

Raspagem de Dados para Avaliação Imobiliária

Um Estudo de Caso para
Mapeamento de Riscos e
Oportunidades



11 A 15
SETEMBRO | 2023
SÃO PAULO - SP

REALIZAÇÃO



IBAPE NACIONAL
Instituto Brasileiro de Avaliações
e Perícias de Engenharia

1. Motivação do Estudo

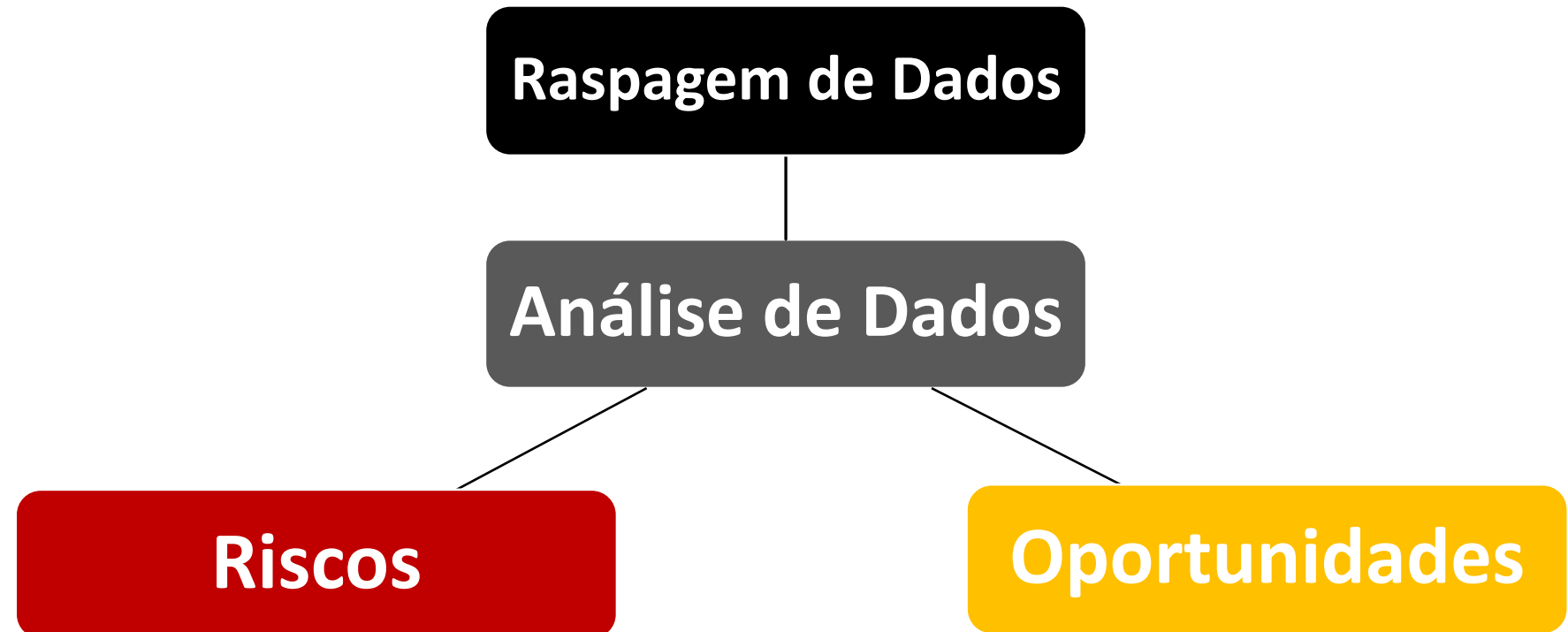
- Em face aos avanços tecnológicos de *Machine Learning*, tem-se o claro desafio quanto à automação de processos rotineiros em diversas áreas;
- Na Engenharia de Avaliações, o processo mais rotineiro e que exige mais tempo de execução mecânica é o *Levantamento de Dados de Mercado*;
- Mirando esta demanda, muitos agentes de mercado passaram a explorar a Raspagem de Dados como alternativa de otimização do Processo. Assim, faz-se necessário estudá-lo para compreender seus *riscos* e *oportunidades* inerentes.

1.1 - Conceituação

- Raspagem de Dados ou *Web Scraping* é o processo de extração automatizada de informações de websites;
- A técnica envolve o uso de um software ou script para (i) percorrer páginas, (ii) identificar e (iii) extrair dados relevantes, tais como texto, imagens, links ou qualquer outro tipo de **conteúdo estruturado**;
- Ainda, tem-se a possibilidade de realização da coleta de *grandes quantidades* de dados de maneira eficiente quanto ao tempo e recursos computacionais gastos.

1.2 - Objetivo

- ✓ O Estudo de Caso desenvolvido teve por objetivo, *através da análise de dados* obtidos por meio de **Raspagem de Dados**, avaliar os *Riscos* e *Oportunidades* do uso da técnica.



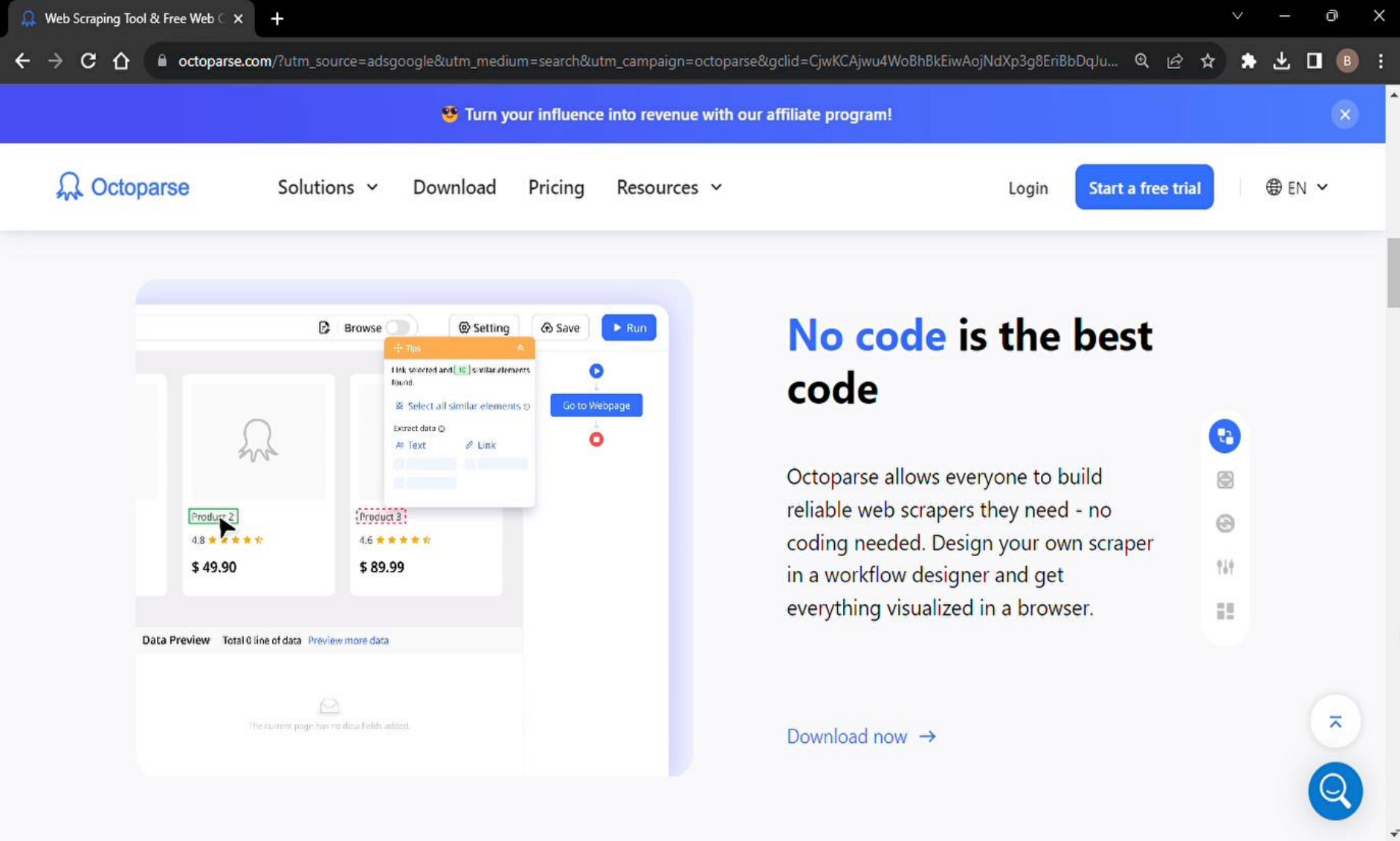
2. Metodologia



2.1 – Plataforma de Webscraping

- Optou-se pela escolha da **Octoparse** (www.octoparse.com), justificada pelas principais características listadas a seguir:
 - ✓ *Ausência de necessidade de escrita de código;*
 - ✓ *Facilidade de reconhecimento de elementos das páginas;*
 - ✓ *Possibilidade de agendamento de rotinas;*
 - ✓ *Possibilidade de progressão e reiteração entre links e páginas;*
 - ✓ *Alto grau de Personalização do fluxo de extração de dados.*

2.1 – Plataforma de *Webscraping*



Web Scraping Tool & Free Web C x +

octoparse.com/?utm_source=adsgoogle&utm_medium=search&utm_campaign=octoparse&gclid=CjwKCAjwu4WoBhBkEiwAojNdXp3g8Eri8bDqJu...

Turn your influence into revenue with our affiliate program!

Octoparse Solutions Download Pricing Resources Login Start a free trial EN

No code is the best code

Octoparse allows everyone to build reliable web scrapers they need - no coding needed. Design your own scraper in a workflow designer and get everything visualized in a browser.

Download now →

2.2 – Portais de Anúncios

- No desenvolvimento do Estudo, fez-se a análise e a opção por acesso à 02 plataformas que agrupam e ordenam anúncios imobiliários: **Viva Real** e **Netimóveis**.
- Ambas as plataformas apresentavam, como características principais e adequadas ao estudo:

Boa
Disponibilidade
de Anúncios

Qualidade de
Estruturação de
Dados

Riqueza na
Declaração de
Características

2.3 – Estrutura de Dados - Ambientes

- Mediante a escolha dos portais de anúncios, tem-se a diferenciação de 02 ambientes principais:

Catálogo Externo

Corpo Interno

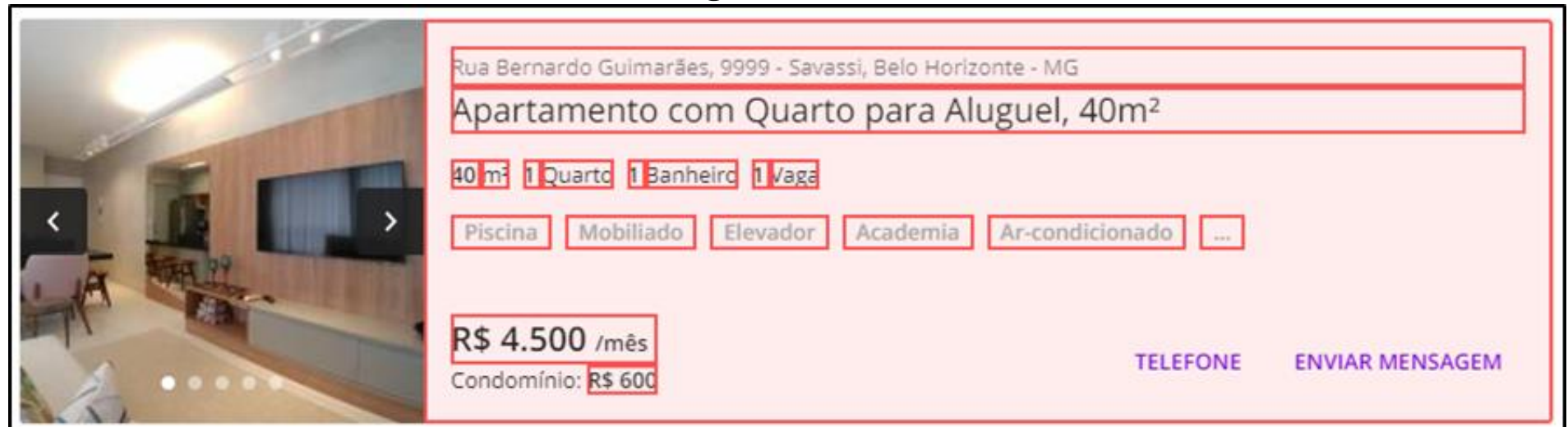
- O *Catálogo Externo* é o ambiente de listagem dos diversos anúncios apresentados na página principal.
- Já o *Corpo Interno* é o ambiente interno de cada um dos links individualizados.

2.3.1 – Catálogo Externo

- Do Catálogo Externo é possível extrair, principalmente:

(i) Pré Visualização de Imagem; (ii) Endereço; (iii) Tipologia; (iv) Comodidades; (v) Contatos;

Catálogo Externo



(vi) Valores; (vii) Metragens; (viii) Dormitórios; (ix) Banheiros; (x) Vagas.

2.3.2 – Corpo Interno

- Individualmente, tem-se no Corpo Interno de cada anúncio:
 - I. Um nível maior de detalhamento dos campos textuais;
 - II. Caixa separada com características completas das facilidades/comodidades da área comum e da unidade privativa;
 - III. A descrição textual independente elaborada pelo Autor do anúncio, conforme exemplo.

2.3.2 – Corpo Interno - Comodidades

- Características apresentadas em link interno do Anúncio

Características		
Academia	Espaço gourmet	Salão de festas
Área de lazer	Armário embutido	Armário na cozinha
Cozinha americana	Janela grande	Sala de jantar
Hidromassagem	Acesso para deficientes	Sauna
Box blindex	Escada	Porcelanato
Sala de almoço		

2.3.2 – Corpo Interno – Descrição do Autor

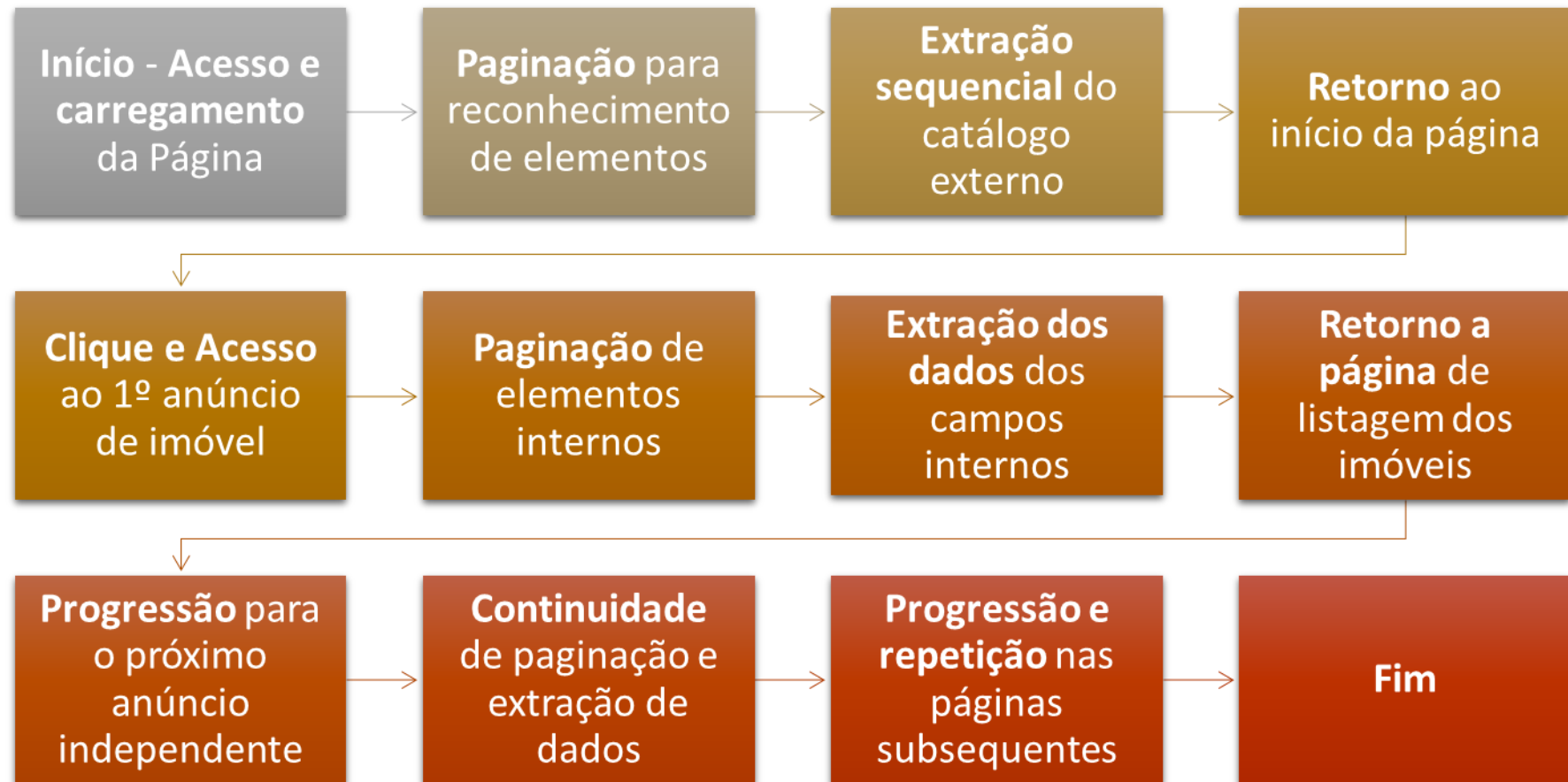
- Ainda no *Corpo Interno* tem-se a descrição textual independente elaborada de forma independente pelo Autor do Anúncio:

Descrição Textual

Apartamento para aluguel com 40 metros quadrados com 1 quarto em Savassi - Belo Horizonte - MG
O apartamento no bairro Savassi tem 40 metros quadrados com 1 quarto sendo 1 suite e 1 banheiro
Possui área de fitness, jardim, playground, salão para festas e eventos.
Vai lhe possibilitar curtir os dias mais quentes na piscina, todo o conforto do ar condicionado nos dias mais quentes. Elevador para mais praticidade no dia-a-dia.
Já tem os móveis incluídos.

2.4 – Sequência Lógica de Obtenção de Dados

- Apresenta-se a seguir o algoritmo desenvolvido para a Raspagem de Dados, separados por etapas:



3. Estudo de Caso

- Para o estudo em questão, fez-se a escolha pela região da Savassi, no município de Belo Horizonte.
- Conhecida por sua localização privilegiada, apresenta boa disponibilidade de anúncios de imóveis tanto para venda quanto locação, bem como de diversas tipologias.
- Para este trabalho, fez-se a coleta de dados de 02 tipologias principais: *Residenciais e Comerciais*.

3.1 – Classe dos Dados

- Para possibilitar a análise dos dados, fez-se a classificação categorizada dos mesmos, aqui apresentada:
 - I. Classe Textual – Dados que contenham informações cadastrais;
 - II. Classe Quantitativa – Dados que contenham informações numéricas;
 - III. Classe Qualitativa – Dados que contenham informações sobre facilidades/comodidades extras dos imóveis.

3.2 – Panorama Geral dos Dados

- Apresenta-se a seguir, o Panorama Geral dos Dados obtidos, de *forma não refinada*, divididos por Plataforma de Anúncio e Tipologia do Imóvel:

Panorama Geral de Imóveis			
	Viva Real	Netimóveis	Total
Residencial	256,00	56,00	312,00
Comercial	502,00	277,00	779,00

- Totalizaram-se, na fase de pré-refinamento, **1.091 imóveis** verificados durante a execução da raspagem de dados.

3.3 – Estruturação e Tratamento

- Diante da obtenção dos dados de forma bruta, faz-se necessária a estruturação dos dados. São etapas:
 - I. Checagem e remoção de possíveis duplicatas;
 - II. Verificação de Preenchimento incorreto dos campos das Classes obtidas;
 - III. Padronização e estruturação das informações obtidas em campos de mesma formatação.

3.4 – Dados Não Estruturados / Refinados

Id	Title1	Field1	Field2	Field3	Field4	Field5	Field6	Field7	Field8	Field9	Field10	Field11	Field12	Field13	Field14	Field15
APART001	Apartamer	Rua Tomé	Código:	964375	Valor de lo	R\$ 26.000,	Condomíni	R\$ 4.531,0	lptu	R\$ 2.048,1	(11x)					
APART002	Apartamer	Savassi – E	Código:	868722	Valor de lo	R\$ 10.000,	Condomíni	R\$ 2.700,0	lptu	R\$ 1.000,0	(11x)	área aprox	4	4	(4	4
APART003	Apartamer	Rua Fernar	Código:	960770	Valor de lo	R\$ 7.000,0	Condomíni	R\$ 1.540,0	lptu	R\$ 722,85	(12x)	área aprox	2	1	(1	2
APART004	Apartamer	Rua Gonça	Código:	805929	Valor de lo	R\$ 6.900,0	Condomíni	R\$ 1.500,0	lptu	R\$ 640,00	(11x)	área aprox	4	2	(2	3
APART005	Apartamer	Rua Tomé	Código:	957819	Valor de lo	R\$ 6.600,0	Condomíni	R\$ 684,00	lptu	R\$ 250,00	(12x)	área aprox	1	2	(1	2
APART006	Apartamer	Rua Tomé	Código:	957776	Valor de lo	R\$ 6.600,0	Condomíni	R\$ 684,00	lptu	R\$ 250,00	(12x)	área aprox	2	1	(1	2
APART007	Apartamer	Rua Levinc	Código:	957811	Valor de lo	R\$ 6.500,0	Condomíni	R\$ 1.500,0	lptu	R\$ 350,00	(12x)					
APART008	Apartamer	Rua Perna	Código:	968564	Valor de lo	R\$ 6.200,0	Condomíni	R\$ 854,00	lptu	R\$ 453,67	(12x)	área aprox	3	2	(1	2
APART009	Apartamer	Rua Perna	Código:	69386	Valor de lo	R\$ 6.000,0	Condomíni	R\$ 1.030,8	lptu	R\$ 779,98	(12x)					
APART010	Apartamer	Rua Sergip	Código:	953081	Valor de lo	R\$ 6.000,0	Condomíni	R\$ 1.382,3	lptu	R\$ 680,00	(12x)					
Id	Field16	Field17	Field18	Field19	Field20	Field21	Field22	Field23	Field24	Field25	Field26	Field27	Field28	Field29	Field30	
APART001	Mais sobre	Você está								Característ	Aquecimer	1 unidade	Piscina	4 quartos	Ar condic	
APART002	Mais sobre	PREDIO:	IMÓVEL:	BENEFICIO	"Os preços					Característ	2 unidades	4 quartos	Ar condic	4 suítes	4 banhos	
APART003	Mais sobre	Excelente	Benefícios	Localizaça	Próximo ac	Fácil acess	Perfeito !!	- Excelente	- Ampla sa	Característ	2 quartos	1 suíte	1 banho	1 varanda	1 sala	
APART004	Mais sobre	PRÉDIO10	IMÓVEL:Ex	Imóvel não	Sala para C	BENEFÍCIO				Característ	4 quartos	2 suítes	2 banhos	Dependênc	1 varanda	
APART005	Mais sobre	Benefícios	Excelente	Excelente	* Apartam	* Apartam	* Cozinha	* Acabame	* Obs: Ser	Característ	Aquecimer	Piscina	1 quarto	Ar condic	1 suíte	
APART006	Mais sobre	Benefícios	Excelente	Excelente	* Apartam	* 02 quarto	* 01 Banhe	* 01 Suite	* Cozinha	Característ	Piscina	2 quartos	1 suíte	1 banho	1 sala	
APART007	Mais sobre	Excelente	Sala de vis	01 quarto	03 quartos	Cozinha es	Área de se	DCE	02 vagas d	Característ	2 unidades	3 quartos	1 suíte	3 banhos	Dependênc	
APART008	Mais sobre	* Imóvel di	Com um es	O apartam	O prédio p	Extremame	Agende a s			Característ	Piscina	3 quartos	1 suíte	2 banhos	1 sala	
APART009	Mais sobre	Apartamer	Esse imóve	INFORMAÇ	- área de l	Aproveite	o preço po			Característ	1 unidade	Piscina	4 quartos	1 suíte	2 banhos	
APART010	Mais sobre	Apartamer	Condomíni	O Condom						Característ	Piscina	3 quartos	1 suíte	3 banhos	Dependênc	
Id	Field31	Field32	Field33	Field34	Field35	Field36	Text	Text1								
APART001	4 suítes	5 banhos	Dependênc	1 varanda	Rua Tomé	Ximenes N	Você está	Aquecimento solar	1 unidade por andar	F						
APART002	1 varanda	3 salas	Fachada fr	4 vagas	Savassi – E	Colonial N	PREDIO:Pr	2 unidades por andar	4 quartos	Ar condic						
APART003	2 vagas				Rua Fernar	Prolar Neti	Excelente	2 quartos	1 suíte	1 banho						
APART004	1 sala	3 vagas	Portaria pe	Hall social	Rua Gonça	Ximenes N	PRÉDIO10	4 quartos	2 suítes	2 banhos						
APART005	2 banhos	1 sala	2 vagas	Sala de ma	Rua Tomé	Prolar Neti	Benefícios	Aquecimento solar	Piscina	1 quarto						
APART006	2 vagas	Playground	Churrasque	Portaria pe	Rua Tomé	Prolar Neti	Benefícios	Piscina	2 quartos	1 suíte						
APART007	2 salas	2 vagas	Portaria pe	Hall social	Rua Levinc	Alphasul N	Excelente	2 unidades por andar	3 quartos	1 suíte						
APART008	2 vagas	Playground	Portaria pe	Sala de ma	Rua Perna	Luiz Vieira	* Imóvel di	Piscina	3 quartos	1 suíte						
APART009	1 varanda	1 sala	3 vagas	Sauna	Rua Perna	Adbens Ne	Apartamer	1 unidade por andar	Piscina	4 quartos						
APART010	1 varanda	1 sala	2 vagas	Playground	Rua Sergip	Adbens Ne	Apartamer	Piscina	3 quartos	1 suíte						

4 – Análise dos Dados

- A análise técnica se baseou inicialmente na **estruturação** dos dados dos 1091 imóveis para posterior identificação de inconsistências. Essa abordagem foi aplicada às classes Textuais e **Numéricas**.
- Inicialmente, foram **(i)** eliminados dados em duplicidade. Em seguida, fez-se o filtro de **(ii)** imóveis com declaração incompleta de endereço, e ainda, da **(iii)** verificação de disponibilidade de imagens para conferência.
- Somente após a eliminação das inconsistências, foi possível avançar para a análise dos campos de classe **Qualitativa**.

4.1 – Inconsistências Textuais e Numéricas

- Do levantamento inicial de dados, foram subtraídos 89 anúncios repetidos, restando 1.002 anúncios válidos.

Panorama de Remoção de Duplicatas de Imóveis						
	Viva Real			Netimóveis		
	Iniciais	Duplicatas	Restantes	Iniciais	Duplicatas	Restantes
Residencial	256,00	7,00	249,00	56,00	0,00	56,00
Comercial	502,00	10,00	492,00	277,00	72,00	205,00

- Verifica-se também, para fins cadastrais, a qualidade da declaração dos endereços de cada um dos anúncios remanescentes

Panorama de Declaração de Informação Cadastral Incompleta de Endereço						
	Viva Real			Netimóveis		
	Bairro	Logradouro	Nº Completo	Bairro	Logradouro	Nº Completo
Residencial	47,00	133,00	69,00	12,00	23,00	21,00
Comercial	59,00	220,00	213,00	60,00	91,00	54,00

4.1 – Inconsistências Textuais e Numéricas

- Verificou-se ainda, os anúncios que não possuem imagens para verificação das características declaradas pelo Anunciante, apresentadas a seguir:

Panorama de Indisponibilidade de Imagens para Conferência				
	Viva Real		Netimóveis	
	Disponíveis	Indisponíveis	Disponíveis	Indisponíveis
Residencial	249,00	0,00	49,00	7,00
Comercial	469,00	23,00	161,00	44,00

- Do total de 1.002 imóveis, já removidas as duplicatas, outros 74 imóveis não dispunham de imagens para conferência.

4.2 – Informações Qualitativas

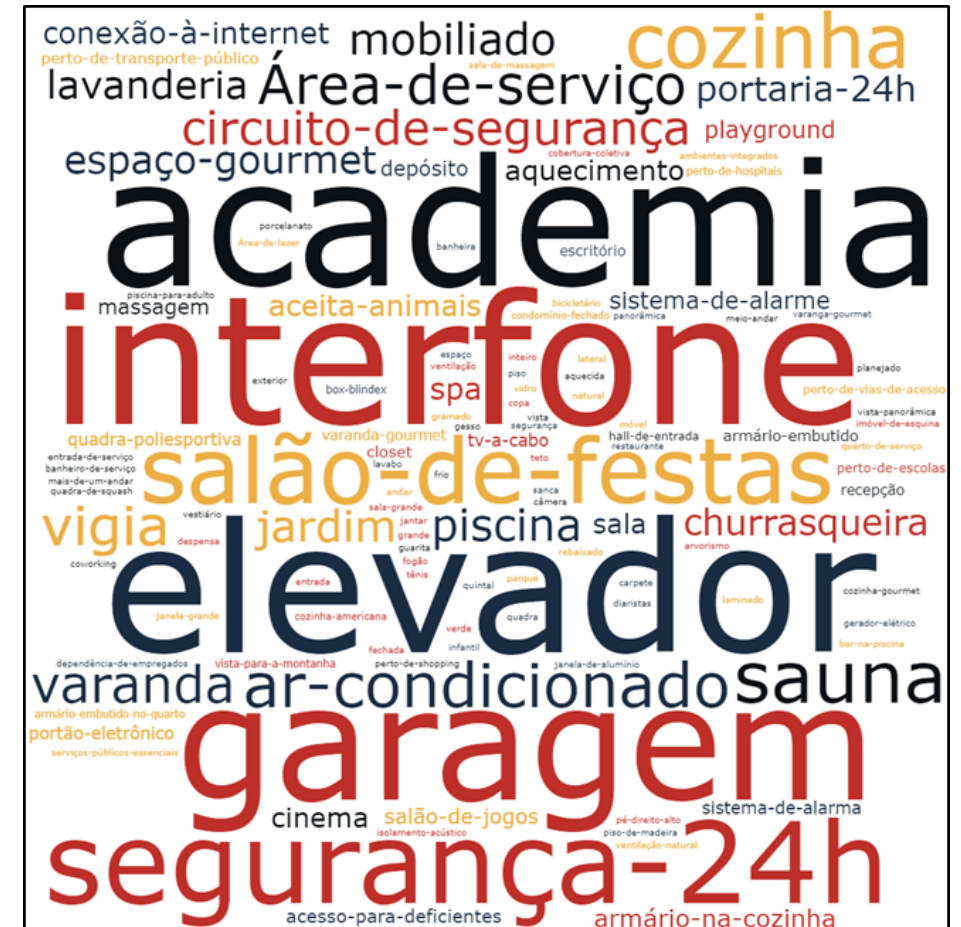
- Realizando-se o levantamento quantitativo das características declaradas na Classe Qualitativa, pode-se observar o seguinte universo de distribuição:

Declaração de Características Qualitativas - Qtde de Termos Independentes		
	Viva Real	Netimóveis
	Caixa de Características Interna	Caixa de Características Interna
Residencial	128,00	60,00
Comercial	86,00	32,00

- Importante observar que cada um dos portais tem uma estrutura distinta de declaração de dados qualitativos, influenciando diretamente na quantidade e qualidade das informações declaradas.

4.3 – Nuvem de Palavras

- Realizou-se, de forma quantitativa, o levantamento do histograma de frequência dos termos, para cada um dos cenários considerados. Visando facilitar a compreensão e dar uma visão global destas informações, traduziu-se este histograma em uma **Nuvem de Palavras** dos termos:



4.4 – Observações Qualitativas

- A *estrutura de campos disponíveis* para declaração de informações em cada portal interfere diretamente na qualidade da disponibilização dos dados;
- Existem muitas inconsistências entre *informações qualitativas declaradas em campos pré-existentes vs. Campo de descrição textual do Autor*;
- Alguns portais permitem que características textuais e **quantitativas** sejam declaradas somente na descrição interna do anúncio, misturando junto as características **qualitativas**;
- Os termos mais comumente declarados internamente ao anúncio, como “elevador”, “interfone” e “garagem” aparecem com frequência muito maior comparado aos demais. 2/3 dos termos qualitativos declarados, aparecem com uma frequência inferior a 25,00% do termo de maior frequência.

5 – Conclusões

Riscos

Dado o Estudo de Caso em tela, pode-se chegar as seguintes constatações, quanto aos **Riscos**:

- A estruturação do portal e das permissões de declaração de dados interfere diretamente na qualidade e disponibilidade de informações;
- Diferentes estruturas de declaração de dados geram um alto nível de incompatibilidade para comparações diretas, sendo um fator de dificuldade, a princípio, para uso em larga escala;
- A qualidade da declaração das informações e o uso correto dos campos adequados, exercida de forma independente pelo Anunciante, é primordial para a qualidade da extração de dados afeta diretamente a qualidade das análises;
- A não realização de conferência visual, técnica que demanda investimentos mais robustos em recursos computacionais e tempo de execução, diminui o nível de confiabilidade da análise dos dados.

5 – Conclusões

Oportunidades

- A análise das informações Textuais e Numéricas, depois de eliminadas inconsistências, podem se tornar fontes valiosas de dados para o uso nas avaliações imobiliárias, pois são de estruturação e tratamento de nível mais fácil e rápido quando comparados a Qualitativa;
- A análise de um histograma de frequência e da nuvem de palavras, pode indicar qualitativamente termos mais declarados, e por consequência, com maior percepção de valor para aquele determinado mercado;
- A realização da Raspagem de Dados feita periodicamente pode reduzir as inconsistências e o descarte de dados, bem como fornecer informações de difícil apuração, tais como (i) a velocidade de comercialização de um determinado imóvel e das (ii) flutuações de preços declarados no mercado, dentre outras.

5.1 – Considerações Finais

- A dependência da declaração dos dados por agentes de mercado pouco qualificados configura ponto crucial no aproveitamento efetivo dos dados;
- Ainda, verifica-se restrições quanto a Raspagem de Dados em mercados onde a disponibilidade de dados é escassa ou mal declarada, reduzindo uma possível escala de produção;
- Assim, o uso da Raspagem de Dados, realizado de forma rápida, demonstra-se muito mais adequado à geração de conhecimento de mercado do que à aplicação direta nos modelos de avaliação imobiliária;

5.2 – Observações para Trabalhos Futuros

Tem-se como sugestões para trabalhos futuros:

- i. A realização de Raspagem de Dados utilizando de maior abrangência de regiões diferentes;
- ii. A execução de rotina de Web Scraping periódicas, para verificação de flutuações e alterações durante um período pré determinado de tempo,
- iii. Por fim, da realização da inclusão da verificação visual das imagens disponibilizadas junto aos anúncios.

11 A 15
SETEMBRO | 2023
SÃO PAULO - SP

REALIZAÇÃO

5.3 – Contatos



Eng. Bruno Henrique Gazzinelli

Linkedin:



www.bhgengenharia.com
bruno@bhgengenharia.com

(31) 99300-9609

(31) 98861-6112

