REVISTA DE TECNOLOGIA APLICADA (RTA)
v.10, n.2, mai-ago 2021, p. 51-64.
ISSN: 2237-3713

http://dx.doi.org/10.48005/2237-3713rta2021v10n2p5164

# Criação de uma base de valores imobiliários geo-referenciados a partir da extração de dados da internet*

## *Creation of a geo-referenced real estate values database from automatic internet data extraction*

**Matheus Aparecido da Silva Roberto**
Universidade Tecnológica Federal do Paraná (UTFPR)
matheroberto@gmail.com

**André Koscianski**
Universidade Tecnológica Federal do Paraná (UTFPR)
matheroberto@gmail.com

## RESUMO

Estão disponíveis na Internet muitas informações de interesse financeiro, porém na maioria dos casos são apresentadas de maneira não adequada para processamento digital. Um caso particular são valores de mercado imobiliário, essenciais em várias tarefas de administração pública e do setor privado. O trabalho teve o objetivo de apresentar técnicas para criar uma base de valores imobiliários geo-referenciados e uma visão geral de uma solução computacional implementada, que pode ser adaptada a situações similares. Como resultados gerou-se uma base de dados imobiliários para a cidade de Ponta Grossa no Paraná, com inicialmente 20 mil registros, com uma taxa de aproveitamento em torno de 90%. Em conclusão, o baixo custo de desenvolvimento e a efetividade do software confirmaram a utilidade desse tipo de solução.

**Palavras-chave:** Bases de Dados. Mercado Imobiliário. Administração. Automação de Processos. Mineração na Web. Geo-referenciamento.

## ABSTRACT

The internet contains a wealth of financial information, but in most cases it is not formatted in a proper way for digital processing. One example is real estate market values, which are essential in various tasks of public administration and the private sector. The objective of this work is to present techniques to create a geo-referenced database of real-estate values. The text gives an overview of a computational solution, which can be adapted to similar situations. The software was used to generate a real-estate database for the city of Ponta Grossa in Paraná, with initially 20 thousand records and a rate of approximately 90% of data extraction. In conclusion, the low development costs and the effectiveness of the software confirmed the usefulness of the proposed approach.

**Keywords:** Databases. Real Estate Market. Process Administration. Process Automation. Web Mining. Georreferencing.

*Received 28 February 2021; accepted in 28 February 2021; published online 03 November 2021.

## 1. INTRODUÇÃO

Uma das bases de dados mais antigas que se conhece são registros contábeis sumerianos, datados de aproximadamente 3000 a.C. (HIGOUNET, 2003). A habilidade de guardar informação revolucionou a história humana, e hoje diversas atividades dependem desse recurso que existe em diferentes formatos. Alguns exemplos são contratos comerciais, divididos em cláusulas, leis e regulamentos que citam outros regulamentos e que assim se aproximam da ideia de hipertexto; e artigos científicos, que incluem gráficos e tabelas.

A Computação trouxe a passagem de registros físicos para arquivos digitais e facilidade para copiar, modificar e transmitir informações. O passo seguinte veio com a Internet e o acúmulo de um volume extraordinário de informação, nas mais diversas áreas, como economia, ciência, saúde, administração pública. Surpreendentemente, porém, a maior parte das fontes de dados digitais não estão prontas ou plenamente adequadas para processamento por computadores (GÓMEZ-PÉREZ; CORCHO, 2002). A história recente de vitrines de lojas virtuais mostra isso. No início da criação de comércio desse tipo usuários precisavam percorrer a rede para comparar produtos, valores e custos de entrega: evidentemente essa tarefa repetitiva pode ser automatizada em um algoritmo. Logo surgiram ferramentas de busca (tais como Bondfaro e Buscapé), empregando algoritmos denominados robôs para vasculhar a rede (GREGG et al, 2006). Infelizmente, adaptar essa tecnologia para outros contextos não é uma tarefa trivial.

O problema genérico de extrair dados a partir de documentos não-estruturados, mesmo que digitais, é bastante complexo e requer recursos sofisticados, como algoritmos para tratar linguagem natural (AGGARWAL et al, 2012). No caso de contextos restritos, são desenvolvidas soluções específicas; tem-se assim o caso exemplificado de buscadores de produtos em lojas (ARRUDA; ROSSI; PENIDO, 2011), tratamento de informações jurídicas (CUNHA; SILVA; TALON, 2013), ou ainda monitoramento de saúde (VELASCO et al 2014). Atualmente não se dispõe ainda de métodos universais para vasculhar a Internet e extrair dados (FERRARA et al, 2014; PARVEZ et al, 2018), mas existem ferramentas e técnicas que resolvem partes do problema e que podem ser combinadas para tratar um determinado domínio de informação.

Este trabalho discute uma aplicação de extração de dados da Internet que inclui reconhecer dados parciais ou com erros, e um problema adicional de georreferenciamento. O objetivo consiste em criar uma base de dados que retrata a evolução temporal de valores de imóveis. Esse tipo de informação é importante em tarefas como administração pública (na gestão de planos diretores e determinação de valores de IPTU), em investimento privado (imobiliárias e construtoras), estudos de caráter demográfico (caracterização de regiões de uma cidade) e econômico (projeções financeiras). No presente caso a base será usada como entrada para estudos de modelagem e simulação de crescimento urbano.

A solução apresentada no artigo implementa uma sequência de passos que inicia com a extração de dados brutos a partir de endereços selecionados, o que dispensa varredura utilizando robôs. Em virtude de muitos anúncios apresentarem logradouros com grafia errada, a simples filtragem de endereços válidos poderia levar a uma perda importante de dados por descarte. Para contornar isso são aplicados métodos de casamento aproximado de *strings* (cadeias de caracteres), procurando minimizar a rejeição de endereços. A lista resultante passa então por georreferenciamento para obter coordenadas em um formato padronizado; este estudo faz uma breve comparação entre os principais provedores desse tipo de serviço. Por fim, os dados são arquivados em uma base estruturada visando processamento por Sistemas de Informação Geográfica (SIG) e softwares de simulação urbana.

O artigo descreve as técnicas mais relevantes nesse contexto, mostrando as escolhas realizadas e como as ferramentas são integradas para implementação de uma solução.

Apresenta-se um comparativo com um trabalho no mesmo setor (dados imobiliários) e uma análise geral de custo e ganho econômico obtido com a automação do processo.

## 2. FUNDAMENTAÇÃO

### 2.1. Contexto do Trabalho

Pesquisas envolvendo Sistemas de Informação Geográfica e estudos de projeção econômica e demográfica requerem volumes significativos de informação. Dados atualizados e com amostragem adequada são difíceis de obter, especialmente informações de natureza financeira (SILVA, 1998; WADDELL; 2000, ROTH; 2019). Essa limitação é por vezes contornada utilizando aproximações e extrapolações para cobrir lacunas, mas isso aumenta a margem de incerteza e traz mais erros de projeção (CELLMER; SZCZEPANKOWSKA, 2014). Por outro lado, a Internet é uma fonte de informações importante em setores que envolvem finanças e economia (BARTELS; BREITNER, 2004; BLAZQUEZ et al, 2018). Anúncios em classificados eletrônicos apresentam uma série de vantagens nesse contexto. Primeiramente, são voltados para o público em geral e por esse motivo não há questões de restrição de consulta. Em seguida, exceto pela necessidade de conexão Internet, o custo de acesso é nulo. Finalmente, as informações são editadas em permanência, reduzindo assim a possibilidade de obter valores já desatualizados.

A publicidade de imobiliárias que se fazia em jornais impressos passou inteiramente para a Internet. A quantidade de informação aumentou, mas por outro lado permaneceu misturada e desorganizada. Com frequência anúncios são apresentados em textos fora de sequência, escritos pensando não na troca de dados entre computadores, mas sim em pessoas que buscam saber sobre imóveis. Não existe um padrão estabelecido entre sites e, além disso, a informação geográfica se limita a um endereço, muitas vezes abreviado ou grafado erroneamente. Isto significa que a criação de uma base de dados de anúncios envolve a identificação e extração das informações, a filtragem de erros, a tarefa de converter logradouros em coordenadas geográficas, e por fim o arquivamento de dados em um formato estruturado previamente determinado.

Esses problemas serão abordados nas próximas seções.

### 2.2. Extração de Dados da Internet: visão geral

A extração de dados a partir da Internet, conhecida comumente como Web Mining, apresenta desafios específicos além dos existentes no processamento tradicional de textos (BHARANIPRIYA; PRASAD, 2011; FERRARA et al, 2014; PARVEZ et al, 2018). O exemplo mais básico disso é a multiplicidade de padrões de comunicação e representação de dados usados na rede (SYME et al, 2012).

O princípio geral para extrair informações da internet é simular a ação humana, seja via protocolo HTTP (Hypertext Transfer Protocol) ou acionamento automatizado de navegadores (NEIL, 2016), implementando o chamado web crawling (OLSTON et al; 2010). O conjunto de técnicas voltadas para o problema é conhecido como web scraping ou web harvesting (SCHRENK, 2012), muito utilizadas para marketing digital e mineração de dados (CHEN et al. 2015; NEIL, 2016). O princípio geral é realizar acessos automáticos a um ou vários sites para carregar documentos HTML (Hypertext Markup Language) e extrair deles as informações de interesse (GLEZ-PEÑA et al, 2013). Entre as técnicas mais utilizadas pode-se citar as seguintes (PARVEZ et al 2018):

*Análise de HTML*, ou *HTML Parsing*. Essa técnica trabalha com a estrutura de uma página HTML para localizar dentro dela as informações necessárias (GLEZ-PEÑA et al; 2013). O HTML recebido do servidor contem palavras chaves que podem auxiliar na extração de dados. Outra maneira de abordar o problema consiste em criar uma árvore sintática (KADAM; PAKLE, 2014). A técnica de *HTML parsing*, além de ser rápida, traz facilidade a desenvolvedores para encontrar links e imagens;

*CSS Selector*: diferente do método anterior, este torna possível a seleção de elementos HTML sem a preocupação com estrutura do documento. Neste caso a identificação de informações é possível se houverem atributos como "class", "type" ou "id". A técnica é mais flexível, no sentido de que torna mais fácil realizar futuras alterações de seleção (FAROOQ; HUSAIN; SUAIB, 2018; YIN; HE; LIU, 2018);

*Expressões Regulares*: esta solução é especialmente indicada se for possível encontrar trechos de texto que seguem um padrão. Exemplos disso são datas, números de CPF e de telefone. Embora de uso simples, a técnica é capaz de tratar padrões complexos e é efetiva para identificar trechos de texto dentro de grandes volumes de HTML (BRIN, 1998).

## 2.3. Georreferenciamento

A localização espacial de informações é um item essencial em Sistemas de Informação Geográfica (SIG). As informações de posicionamento são utilizadas em situações como estudo de epidemias (PALANIYANDI, 2014), gerenciamento de reservas naturais (CORRÊA et al, 1996), e planejamento e investimento urbano (PÉRICO; CEMIN, 2006). Esse dado é representado em um sistema de coordenadas com latitude e longitude que deve se referir a uma determinada geometria de referência (CROUSE, 2016). O padrão mais comum é conhecido como WGS84 e foi adotado neste trabalho (CASTILHO; FRANZOSO, 2015).

A tarefa de georreferenciar endereços, que no passado manual, hoje conta com serviços automatizados (FLORCZYK et al, 2010; OZIMEK; MILES, 2011; SOLINA; RAVNIK, 2010; WILSON; SWIFT; GOLDBERG, 2008). Os provedores disponíveis apresentam diferenças principalmente de precisão e custo. Para transformar endereços em localização espacial, este trabalho testou dois principais serviços de georreferenciamento: o serviço mantido pela empresa Google com o Geocoding API; e a ferramenta Nominatin, mantida por voluntários e pela fundação OpenStreetMap (OSMF).

O serviço Geocoding da Google mantem uma base de dados geográficos atualizada e de abrangência global. Oferece mapas e serviços cartográficos para vários sites ou empresas (CIEPŁUCH et al, 2010; SILVEIRA; OLIVEIRA; JUNGER, 2017; SZTUTMAN, 2014). A ferramenta Nominatim, por sua vez, tem a base de dados mantida por voluntários, o que pode torná-la um pouco defasada em alguns locais (CIEPŁUCH et al, 2010). Entretanto, ela se destaca pelo uso gratuito.

# 3. METODOLOGIA

A solução foi escrita na linguagem Python, dada a simplicidade de sintaxe e a grande disponibilidade de recursos para desenvolver aplicações voltadas à Internet (DOLGERT; GIBBONS; KUZNETSOV, 2008). O software desenvolvido neste trabalho é classificado como limitado por E/S e não por CPU (GAGNE; GALVIN; SILBERSCHATZ, 2015), e assim o uso de uma linguagem interpretada não foi negativo para desempenho.

A Figura 1 apresenta a sequência de etapas de processamento, que corresponde aproximadamente à estrutura do software desenvolvido.

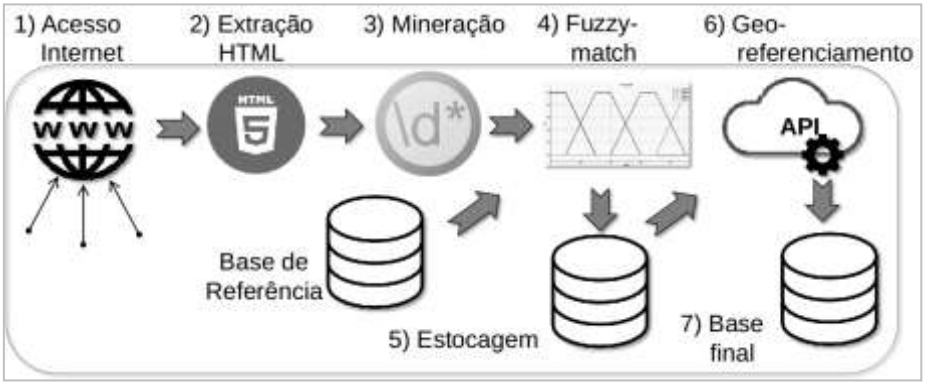


**Figura 1 – Estrutura geral da solução.**

Fonte: autoria própria.

A sequência de processamento pode ser resumida nos seguintes passos:

1. acesso à Internet usando processamento paralelo para ler dados de sites imobiliários;
2. aplicação de HTML *parsing* com a biblioteca Beautiful Soup;
3. extração de texto a partir de HTML e mineração com expressões regulares;
4. verificação de endereços com biblioteca fuzzywuzzy e lista de logradouros;
5. arquivamento de base dados inicial no formato JSON (JavaScript Object Notation);
6. chamada à API (Application Programming Interface) Nominatin;
7. resultado final em um arquivo JSON com todos anúncios georreferenciados.

As seções a seguir apresentam mais detalhes sobre as etapas de processamento.

### 3.1. Aquisição dos dados e extração de endereços

A leitura de dados a partir de um site imobiliário foi implementada com a biblioteca Python requests-html. Ela contem a lógica para criar uma conexão e obter o HTML da página em que se encontram agrupados os anúncios. Há sites organizados como listas de dados e outros como uma série de links, e então é necessário extrair as URL (Uniform Resource Locator) para concluir a primeira etapa.

Nesse momento é preciso tratar links quebrados e erros de sintaxe. Além disso páginas podem conter laços (links apontando para páginas já visitadas), o que é contornado implementando um histórico de visita. Parte dos links contem informações sem importância, tais como anúncios diversos ou links para WhatsApp e Facebook. Para evitar a navegação para esses enlaces e a consequente perda de tempo de E/S, foram incluídas no código informações de filtragem usando expressões regulares.

Para otimizar a ferramenta no processo de extração de anúncios, o processo foi multiplexado em threads. Cada thread recebe um pedaço da lista de enlaces, encontra anúncios e extrai os dados necessários do HTML da página. Foram instanciadas quatros threads, obedecendo à limitação de requisições simultâneas impostas por alguns sites usados no estudo. Como as requisições Internet tem um atraso acentuado comparado com a velocidade da CPU, e o número de pedidos simultâneos é pequeno (apenas quatro), obtem-se um ganho com o uso de multi-thread neste caso.

No passo seguinte, a técnica CSS Selector foi usada para identificar *tags* HTML por atributos. Esse processo foi implementado empregando a biblioteca Beautiful Soup e é uma maneira eficiente para obter os dados desejados se os mesmos estão destacados em documentos HTML (RICHARDSON, 2007). Neste trabalho os seletores foram ajustados conforme os sites imobiliários selecionados, permitindo assim filtrar textos de interesse para o projeto. Foram também utilizadas expressões regulares para quebrar endereços em partes:

logradouro, número, bairro e vila. Esses endereços apresentam com frequência erros de escrita ou uso de abreviaturas, requerendo assim mais uma etapa de tratamento.

Para identificar um nome de rua modificado por uma abreviatura ou erro de digitação, foi utilizado um algoritmo para verificar casamento aproximado de cadeias de caracteres. O algoritmo bitap é capaz de reconhecer uma string em que houveram inserções e remoções de caracteres (MYERS, 1999). Basicamente, o código procura trechos de uma cadeia, em sequência, que correspondam a trechos idênticos em um texto de referência fornecido. Os textos de referência neste trabalho consistem de uma lista de nomes de logradouros obtidos do projeto OpenStreetMaps (HAKLAY, 2008). Um arquivo ShapeFile contendo o mapa de Ponta Grossa foi lido a partir do site OpenStreetMaps e a lista de logradouros foi extraída por meio do software Quantum-GIS. A implementação para este projeto usou o algoritmo bitap disponível em Python na biblioteca fuzzywuzzy (COHEN, 2011). Essa biblioteca retorna uma pontuação de semelhança, o que implicou realizar uma iteração em lista para selecionar o resultado com valor mais alto.

Com as informações de um anúncio finalmente separadas, é preparado um objeto JSON para armazenamento. Esse formato foi escolhido pela interoperabilidade, simplicidade de sintaxe e indicações de desempenho comparado com XML (NURSEITOV et al, 2009). Os anúncios de cada imobiliária contem uma referência, ou código interno de identificação. Essa referência foi utilizada para identificar no arquivo de saída instâncias do mesmo anúncio, de maneira que possam ser registradas variações de preço mês a mês.

## 3.2. Georreferenciamento

A solução adotou o serviço de georreferenciamento Nominatin. Essa escolha se justificou inicialmente pela ferramenta ser gratuita, e pela fácil integração no código (CIEPŁUCH et al, 2010). Testes de desempenho confirmaram que o tempo de resposta era adequado para o projeto. Para implementar as requisições e tratamentos do serviço de georreferenciação foi utilizada a biblioteca geopy (GEOPY, 2018). O serviço Nominatin não permite mais de uma requisição simultânea; isto foi resolvido estabelecendo um intervalo de tempo entre requisições, por meio do módulo RateLimiter da biblioteca geopy.

Nesse ponto termina o processamento inicial de um anúncio. Em execuções seguintes a ferramenta deve registrar variações de preço de um mesmo imóvel na base de dados, e não haverá necessidade de fazer nova consulta ao Nominatim. Entretanto, esse objetivo esbarra numa dificuldade: mudanças no anúncio podem incluir preço, mas também edição do endereço. Outra situação que pode levar a isso é o proprietário do imóvel contratar os serviços de outra imobiliária. Para garantir o melhor possível que um imóvel seja reconhecido e não haja duplicidade, são feitas tentativas em sequência para georreferencialmento. O código emprega combinações de informações (por exemplo, incluindo ou não nome de bairro) procurando uma coincidência. A cada tentativa, busca-se obter a localização geográfica do endereço em latitude e longitude.

A Figura 3 sumariza a divisão da ferramenta em módulos.

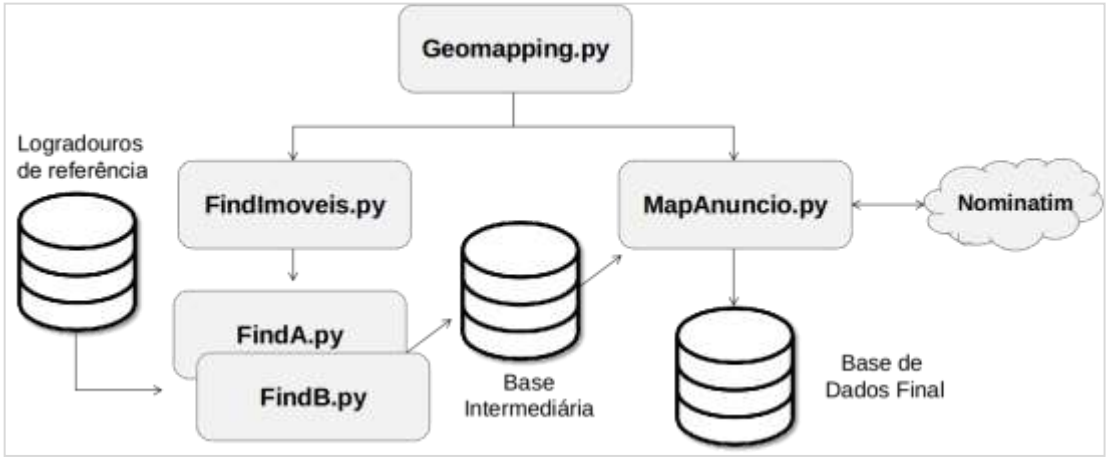


**Figura 3 – Divisão em módulos da ferramenta.**

Fonte: autoria própria.

Para finalização do processo tem-se dois arquivos em formato JSON, um com todos os anúncios extraídos (georreferenciados ou não) e um arquivo com todos os anúncios georreferenciados e prontos para uso.

## 4. RESULTADOS

Para os objetivos do trabalho optou-se por um aplicativo que pudesse funcionar sem supervisão humana. A ferramenta desenvolvida conseguiu tratar cerca de 139 anúncios por minuto; em uma comparação com uma solução semi-automatizada (Batista e Xavier 2018), obteve-se uma aceleração de 30 vezes, sem incluir a redução de taxa de requisições imposta pelo serviço Nominatim. É importante destacar que o objetivo dessa comparação é ilustrar o ganho de produtividade obtido pelo investimento na automação do processo. A realização manual da triagem e arquivamento dos dados tem com um custo cumulativo mais alto do que o desenvolvimento de uma ferramenta dedicada como no presente caso.

Em termos financeiros pode-se considerar os seguintes dados:

- custo por hora para realização manual da tarefa = C;
- custo por hora de um programador = $\alpha \cdot C$;
- anúncios em processo semi-automático (Batista e Xavier, 2018) = 278 por hora;
- aceleração do sistema informatizado = $\beta$ (30 vezes no presente caso);
- tempo de desenvolvimento do sistema = 120 horas (distribuídas em 3 meses);

A comparação entre os dois casos é dada pelo Quadro 1.

**Quadro 1 - Análise comparativa de custo de operação.**

| Item | Semi-Automatizado | Sistema Proposto |
|---|---|---|
| Custo fixo | 0 | $120 \cdot \alpha \cdot C$ |
| Custo variável | $C \cdot t$ | 0 |
| Anúncios por hora | 278 | $\beta \cdot 278$ |
| Custo por anúncio. | $(C \cdot t) / 278 \cdot t$ | $120 \cdot \alpha \cdot C / (\beta \cdot 278 \cdot (t - 120)$ |
| Ponto de Equilíbrio | $t = 120 \cdot (1 + \alpha / \beta)$ | |

Fonte: Autoria Própria.

Supondo que o preço por hora do trabalho de um programador seja $\alpha$=10 vezes o de um operador realizando a tarefa de extração de dados de anúncios (um valor exagerado considerando os valores médios de salários do mercado), para uma aceleração de $\beta$=30 vezes, o sistema informatizado se paga em menos de 40 horas de uso. Considerando que a referência usada como base de comparação não incluia a etapa de georreferenciamento, a economia é ainda maior.

Como o objetivo da ferramenta era obter uma amostra de dados e operar sem nenhuma intervenção humana, anúncios que apresentam problemas foram descartados. Em outras situações de mineração de dados em que isso não seja desejado, as informações podem ser reservadas à parte para uma etapa separada de processamento.

Utilizando apenas três sites, com extração de anúncios a cada 40 dias num intervalo de 9 meses foi possível criar uma base com 9500 anúncios, dos quais 8900 georreferenciados. Isso representa uma taxa de aproveitamento de 93% dos dados de entrada. Todos os arquivos de saída foram gerados em formato JSON, o que possibilita realizar a importação dos dados para um banco de dados não relacional e torna fácil a adaptação para outros formatos se necessário.

## 5. CONCLUSÃO

Bases de dados são um elemento essencial em praticamente todos os setores de atividade. Em muitos casos, uma fonte importante são repositórios não estruturados na Internet, que embora contenham grandes volumes de informação, apresentam uma série de obstáculos para o processamento digital. Isso inclui repetição de informações, descrições inexatas e parciais, tais como abreviaturas, ou simples erros de registro e inconsistências.

Este trabalho visou extrair informações não estruturadas armazenadas em sites imobiliários, visando criar uma base de dados padronizada e flexível, para uso em estudos de simulação urbana. Mesmo utilizando apenas três sites de imobiliárias de uma cidade de médio porte, obteve-se uma amostra significativa de anúncios, e uma alta eficiência no reconhecimento de informações apesar de variações na grafia e abreviaturas de endereços.

Existem possibilidades para acelerar operação da solução proposta, sem grandes mudanças na arquitetura; isso inclui: usar instâncias em paralelo para leitura de diferentes sites de anúncios; usar diferentes endereços IPs para requisições ao Nominatim, ou contratação de um serviço de georreferenciamento. O tempo de desenvolvimento citado no texto (120h) é uma aproximação. Se for descartado o prazo envolvido com estudos iniciais para seleção de métodos e preparação de relatório técnico, esse período é ainda menor. A estrutura geral apresentada neste artigo pode ser adaptada para outras situações e a utilização de bibliotecas padronizadas torna fácil reproduzir a experiência. Uma extensão interessante seria tornar a ferramenta capaz de se adaptar automaticamente a mudanças de estruturas e layouts utilizados por um site. Técnicas de Inteligência Artificial para tratamento de textos genéricos são uma possibilidade. Outro caminho seria rever as técnicas de *web scraping*, diminuindo a atenção na estrutura de documentos HTML e dando mais prioridade a métodos como regex.

O código do projeto é disponibilizado como OpenSource no repositório https://github.com/MatheusRoberto/GeoMappingPG.

## 6. REFERÊNCIAS

AGGARWAL, C. C., ZHAI, C. X., **Mining text data**, Springer Science & Business Media, 2012.

AIJUN, X.; LICHUAN, G. **Encoding & decoding of Chinese address and development of algorithms for intelligent address search**. 2010 International Conference on Computer Application and System Modeling. **Anais**...IEEE, out. 2010. Disponível em: <http://ieeexplore.ieee.org/document/5620171/>. Acesso em: 03 fev 2021.

ARRUDA, C.; ROSSI, A.; PENIDO, E., Buscapé: Do empreendendorismo à inovação aberta, **Casos FDC**, 2011.

BARTELS, P., BREITNER, M. H., Finance applications with the web mining software agent PISA. In **Impulse aus der Wirtschaftsinformatik** , p. 135-149, 2004.

BATISTA, B. **Aprenda por definitivo a usar CSS Selector(Adeus Xpath).** Disponível em: <https://medium.com/automação-com-batista/aprenda-por-definitivo-a-usar-css-selector-adeus-xpath-1f3956763c2>. Acesso em: 03 fev 2021.

BATISTA, J. DA S.; XAVIER, E. S. **Criação de um banco de dados não relacional a partir de informação extraída de textos**Ponta GrossaUniversidade Tecnológica Federal do Paraná, , 29 maio 2018. Disponível em: <http://repositorio.roca.utfpr.edu.br/jspui/handle/1/9729>. Acesso em: 12 fev 2021

BHARANIPRIYA, V.; PRASAD, V. K. **WEB CONTENT MINING TOOLS: A COMPARATIVE STUDYInternational Journal of Information Technology**. [s.l: s.n.]. Disponível em: <http://csjournals.com/IJITKM/PDF 4-1/43.V. Bharanipriya1 %26 V. Kamakshi Prasad2.pdf>. Acesso em: 03 fev 2021.

BLAZQUEZ, D., DOMENECH, J., Big Data sources and methods for social and economic analyses, **Technological Forecasting and Social Change**, v. 130, p. 99-113, 2018.

BRIN, S., Extracting patterns and relations from the world wide web. In **International Workshop on The World Wide Web and Databases**, Springer, Berlin, Heidelberg, p. 172-183, 1998.

BUSCAPÉ COMPANY INFORMAÇÃO E TECNOLOGIA LTDA. **Buscapé - Conheça o Buscapé**. Disponível em: <https://www.buscape.com.br/conheca-o-buscape>. Acesso em: 03 mar 2020.

CASCÓN-KATCHADOURIAN, J.; RUIZ-RODRÍGUEZ, A.-Á.; ALBERICH-PASCUAL, J. Uses and applications of georeferencing and geolocation in old cartographic and photographic document management. **El Profesional de la Información**, v. 27, n. 1, p. 202, 2018.

CASTILHO, S. D., FRANZOSO, L. F. F., Análise de Ferramentas e Processos Utilizados em Levantamento Georreferenciado, **RETEC-Revista de Tecnologias**, v. 7, n. 1, 2015

CELLMER, R., SZCZEPANKOWSKA, K., Simulation modeling in a real estate market, In **Proceedings of the International Conference on Environmental Engineering**. ICEE, v. 9,

Vilnius Gediminas Technical University, Department of Construction Economics & Property, 2014.

CHEN, Z.-H. et al. **Big data: Open data and realty website analysis**. 2015 8th International Conference on Ubi-Media Computing (UMEDIA). **Anais**...IEEE, ago. 2015Disponível em: <http://ieeexplore.ieee.org/document/7297433/>. Acesso em: 24 jun. 2020

CIDADE DE SÃO PAULO. **CopiCola**. Disponível em: <https://copicola.prefeitura.sp.gov.br/#sobre>. Acesso em: 28 nov. 2020.

CIEPŁUCH, B. et al. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. **Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010**, p. 337–341, 20 jul. 2010.

COELHO, A. L. N. Sistema De Informações Geográficas (Sig) Como Suporte Na Elaboração De Planos Diretores Municipais, **Caminhos De Geografia**, v. 10, n. 30, p. 93–110, 2009.

COHEN, A. **FuzzyWuzzy: Fuzzy String Matching in Python - ChairNerd**. Disponível em: <https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/>. Acesso em: 2 jul. 2020.

COMPUTERWORLD. **Prefeitura de São Bernardo reduz burocracia com transformação digital | Computerworld**. Disponível em: <https://computerworld.com.br/2019/05/13/prefeitura-de-sao-bernardo-reduz-burocracia-com-transformacao-digital/>. Acesso em: 28 nov. 2020.

CORRÊA, T., COSTA, C., SOUZA, M. G., & BRITES, R. S., Delimitação e caracterização de áreas de preservação permanente por meio de um sistema de informações geográficas (SIG). **Revista Árvore**, v. 20, n. 1, p. 129-135, 1996.

CROUSE, D. F. **An Overview of Major Terrestrial, Celestial, and Temporal Coordinate Systems for Target Tracking**. No. NRL/FR/5344-16-10. Naval Research Lab, Washington DC Surveillance Technology Branch, 2016.

CUNHA, J. F. T; SILVA, W. F.; TALON, A. F., Aplicação da Técnica de Mineração de Dados na Análise de Processos Jurídicos do Estado de São Paulo. **Caderno de Estudos Tecnológicos**, v. 1, n. 1, 2013.

DESAI, K. et al. Web Crawler : Review of Different Types of Web Crawler, Its Issues, Applications and Research Opportunities. **International Journal of Advanced Research in Computer Science**, v. 8, n. 3, p. 1199–1202, 2017.

DOLGERT, A., GIBBONS, L., KUZNETSOV, V., Rapid web development using AJAX and Python. **Journal of Physics: Conference Series**, v.119, n. 4, IOP Publishing, 2008.

E-GESTÃO PÚBLICA. **O que é preciso para digitalizar a gestão de prefeitura?** Disponível em: <https://www.e-gestaopublica.com.br/gestao-de-prefeitura/>. Acesso em: 28 nov. 2020.

FAROOQ, B.; HUSAIN, M. S.; SUAIB, M. **CRAWLING OF JAPANESE REAL-ESTATE WEBSITES USING SCRAPY**. International Journal of Advanced Research in Computer Science. **Anais**...2018Disponível em: <www.ijarcs.info>. Acesso em: 23 set. 2020.

FERRARA, E. et al. Web data extraction, applications and techniques: A survey. **Knowledge-Based Systems**, v. 70, p. 301–323, 2014.

FERREIRA, R. V.; RAFFO, J. DA G. O USO DOS SISTEMAS DE INFORMAÇÃO GEOGRÁFICA (SIG) NO ESTUDO DA ACESSIBILIDADE FÍSICA AOS SERVIÇOS DE SAÚDE PELA POPULAÇÃO RURAL: REVISÃO DA LITERATURA. **Revista Brasileira de Geografia Médica e da Saúde**, v. 8, n. 15, p. 178–189, 2012.

FETTERLY, D.; MANASSE, M.; NAJORK, M. A Large-Scale Study of the Evolution of Web Pages. **Software: Practice & Experience**, v. 34, n. 2, p. 213–237, 2004.

FITZ, P. R. **Geoprocessamento Sem Complicação**. Oficina de Textos ed. São Paulo: Oficina de Textos, 2008.

FLORCZYK, A. J. et al. Semantic selection of georeferencing services for urban management. **Electronic Journal of Information Technology in Construction**, v. 15, p. 111–121, 2010.

FOURSQUARE. **Foursquare - A empresa confiável de inteligência de dados de localização**. Disponível em: <https://pt.foursquare.com/>. Acesso em: 30 jul. 2020.

FUNDAÇÃO OPENSTREETMAP. **OpenStreetMap**. Disponível em: <https://www.openstreetmap.org/copyright>. Acesso em: 30 nov. 2020.

GEOPY, C. **Welcome to GeoPy's documentation!** Disponível em: <https://geopy.readthedocs.io/en/stable/>. Acesso em: 17 set. 2020.

GLEZ-PEÑA, D. et al. Web scraping technologies in an API world. **Briefings in Bioinformatics**, v. 15, n. 5, p. 788–797, 30 abr. 2013.

GÓMEZ-PÉREZ, A, CORCHO, O. Ontology languages for the semantic web, **IEEE Intelligent systems**, v. 17, n. 1, p. 54-60, 2002.

GOOGLE INC. **Google**. Disponível em: <https://www.google.com/webhp?hl=pt-BR&sa=X&ved=0ahUKEwj6i5-1-ZLmAhUcDrkGHb18DLMQPAgH>. Acesso em: 30 nov. 2020.

GOOGLE MAPS, P. **Plataforma do Google Maps | Google Developers**. Disponível em: <https://developers.google.com/maps/documentation?hl=pt-br>. Acesso em: 4 nov. 2020.

GREEN, D. **Prefeitura de São Cristóvão é Pioneira na Digitalização dos Documentos**. Disponível em: <https://suporte.greendoc.com.br/noticia/2/prefeitura-de-sao-cristovao-e-pioneira-na-digitalizacao-dos-documentos>. Acesso em: 28 nov. 2019.

GUIMARÃES, J. W. Elaboração e construção de um protótipo mínimo viável para o Tingoram : um sistema de mineração de dados web baseado em georreferenciamento para

sugestão semi automatizada de doação de alimentos. 2018.

GREGG, D. G., WALCZAK, S. Adaptive web information extraction, **Communications of the ACM**, v. 49, n. 5, p. 78-84, 2006.

HAKLAY, M., WEBER, P., Openstreetmap: User-generated street maps, **IEEE Pervasive Computing**, v. 7, n. 4, p. 12-18, 2008.

HIGOUNET, C. **História Concisa da Escrita**. Ed. Parábola, 2003.

IDEAL MARKETING. **O que é sitemap XML e por que usar um mapa no seu site?** Disponível em: <https://www.idealmarketing.com.br/blog/o-que-e-sitemap/>. Acesso em: 29 nov. 2020.

KADAM, V. B., PAKLE, G. K., A survey on HTML structure aware and tree based web data scraping technique, **International Journal of Computer Science and Information Technologies**, v. 5, n. 2 , p. 1655-1658, 2014.

KAUSAR, M. A.; DHAKA, V. S.; SINGH, S. K. Web Crawler: A Review. **International Journal of Computer Applications**, v. 63, n. 2, p. 31–36, 15 fev. 2013.

MALIK, S. K.; RIZVI, S. **Information extraction using web usage mining, web scrapping and semantic annotation**. Proceedings - 2011 International Conference on Computational Intelligence and Communication Systems, CICN 2011. **Anais**...IEEE, out. 2011Disponível em: <http://ieeexplore.ieee.org/document/6112910/>. Acesso em: 3 fev. 2021

MYERS, G., A fast bit-vector algorithm for approximate string matching based on dynamic programming, **Journal of the ACM**, v. 46, n. 3, May 1999.

MITCHELL, R. E. **Web Scraping with Python Collecting Data from the Modern Web**. First ed. Sebastopol: O'Reilly Media, 2015.

MONGODB, I. **The most popular database for modern apps | MongoDB**. Disponível em: <https://www.mongodb.com/>. Acesso em: 3 fev. 2021.

NEDER, H. D. et al. Índice de defasagem do Imposto Predial e Territorial Urbano (IPTU) dos Municípios de Minas Gerais : um estudo de caso para Uberlândia ( MG ). Brasil. **Revista ESPACIOS**, v. 38, n. 46, p. 25–39, 23 jun. 2017.

NEIL, Y. Web Scraping the Easy Way. **University Honors Program Theses**, 1 jan. 2016.

NIANTIC, I. **Pokémon GO**. Disponível em: <https://pokemongolive.com/pt_br/>. Acesso em: 30 nov. 2019.

NURSEITOV, N., PAULSON, M., REYNOLDS, R., & IZURIETA, C. Comparison of JSON and XML data interchange formats: a case study. **Caine**, 9, 157-162, 2009.

OLIVEIRA FILHO, P. C. DE; SILVA, S. V. K. DA K. DA. Um sistema de informações para suporte espacial e de decisões à gestão da arborização urbana no município de Guarapuava, Paraná. **Revista da sociedade brasileira de arborização urbana**, v. 5, n. 3, p. 82–96, 2010.

OLSTON, C., NAJORK, M., Web crawling, **Foundations and Trends in Information Retrieval**, v. 4, n. 3, p. 175-246, 2010.

OZIMEK, A.; MILES, D. Stata utilities for geocoding and generating travel time and travel distance information. **The Stata Journal: Promoting communications on statistics and Stata**, v. 11, n. 1, p. 106–119, 19 mar. 2011.

PALANIYANDI, M., The environmental aspects of dengue and chikungunya outbreaks in India: GIS for epidemic control. **International Journal of Mosquito Research**, v. 1, n. 2, p. 38-44, 2014.

PARVEZ, M. S. et al. **Analysis of Different Web Data Extraction Techniques**. 2018 International Conference on Smart City and Emerging Technology, ICSCET 2018. **Anais**...Mumbai, India: IEEE, jan. 2018Disponível em: <https://ieeexplore.ieee.org/document/8537333/>. Acesso em: 3 fev. 2021

Périco, E., Cemin, G., Planejamento do uso do solo em ambiente SIG: alocação de um distrito industrial no município de Lajeado RS Brasil, **Estudos Geográficos: Revista Eletrônica de Geografia**, v. 4 n. 1, p. 41-52, 2006.

PYTHON SOFTWARE FOUNDATION. **Welcome to Python.org**. Disponível em: <https://www.python.org/>. Acesso em: 30 nov. 2019.

REITZ, K. **Requests-HTML: HTML Parsing for Humans (writing Python 3)!** . Disponível em: <https://requests-html.kennethreitz.org/>. Acesso em: 2 dez. 2020.

RICHARDSON, L. **Documentação Beautiful Soup**. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc.ptbr/>. Acesso em: 2 dez. 2020.

SCHRENK, M. **Webbots, spiders, and screen scrapers : a guide to developing Internet agents with PHP/CURL**. [s.l.] No Starch Press, 2012.

SILVA, A. N. R. DA. **Sistemas de Informações geográficas para planejamento de transportes**. [s.l.] Universidade de São Paulo, 1998.

SILVA, M. C. Sistemas De Informações Geográficas Na Identificação De Doenças E Epidemias. **Tekhne e Logos**, v. 8, n. 4, p. 94–106, 2017.

SILVEIRA, I. H. DA; OLIVEIRA, B. F. A. DE; JUNGER, W. L. Utilização do Google Maps para o georreferenciamento de dados do Sistema de Informações sobre Mortalidade no município do Rio de Janeiro, 2010-2012*. **Epidemiologia e Serviços de Saúde**, v. 26, n. 4, p. 881–886, nov. 2017.

SOLINA, F.; RAVNIK, R. **Georeferencing works of literature**. Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces. **Anais**...Cavtat, Croatia: IEEE, 2010Disponível em: <https://ieeexplore.ieee.org/document/5546410>

SYME, D., BATTOCCHI, K., TAKEDA, K., MALAYERI, D., FISHER, J., HU, J., CHAE,

W., Strongly-typed language support for internet-scale information sources. **Technical Report** MSR-TR-2012–101, Microsoft Research., 2012

SZTUTMAN, P. **Análise da qualidade posicional das bases do Google Maps, Bing Maps e da Esri para referência espacial em projetos em SIG: aplicação para o município de São Paulo.** São Paulo: Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, 9 dez. 2014.

TRIPADVISOR LLC. **TripAdvisor**. Disponível em: <https://www.tripadvisor.com.br/?fid=721dacfd-5ffb-4513-b1b3-2e53e84f53b7>. Acesso em: 3 fev. 2021.

TUPÃ ESTÂNCIA TURÍSTICA. **Digitalização possibilita preservação de documentos do município - Prefeitura de Tupã**. Disponível em: <https://www.tupa.sp.gov.br/noticia/225/digitalizacao-possibilita-preservacao-de-documentos-do-municipio.html>. Acesso em: 12 fev. 2021.

VELASCO, E., AGHENEZA, T., DENECKE, K., KIRCHNER, G., ECKMANNS, T. Social media and internet-based data in global systems for public health surveillance: a systematic review. **The Milbank Quarterly**, v. 92, n. 1, p. 7-33, 2014.

WADDELL, P. A., A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim, **Environment and Planning B: Planning and Design**, v. 27, p. 247–263, 2000.

WALLAPOP. **wallapop, Local Free Classified Ads**. Disponível em: <https://www.wallapop.com/>. Acesso em: 30 nov. 2019.

WANG, J. et al. **The crawling and analysis of agricultural products big data based on Jsoup**. 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2015. **Anais**...IEEE, ago. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7382112/>. Acesso em: 03 fev. 2021

WIKIMAPIA. **WikiMapia - Vamos descrever o mundo todo!** Disponível em: <http://wikimapia.org/#lang=pt&lat=-25.090885&lon=-50.157394&z=13&m=w>. Acesso em: 25 set. 2019.

WILSON, J. P.; SWIFT, J. N.; GOLDBERG, D. W. **Geocoding best practices: Review of eight commonly used geocoding systems**. Los Angeles, CA: [s.n.].
YIN, F.; HE, X.; LIU, Z. **Research on Scrapy-Based Distributed Crawler System for Crawling Semi-structure Information at High Speed**. 2018 IEEE 4th International Conference on Computer and Communications (ICCC). **Anais**...IEEE, dez. 2018 Disponível em: <https://ieeexplore.ieee.org/document/8781062/>. Acesso em: 3 fev. 2021.