INSTITUTO DE INFORMÁTICA
Universidade Federal de Goiás

Vagas remotas & skills emergentes

Especialização em Sistemas e Agentes Inteligentes
Disciplina: Extração Automática de Dados
Prof. Otávio Calaça Xavier

Carolina Pastor Humpiri
Hugo Ferreira Ginu
Pedro Moacir de Carvalho
Rafael Ferreira Peixoto

21/06/2025

INF
INSTITUTO DE INFORMÁTICA

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Agenda

- Problema & importância.
- Arquitetura do pipeline (extração → engenharia → análise).
- Principais descobertas (gráficos/insights).
- Reflexões éticas e legais (LGPD, direitos autorais, robots.txt).

Problema & importância.

Problema & importância.

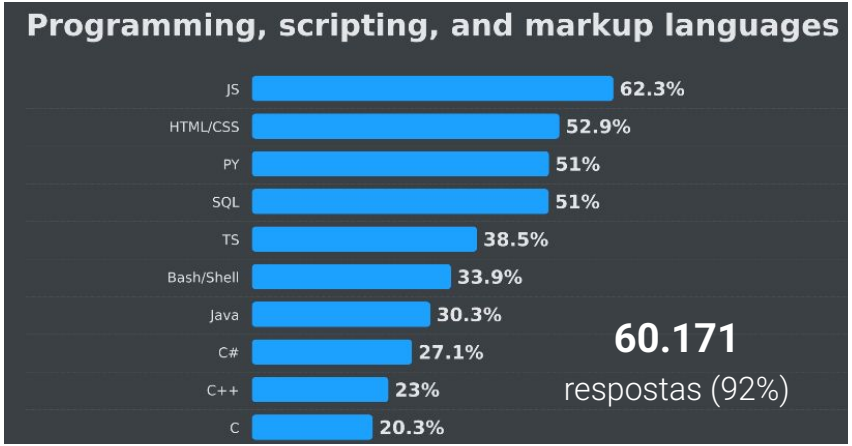
- Mercado de trabalho x Skills
- Uso de scrap para busca pessoal de vagas

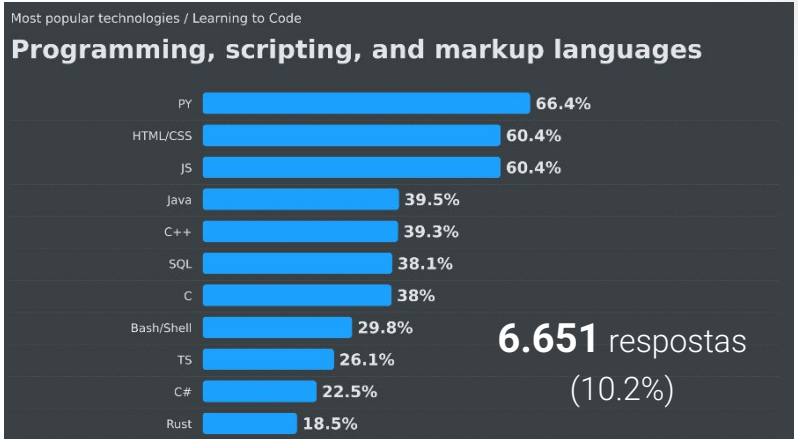
Objetivo do trabalho propõe uma aplicação prática das técnicas para coletar e estruturar dados disponíveis na web, com foco na análise de vagas de emprego remotas e nas habilidades (skills) mais exigidas no mercado.

Stack Overflow Annual Developer Survey

Pesquisa Anual Stackoverflow

Pesquisa Anual com **65.437** respondentes em 2024 de **185** países.

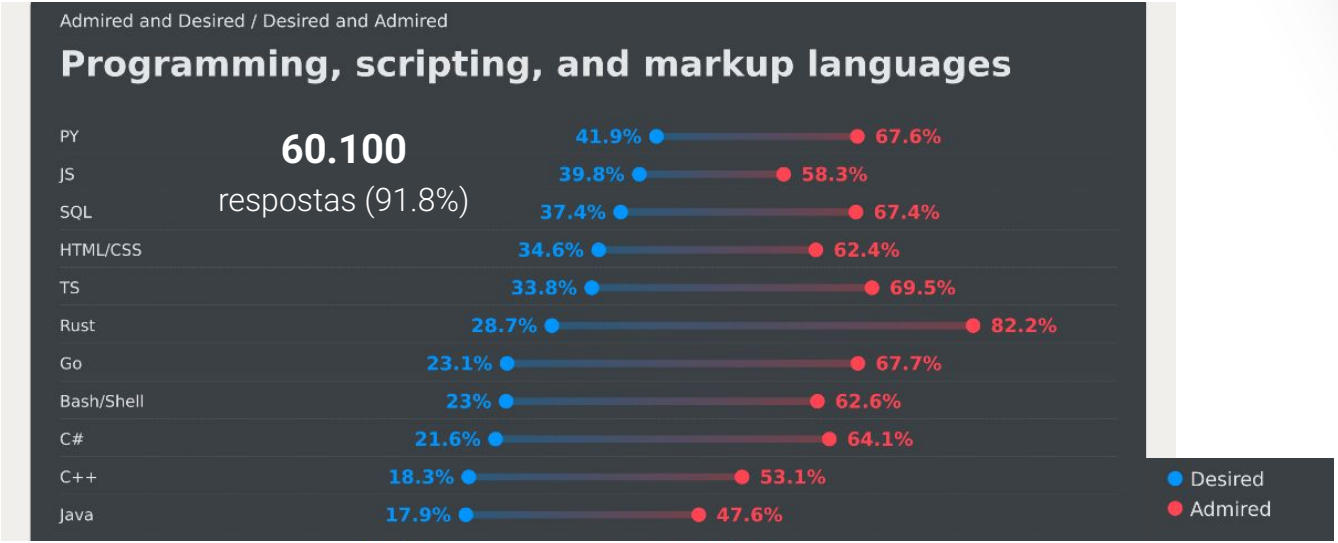




"Em quais linguagens de programação, script e marcação você realizou trabalho de desenvolvimento extensivo no último ano e em quais você deseja trabalhar no próximo ano? (Se você tanto trabalhou com a linguagem quanto deseja continuar trabalhando, marque as duas caixas nessa linha.)"

Fonte: https://survey.stackoverflow.co/2024/technology. Acesso em 21/06/2025.

Pesquisa Anual Stackoverflow



"Em quais linguagens de programação, script e marcação você realizou trabalho de desenvolvimento extensivo no último ano e em quais deseja trabalhar no próximo ano? (Se você tanto trabalhou com a linguagem quanto deseja continuar trabalhando, marque ambas as opções nessa linha.)"

Fonte: https://survey.stackoverflow.co/2024/technology. Acesso em 21/06/2025.

remoteok.com

robots.txt

User-agent: *

Crawl-delay: 1

Allow: /

Sitemap: https://remoteok.com/sitemap.xml

User-agent: AhrefsBot

Disallow: /

Disallow: /*?action=get_jobs

Disallow: /*?*action=get_jobs

Disallow: /*?*&action=get_jobs

Disallow: /*?*&action=get_jobs&*

robots.txt

User-agent: *

Crawl-delay: 1

Allow: /

Sitemap: https://remoteok.com/sitemap.xml

Disallow: /*?action=get_jobs

robots.txt - SEO (Search Engine Optimization) - Otimização de Mecanismos de Busca

```
User-agent: SemrushBot
Disallow: /

User-agent: SemrushBot-BA
Disallow: /

# Majestic - SEO
User-agent: MJ12bot
Disallow: /

User-agent: Screaming Frog SEO Spider
Disallow: /

# Seozoom - SEO
User-Agent: ZoomBot
Disallow: /

User-agent: sistrix
Disallow: /

User-agent: serpstatbot
Disallow: /

User-agent: MozBot
Disallow: /

#
https://moz.com/help/moz-procedures/crawlers/rogerbot
User-agent: rogerbot
Disallow: /

#
https://moz.com/help/moz-procedures/crawlers/dotbot
User-agent: dotbot
Disallow: /

User-agent: DataForSeoBot
Disallow: /

Disallow: /l/
Disallow: /*ou-tout-autre*
Disallow: /*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
```

sitemap.xml


remoteok.com/sitemap.xml
This XML file does not appear to have any style information associated with it. The document tree is shown below.
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
<sitemap>
<loc>https://remoteok.com/sitemap-predefined-1.xml</loc>
</sitemap>
<sitemap>
<loc>https://remoteok.com/sitemap-jobs-1.xml</loc>
</sitemap>
<sitemap>
<loc>https://remoteok.com/sitemap-jobs-2.xml</loc>
</sitemap>
<sitemap>
<loc>https://remoteok.com/sitemap-jobs-3.xml</loc>
</sitemap>
<sitemap>
<loc>https://remoteok.com/sitemap-jobs-4.xml</loc>

sitemap.xml

```
[ ] # URL do Sitemap
    sitemap_url = "https://remoteok.com/sitemap.xml"

    response = requests.get(sitemap_url, headers=headers)
    if response.status_code == 200:
        soup = BeautifulSoup(response.content, 'xml') # xml parser
        urls = [loc.get_text(strip=True) for loc in soup.find_all('loc')]

        print(f"Encontradas {len(urls)} URLs no sitemap:")
        for url in urls:
            print(url)

    else:
        print(f'Erro na requisição do sitemap. Status code: {response.status_code}')
```

```
Encontradas 75 URLs no sitemap:
https://remoteok.com/sitemap-predefined-1.xml
https://remoteok.com/sitemap-jobs-1.xml
https://remoteok.com/sitemap-jobs-2.xml
https://remoteok.com/sitemap-jobs-3.xml
https://remoteok.com/sitemap-jobs-4.xml
https://remoteok.com/sitemap-jobs-5.xml
https://remoteok.com/sitemap-jobs-6.xml
```

Sitemap.xml - Sitemap 0.90

https://www.sitemaps.org/pt_BR/index.html

O que são Sitemaps?

Os Sitemaps são uma forma fácil para os webmasters de informar os mecanismos de pesquisa sobre seus sites disponíveis para indexação. Em sua composição mais simples, um Sitemap é um arquivo XML que relaciona os URLS de um site junto com metadados adicionais sobre cada URL (quando ele foi atualizado pela última vez; com que freqüência ele é alterado; qual a sua importância em relação a outros URLs no site) para que os mecanismos de pesquisas possam indexar o site de maneira mais inteligente.

Em geral, indexadores descobrem páginas com base em links no site e outros sites. Os Sitemaps complementam esses dados para permitir que os indexadores com suporte para Sitemaps peguem todos os URLs no Sitemap e aprendam sobre esses URLs usando os metadados associados. O uso do protocolo de Sitemap não garante que as páginas da Web sejam incluídas nos mecanismos de pesquisa, mas fornece dicas para que os indexadores sejam mais eficientes na indexação do seu site.

O Sitemap 0.90 é oferecido de acordo com os termos da Attribution-ShareAlike Creative Commons License e é amplamente empregado, contando com o apoio da Google, Yahoo! e Microsoft.

sitemap.xml

https://remoteok.com/sitemap-predefined-1.xml

https://remoteok.com/sitemap-jobs-1.xml (...)

https://remoteok.com/sitemap-jobs-13.xml[1-13]

https://remoteok.com/sitemap-tags-1.xml

https://remoteok.com/sitemap-companies-1.xml (...)

https://remoteok.com/sitemap-companies-4.xml [1-4]

https://remoteok.com/sitemap-users-1.xml (...)

https://remoteok.com/sitemap-users-54.xml [1-54]

https://remoteok.com/sitemap-countries-1.xml

https://remoteok.com/sitemap-cities-1.xml

Limitações com scraping - sitemap.xml

65.000+ vagas

1s -> 18h+

Após 1h -> **Erro 429 Too Many Requests**, mesmo respeitando 1s de Crawl-delay (foi usado 2s)

Colab -> Após ~2h perde conexão

1000/65000

Limitações com Crawling

Click do selenium barrado pela Cloudfare,

embora rolagem tenha sido permitida

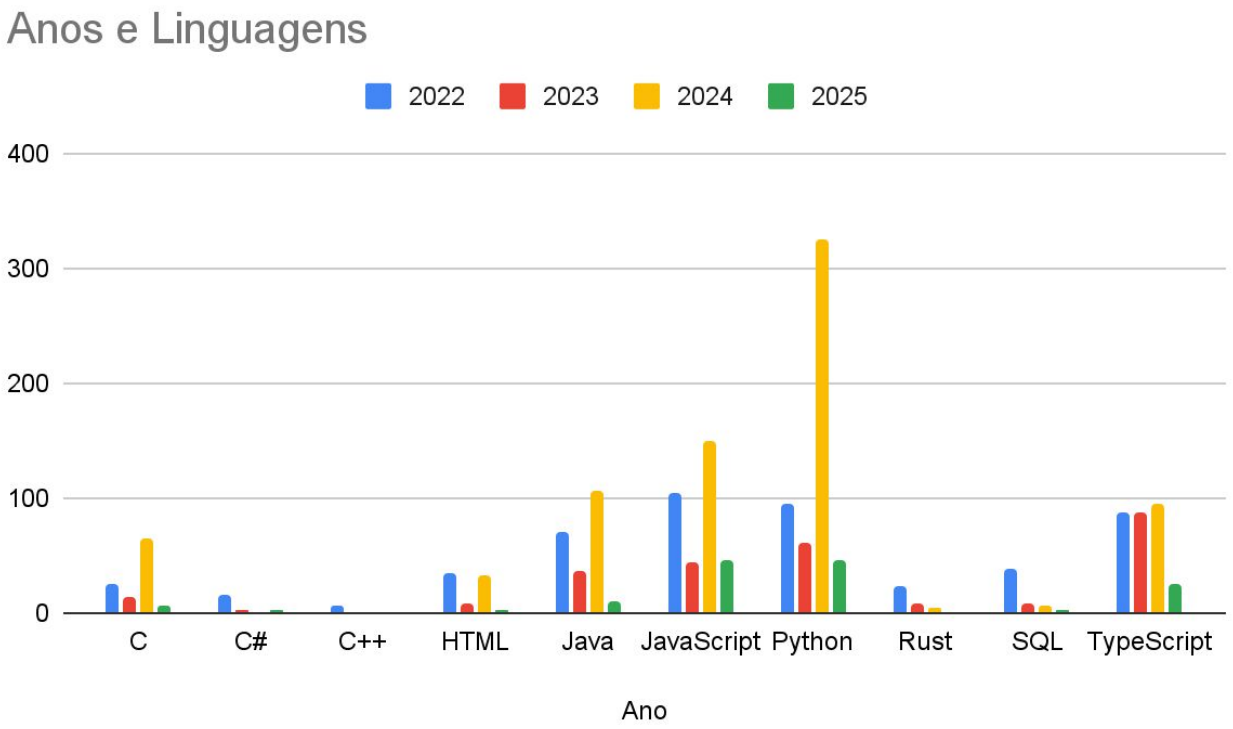
Resumo das estratégias

Estratégia	Comentários
Usando o sitemap.xml, ler todos as páginas de anúncio (65000) e extrair o texto da vaga e data.	Conseguimos 1000 antes do site blockear, agora recebemos Erro 429. Demoraria 18h sem o bloqueio. Mas teriamos um **dataset completo de vagas do site**.
Usando o sitemap.xml ou tags do próprio site, fazer crawling (rolagem de barra para carregar site) e ler para cada linguagem em ordem cronológica.	Mais simples e serviria para contar a frequência de cada linguagem nas datas.
Crawling + Scrap	O site a partir do segundo clique é bloqueado pelo Cloudfare.
API Oculta	Politeness: robots.txt

Resultado do sitemap + crawling + scrap

Ano	C	C#	C++	HTML	Java	JavaScript	Python	Rust	SQL	TypeScript
2022	25	15	6	35	71	104	95	24	39	88
2023	14	3	0	9	37	45	62	8	9	88
2024	65	0	0	32	107	149	326	4	6	95
2025	6	2	0	3	11	46	46	1	2	25

Resultado do sitemap + crawling + scrap



Resultado

- Para o artigo melhorar as análises
- Resultado
  - Python
  - JavaScript
  - Java
  - HTML
  - TypeScript

Obrigado