Geocodificação de Endereços Brasileiros com Abordagens Lexicais, Indexadas e Semânticas

Sabrina Araújo da Silva Samantha Miyahira
Orientadora: Profa. Dra. Kelly Braghetto

Bacharelado em Ciência da Computação
Instituto de Matemática, Estatística e Ciência da Computação

Introdução

A ausência de padronização na forma como endereços são registrados no Brasil dificulta a integração de bases de dados e compromete análises territoriais essenciais para políticas públicas. Plataformas como a **CulturaEduca**, que utilizam informações espaciais para apoiar decisões locais, dependem da **correta localização geográfica de endereços** provenientes de fontes diversas e frequentemente inconsistentes.

Para enfrentar esse desafio, este trabalho desenvolveu uma ferramenta de **software livre capaz de padronizar e geolocalizar endereços** a partir de entradas textuais incompletas ou com erros, utilizando como referência o **Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE/IBGE)**. Nossa solução combina comparação léxica (*RapidFuzz*), busca indexada (*Elasticsearch*) e modelos de linguagem, além da aplicação de critérios de desempate de pontos para um mesmo endereço, com o objetivo de melhorar a acurácia territorial na geolocalização.

A seguir, apresentamos as abordagens utilizadas e seus resultados quando aplicados a endereços no **Município de São Paulo (SP)**, analisando sua eficiência e a **acurácia** em comparação com as do *GeoCodeBR* (ferramenta desenvolvida pelo IPEA).

CulturaEduca

O **CulturaEduca** é uma plataforma pública que reúne dados de escolas, unidades de saúde e equipamentos culturais em mapas interativos. Ela apoia a gestão municipal ao oferecer informações territoriais que ajudam no planejamento e na tomada de decisões.
Para isso, depende de uma **geolocalização precisa**, pois muitos indicadores e visualizações são construídos a partir dos endereços das bases governamentais.

Visão Geral da Solução

O processo de correspondência foi estruturado em um **módulo** que opera em quatro etapas principais:

1. **Normalização dos endereços** — limpeza, expansão de abreviações, padronização textual, tratamento de números e separação em componentes.
2. **Correspondência** — busca de candidatos no CNEFE e cálculo de similaridade entre a entrada e cada candidato.
3. **Desempate** — endereços que tiveram mais de um ponto resultante da etapa 2 passam pelos critérios de desempate para obter apenas uma coordenada para cada endereço.
4. **Avaliação** — comparação espacial das coordenadas, cálculo do erro de distância e verificação territorial em nível de setor censitário.

Módulo de Correspondência

O módulo recebe um **endereço normalizado**, busca candidatos no CNEFE usando uma das abordagens (léxica, indexada ou semântica) e calcula uma **pontuação de similaridade** entre a entrada e cada candidato.

Essa similaridade combina informações de **logradouro**, **número** e **bairro** em um **score final**, e o módulo seleciona o(s) candidato(s) de maior pontuação como endereço correspondente. As três abordagens reutilizam essa mesma estrutura, mas calculam a similaridade de formas distintas.

1. Abordagem Léxica (RapidFuzz)

A abordagem léxica calcula a similaridade entre textos usando **distância de edição** e **conjuntos de tokens**. Ela captura variações leves de escrita, como abreviações, acentuação e mudanças na ordem dos termos.

Funciona bem quando a forma textual é parecida, mas perde desempenho quando há diferenças estruturais mais profundas.

2. Abordagem Indexada (Elasticsearch)

A abordagem indexada utiliza um **índice invertido** e a função de ranqueamento **BM25** para recuperar rapidamente os endereços mais prováveis. Ela permite *fuzzy search*, tolerando grafias próximas.

É eficaz em bases extensas e em variações moderadas de escrita.

3. Abordagem Semântica (LLM)

A abordagem semântica representa cada endereço por meio de **embeddings vetoriais**. A similaridade é medida pela proximidade entre os vetores, permitindo reconhecer expressões que têm o mesmo significado mesmo quando escritas de formas muito diferentes.

É a abordagem mais robusta para variações profundas de escrita, mas seu custo computacional inviabilizou a execução em bases extensas, motivo pelo qual não aparece nos resultados a seguir.

Resultados

Os testes foram realizados utilizando a base de endereços de estabelecimentos de educação básica do **CulturaEduca**, buscando localizar esses registros na base do **CNEFE do Estado de São Paulo**.

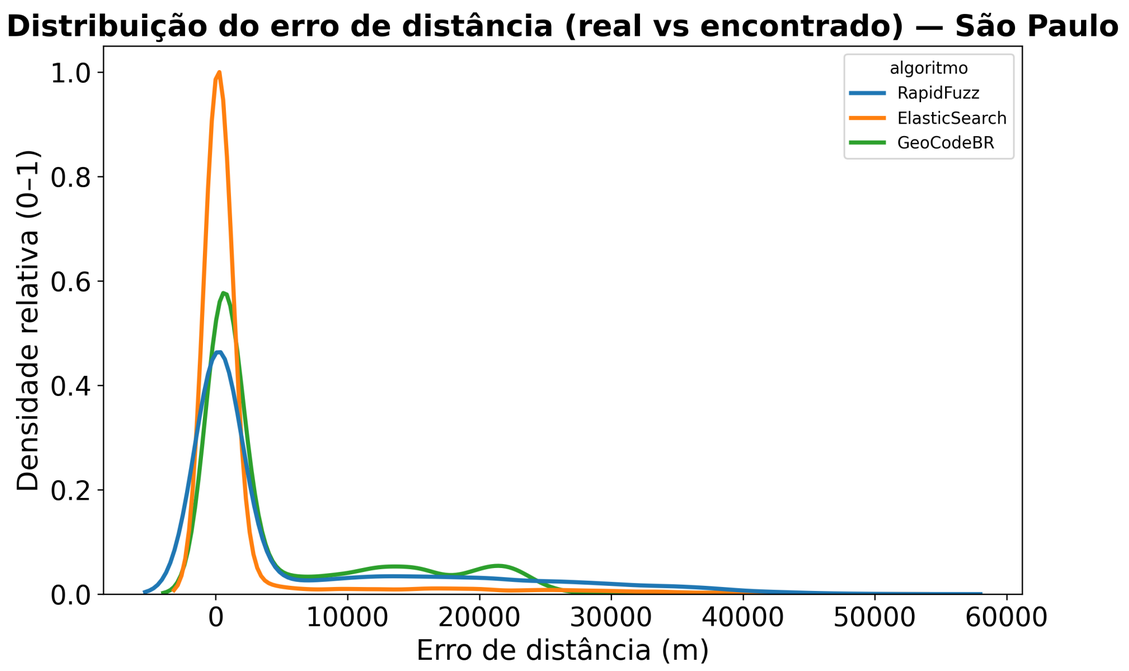

Figura 1 — Município de São Paulo (SP): distribuição do erro de distância (KDE).

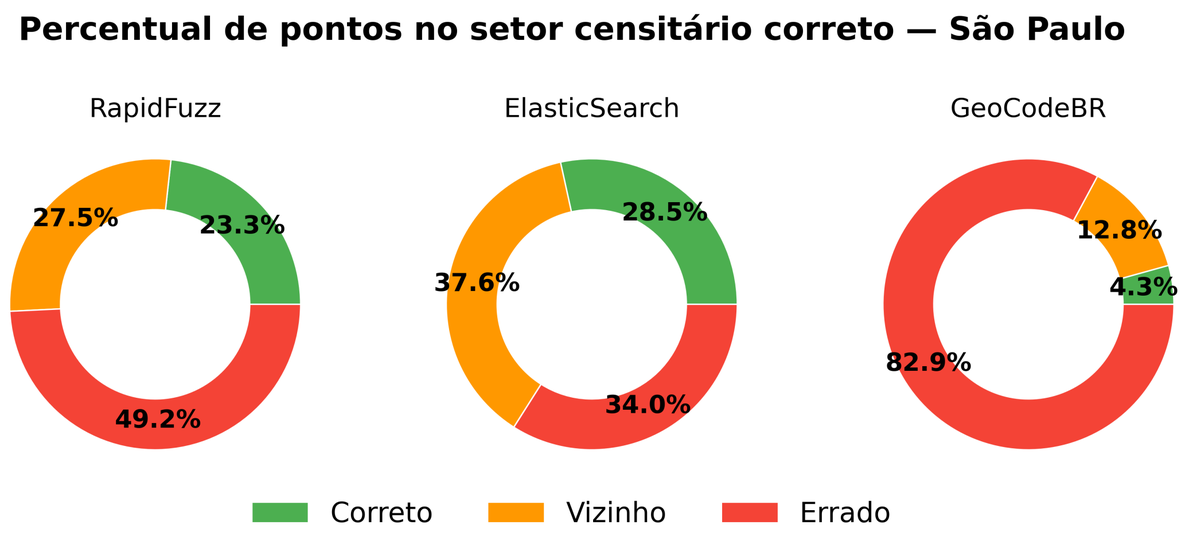

Figura 2 — Município de São Paulo (SP): proporção entre resultados corretos, vizinhos e errados.

Conclusão

A API desenvolvida apresentou desempenho superior ao *GeoCodeBR*, tanto em relação ao **desvio dos pontos geocodificados** quanto à identificação correta dos **setores censitários**. A abordagem indexada demonstrou ser mais eficaz para a maioria dos registros, tanto em tempo de execução quanto em acurácias nos resultados.

Os resultados para o Município de São Paulo evidenciam que essas técnicas, aliadas à normalização dos endereços, reduzem significativamente erros de geocodificação e melhoram a identificação do setor censitário correto.

Assim, a solução proposta avança na direção de um **geocodificador livre**, mais preciso e adequado às demandas do **CulturaEduca**.

Referências

[1] "Portal culturaeduca." `https://culturaeduca.cc/sobre/`, s.d.

[2] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

sabrinaaraujo@usp.br Trabalho de Conclusão de Curso samanthamiyahira@usp.br