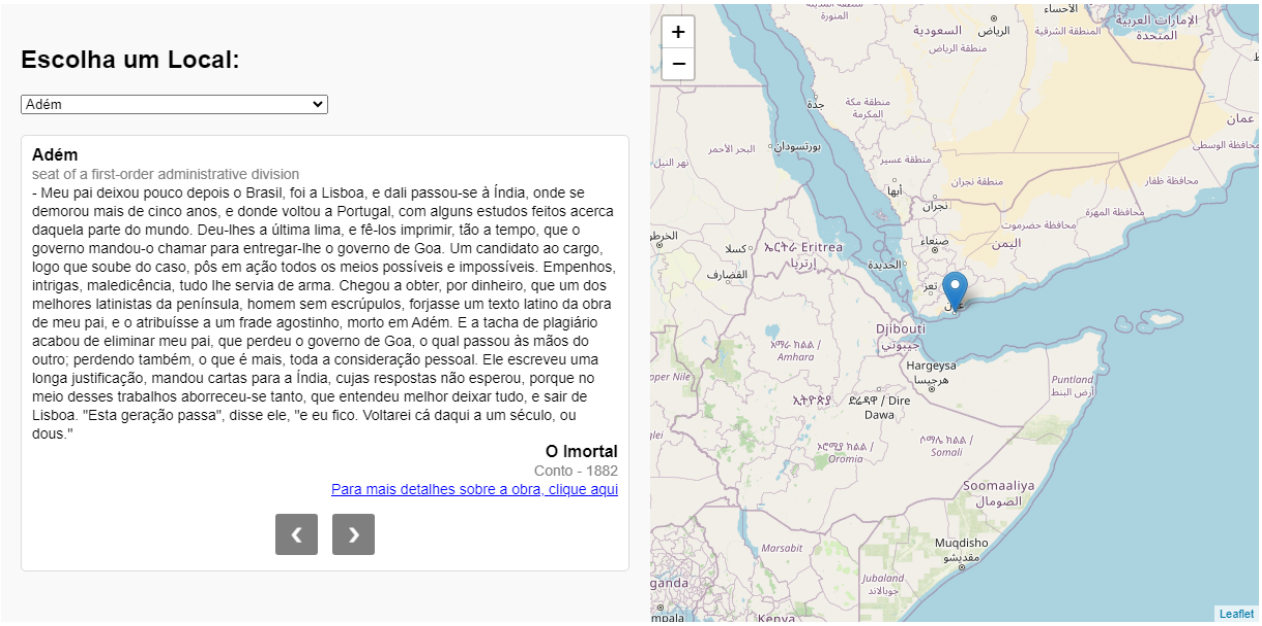

Imagem 6: Mapa de citações por local.[xix]

MAPA DE CALOR COM O CONJUNTO DAS CITAÇÕES

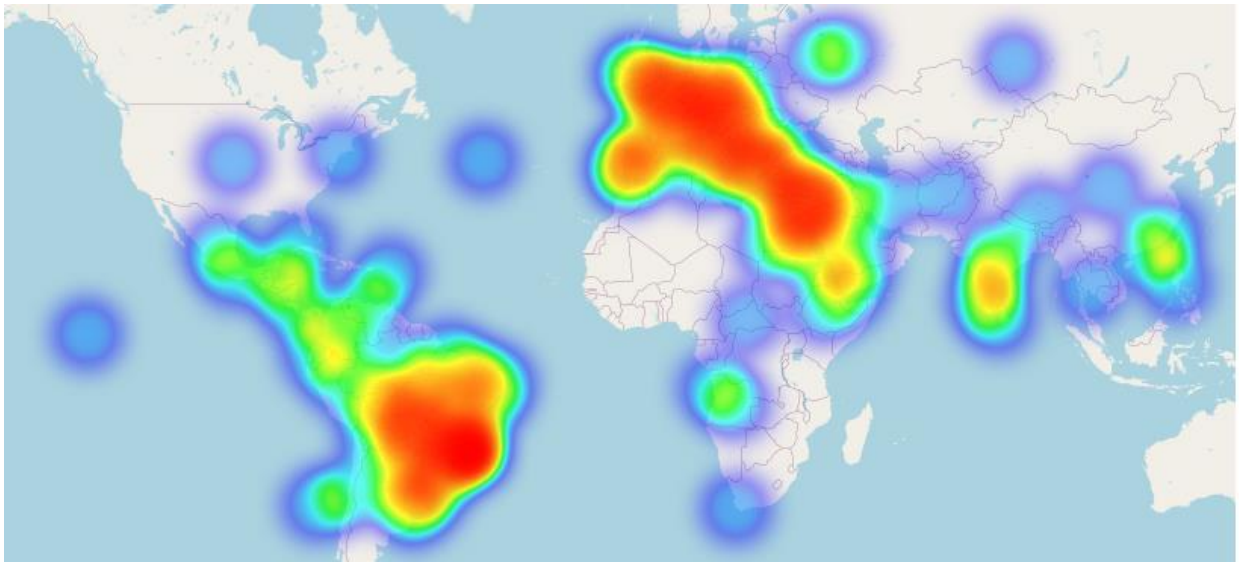

Imagem 7: Mapa de calor com a frequência de locais citados no conjunto da obra.

STREET VIEW PANORAMA

Imagem 8: Visão de local e citação sobre Copacabana com Google StreetView.

CITAÇÕES POR OBRA

Escolha uma Obra:

Imagem 9: Mapa de citações por obra

MAPA DE CALOR POR OBRA

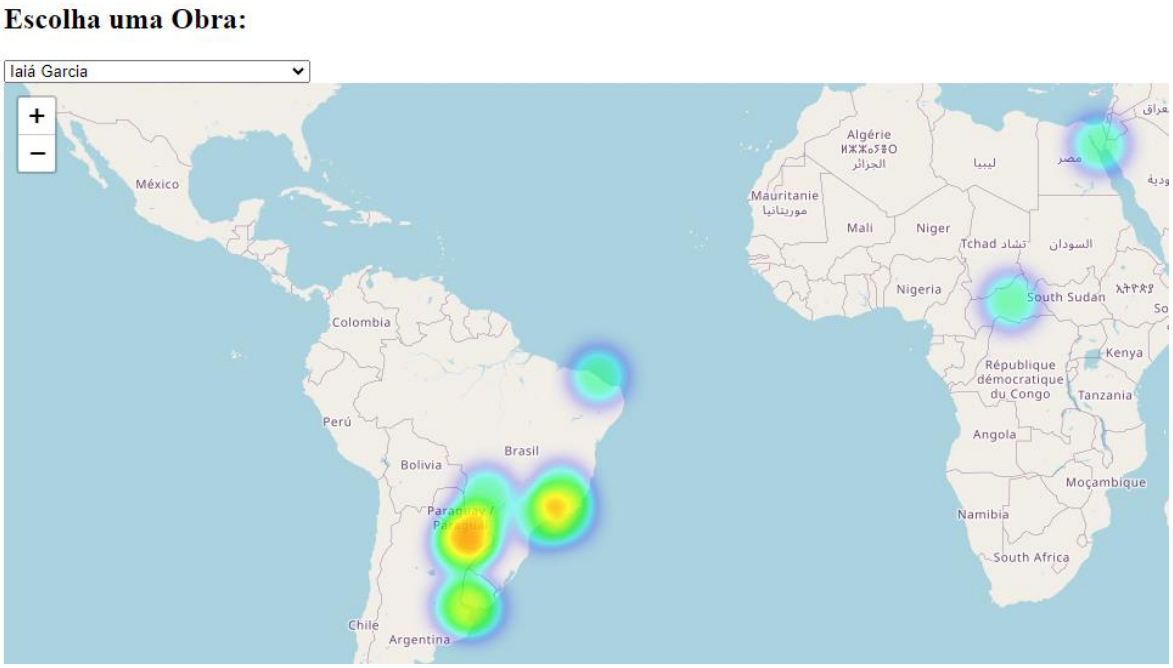

Imagem 10: Mapa de calor de citações por obra

Considerações finais

Como tentamos demonstrar, a intersecção entre tecnologia, literatura e geolocalização pode oferecer experiências de leitura interessantes, muitas delas inexploradas nesse estudo, o que proporciona um terreno fértil para o desenvolvimento das chamadas humanidades digitais. Ao unir essas diferentes áreas do conhecimento, pode-se enriquecer as narrativas literárias com uma dimensão espacial palpável, bem como explorar novas maneiras de engajar e interpretar textos clássicos e contemporâneos. Nesta atividade, por exemplo, pudemos mergulhar em algumas obras de Machado de Assis, não apenas seguindo a trama de forma imaginária, mas também acompanhando os passos dos personagens em mapas digitais que falam do universo de possibilidades conhecidas e imaginadas por um dos grandes expoentes do Século XIX no Brasil.

À medida que a história se desenrola, o leitor pode explorar os locais reais ou fictícios mencionados, compreendendo melhor o contexto cultural e histórico que molda a narrativa. Essa imersão geográfica concede uma camada de compreensão que permite que a literatura respire fora das páginas dos livros e ganhe vida em outros suportes. Nesse contexto, a tecnologia é apenas uma lente através da qual podemos reexaminar obras clássicas sob uma luz diferente. Com isso,

não se pretende substituir o contato com o livro, mas complementar a experiência de leitura a partir de outros métodos e ferramentas para questionar, compreender e apreciar a literatura.

Do ponto de vista dos estudos acadêmicos em tecnologia da informação o presente artigo busca oferecer conceitos, abordagens e mesmo incentivar projetos multidisciplinares que envolvam a estruturação de dados segundo os parâmetros da chamada Web Semântica a fim de enriquecer o corpo de conhecimento qualificado disponível. Ao empregar padrões como RDF (Resource Description Framework), SPARQL (uma linguagem de consulta) e ontologias OWL (Web Ontology Language), é possível estruturar dados literários de maneira que máquinas possam "entender" e processar relações complexas, facilitando a pesquisa interdisciplinar e permitindo a criação de redes de conhecimento que podem revelar padrões e tendências previamente ocultos dentro de grandes volumes de texto.

Em um aspecto mais técnico, o desafio está em como representar e interligar dados de forma que eles sejam ao mesmo tempo acessíveis para análise computacional e visualização humana. O uso de Linked Data e padrões de interoperabilidade assegura que diferentes conjuntos de dados possam ser combinados, fornecendo um quadro mais completo e multidimensional de informações.

Por fim, a Web Semântica não é apenas um instrumento para melhorar a pesquisa e educação nas humanidades digitais, mas também um convite à colaboração transdisciplinar. Ela encoraja o diálogo entre cientistas da computação, bibliotecários digitais, historiadores, geógrafos e literatos, todos contribuindo com suas perspectivas únicas para a construção de uma infraestrutura de conhecimento mais rica e conectada.

REFERÊNCIAS

ADIBOZZI, Daniel; et al. Towards a Human-like Open-Domain Chatbot. [S.l.]: **Google Research**, 2020. Disponível em: https://research.google/pubs/towards-a-human-like-open-domain-chatbot/. Acesso em: 22 de setembro de 2023.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. In: _____. **Semantic Web – Interoperability, Usability, Applicability**. 2011. Disponível em: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. Acesso em: 3 de Outubro de 2023.

CHALHOUB, Sidney. **Machado de Assis**: historiador. São Paulo: Companhia das Letras, 2003.

DIEGO, Marcelo. Entrevista com Marta de Senna. **Machado de Assis em Linha**, São Paulo, v. 13, n. 29, p. 181-189, abr. 2020. DOI: 10.1590/1983-68212020132913. Disponível em: https://doi.org/10.1590/1983-68212020132913. Acesso em: 20 de agosto de 2023.

DO NASCIMENTO, João Gabriel. O branco imposto e o negro conquistado: Machado de Assis na propaganda da Caixa Econômica Federal. **Revista da Associação Brasileira de Pesquisadores/as Negros/as** (ABPN), v. 8, n. 20, p. 74-85, 2016.

GROVER, Claire; TOBIN, Richard. A Gazetteer and Georeferencing for Historical English Documents. In: _____. **Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities** (LaTeCH) @ EACL 2014. Gothenburg, Sweden: Association for Computational Linguistics, 26 Abril 2014.

HETLAND, Magnus. Python and the Web. In: _____ **Beginning Python From Novice to Professional**. New York: Apress, 2005. pp. 313–339. Disponível em: https://doi.org/10.1007/978-1-4302-0072-7_15. Acesso em: 19 de Novembro de 2023.

ILIAIDIS, A.; ACKER, A.; STEVENS, W. **One schema to rule them all**: How Schema.org models the world of search. Journal of the Association for Information Science and Technology, 2022. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24744. Acesso em: 12 de Setembro de 2023.

MORETTI, Franco. **Atlas do Romance Europeu 1800-1900**. São Paulo: Boitempo, 2003. Disponível em: https://www.boitempoeditorial.com.br/produto/atlas-do-romance-europeu-1800-1900-73. Acesso em: 17 de setembro de 2023.

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; et al. Training language models to follow instructions with human feedback. In: _____ **Advances in Neural Information Processing Systems** (NeurIPS). 2022. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf. Acesso em: 25 de Novembro de 2023.

PÉREZ, J.; ARENAS, M.; GUTIERREZ, C. Semantics and complexity of SPARQL. **ACM Transactions on Database Systems**, 2009. Disponível em: https://dl.acm.org/doi/abs/10.1145/1567274.1567278. Acesso em: 20 de Setembro de 2023.

PENG, B.; LI, C.; HE, P.; GALLEY, M.; GAO, J. **Instruction tuning with GPT-4**. ArXiv preprint arXiv:2304.03277. 2023. Disponível em: https://arxiv.org/abs/2304.03277. Acesso em: 14 de Dezembro de 2023.

SANTOS, D. **Futuro risonho**: prolegómenos para uma colaboração entre a Linguateca e o NuPILL. 2022. Disponível em: https://www.duo.uio.no/bitstream/handle/10852/98444/1/SantosNuPILL.pdf. Acesso em: 8 de Dezembro de 2023.

SANTOS, Matheus. **Chatterbot baseado em obras de Machado de Assis**: uma plataforma para o estímulo a leitura de literatura clássica. Bauru: UNISAGRADO, 2021. Disponível em: https://secure.usc.br/handle/handle/102. Acesso em: 22 de setembro de 2023.

SCHWARZ, Roberto. **Um mestre na periferia do capitalismo**: Machado de Assis. São Paulo: Duas Cidades, 1990.

SEGARAN, Toby; EVANS, Colin; TAYLOR, Jamie. **Programming the Semantic Web**. Sebastopol: O'Reilly, 2009. pp. 23-26.

SIEMER, S. **Exploring the Apache Jena Framework**. George August University, Göttingen, 2019. Disponível em: http://www.dbis.informatik.uni-goettingen.de/Teaching/Theses/PDF/FPrakt-Siemer-MSc-jun-2019.pdf. Acesso em: 22 de Agosto de 2023.

NOTAS

[i] Os mapas que resultaram deste estudo estão disponíveis no link a seguir: https://huggingface.co/spaces/histlearn/MachadodeAssis.

[ii] SEGARAN, Toby. EVANS, Colin. TAYLOR, Jamie. **Programming the Semantic Web**. Sebastopol: O'Reilly, 2009. pp. 23-26

[iii] HETLAND, Magnus. Python and the Web. In: **Beginning Python From Novice to Professional**. New York: Apress, 2005. pp. 313–339. https://doi.org/10.1007/978-1-4302-0072-7_15. Acesso em 19 de Novembro de 2023.

[iv] OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; et al. Training language models to follow instructions with human feedback. In: **Advances in Neural Information Processing Systems** (NeurIPS). 2022. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf . Acesso em: 25 de Novembro de 2023.

[v] PENG, B.; LI, C.; HE, P.; GALLEY, M.; GAO, J. **Instruction tuning with GPT-4**. ArXiv preprint arXiv:2304.03277. 2023. Disponível em: https://arxiv.org/abs/2304.03277 . Acesso em: 14 de Dezembro de 2023.

[vi] Sobre o trabalho do NuPILL (Núcleo de Pesquisas em Informática, Literatura e Linguística) da Universidade Federal de Santa Catarina, ver: SANTOS, D. **Futuro risonho**: prolegómenos para uma colaboração entre a Linguateca e o NuPILL. 2022. Disponível em: https://www.duo.uio.no/bitstream/handle/10852/98444/1/SantosNuPILL.pdf . Acesso em: 8 de Dezembro de 2023.

[vii] BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. In: **Semantic Web – Interoperability, Usability, Applicability**. 2011. Disponível em: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. Acesso em: 3 de Outubro de 2023.

[viii] DO NASCIMENTO, João Gabriel. O branco imposto e o negro conquistado: Machado de Assis na propaganda da Caixa Econômica Federal. **Revista da Associação Brasileira de Pesquisadores/as Negros/as** (ABPN), v. 8, n. 20, p. 74-85, 2016.

[ix] CHALHOUB, Sidney. **Machado de Assis, historiador**. São Paulo: Companhia das Letras, 2003.

[x] ADIBOZZI, Daniel; et al. Towards a Human-like Open-Domain Chatbot. [S.l.]: **Google Research**, 2020. Disponível em: https://research.google/pubs/towards-a-human-like-open-domain-chatbot/. Acesso em: 22 de setembro de 2023.

[xi] SANTOS, Matheus. **Chatterbot baseado em obras de Machado de Assis**: uma plataforma para o estímulo a leitura de literatura clássica. Bauru: UNISAGRADO, 2021. Disponível em: https://secure.usc.br/handle/handle/102 Acesso em: 22 de setembro de 2023.
[xii] Há ainda um dataset no kaggle com as obras completas do escritor e outras experiências de tecnologia aplicadas ao conjunto: https://www.kaggle.com/datasets/luxedo/machado-de-assis/. Acesso em: 20 de agosto de 2023.
[xiii] SCHWARZ, Roberto. **Um mestre na periferia do capitalismo**: Machado de Assis. São Paulo: Duas Cidades, 1990.
[xiv] MORETTI, Franco. **Atlas do Romance Europeu**, 1800-1900. São Paulo: Boitempo, 2003.
[xv] Ver: https://www.boitempoeditorial.com.br/produto/atlas-do-romance-europeu-1800-1900-73 . Acesso em 17 de setembro de 2023.
[xvi] Para leitura da documentação oficial da OpenAI, ver: https://platform.openai.com/docs/api-reference/introduction
[xvii] CORDEIRO, A. H. D. N. Apache Jena. 2019. Disponível em: https://www.cin.ufpe.br/~in1099/132/Apache%20Jena.pdf . Acesso em: 15 de Outubro de 2023
[xviii] A consulta tem a seguinte sintaxe:

```
PREFIX schema: <http://schema.org/>
SELECT * WHERE {
  ?s schema:name "África" .
  ?s ?p ?o . }
```

[xix] Todos os mapas gerados a partir da obtenção dos dados estão disponíveis em: https://huggingface.co/spaces/histlearn/MachadodeAssis. Apenas a implementação da visão com a API do Google StreetView não está disponível online em razão dos custos associados.

Authors' Contributions

Dilvan de Abreu Moreira: https://orcid.org/0000-0002-4801-2225

- Metodologia: Planejamento da abordagem, métodos e procedimentos da pesquisa.
- Administração de projetos: Coordenação das atividades de pesquisa.
- Supervisão: Orientação e acompanhamento do progresso da pesquisa.

Davi Machado da Rocha: https://orcid.org/0009-0004-3326-6881

- Conceitualização: Desenvolvimento de ideias e estrutura da pesquisa.
- Investigação: Realização dos experimentos e coleta de dados.
- Curadoria de dados: Organização, manutenção e controle dos dados de pesquisa.
- Escrita – rascunho original: Redação inicial do manuscrito.
- Escrita – revisão e edição: Revisão e aperfeiçoamento do manuscrito.

Conflicts of Interest

Os autores declaram não haver conflitos de interesse em relação à publicação deste artigo.

The authors declare no conflicts of interest regarding the publication of this paper.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.