

Web Scraping e Datasets

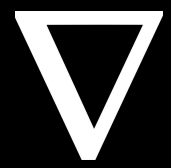
Introdução ao web
scraping com python



Nicolas de Sousa Maia
nicolasdesousamaia@usp.br

The background features a large, stylized 'V' shape composed of three overlapping triangles. The outermost triangle is light blue, the middle one is light purple, and the innermost one is light pink. A blue line starts from the left, extends horizontally, and then turns 90 degrees downward. A pink line starts from the right, extends horizontally, and then turns 90 degrees upward. Both lines meet at the 'Kaggle' text.

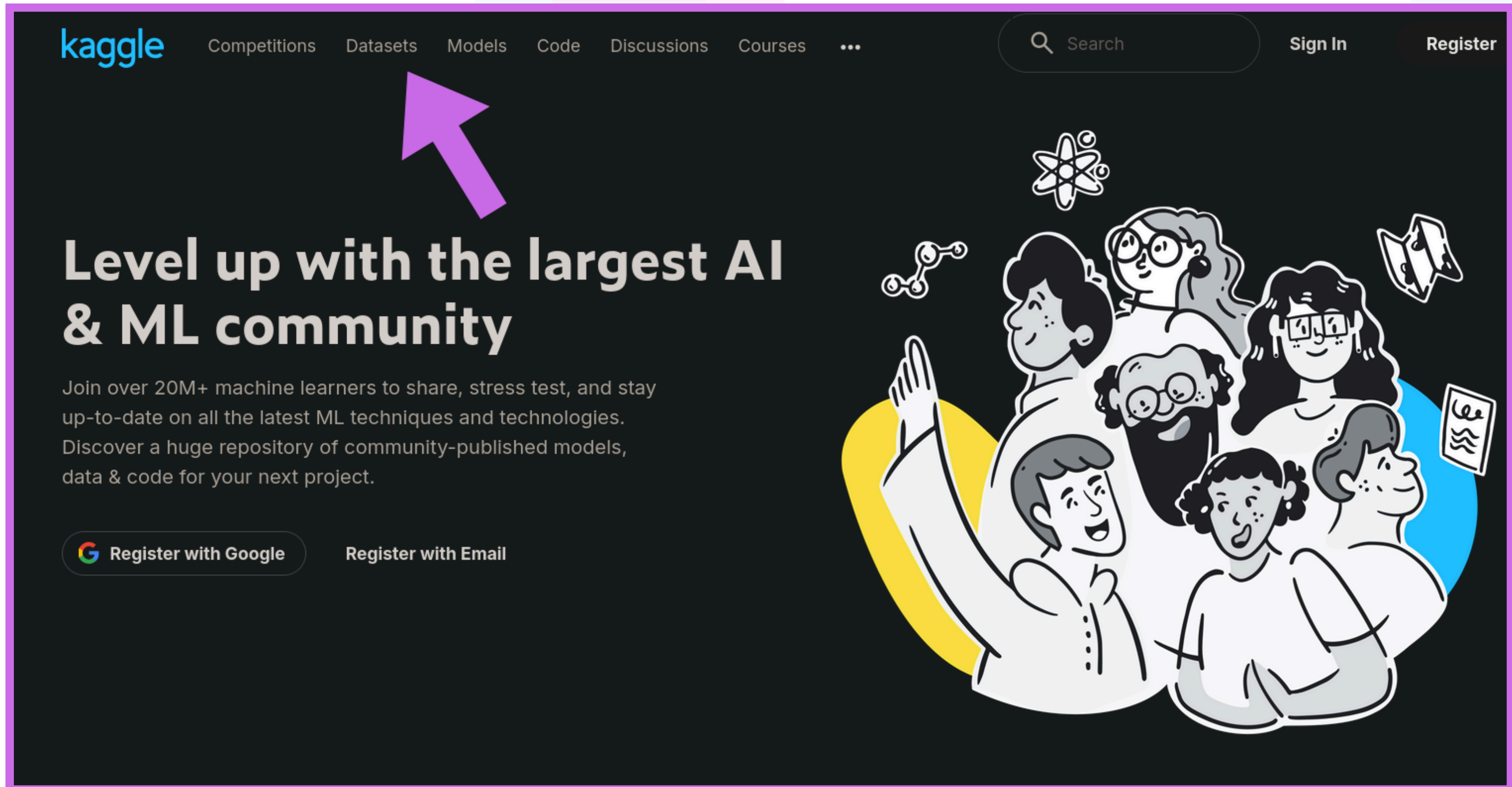
Kaggle



Um pouco sobre o Kaggle

- subsidiária da Google
- comunidade de cientistas de dados e engenheiros de machine learning
- miríade de **datasets**
- competições e discussões de data science

Encontrando Datasets no Kaggle




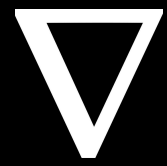
The image shows the Kaggle website homepage. At the top, there is a navigation bar with the Kaggle logo, links for Competitions, Datasets, Models, Code, Discussions, Courses, and a search bar. A purple arrow points to the 'Datasets' link. Below the navigation bar, the main heading reads 'Level up with the largest AI & ML community'. Underneath this, a paragraph states: 'Join over 20M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.' At the bottom of the main content area, there are two buttons: 'Register with Google' and 'Register with Email'. On the right side of the page, there is a large illustration of a diverse group of people, some wearing glasses, with various icons like a brain, a lightbulb, and a document floating around them.

kaggle Competitions Datasets Models Code Discussions Courses ... Search Sign In Register

Level up with the largest AI & ML community

Join over 20M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

 Register with Google Register with Email




Variadade de Datasets

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset

 Search datasets

 Filters

All datasets

Computer Science

Education

Classification

Computer Vision

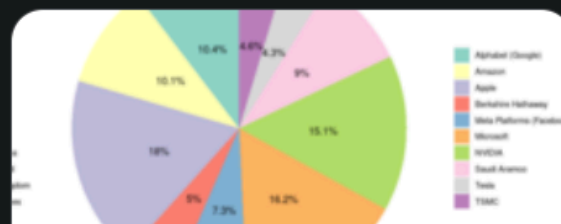
NLP

Data Visualization

Pre-Trained Model

Trending Datasets

[See All](#)




World Top Companies: Key Financial Analysis

Patrick L Ford · Updated 18 days ago
Usability 10.0 · 1 MB
5 Files (CSV)

▲ 13



Nvidia Daily Stock Price Data

 Julia Zwittlinger · Updated 24 days ago
Usability 10.0 · 123 kB
1 File (CSV)

▲ 9



Crime Dataset

Haseef Alam · Updated 22 days ago
Usability 10.0 · 50 MB
1 File (CSV)

▲ 13

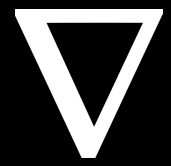


Forecasting Disaster Management in 2024

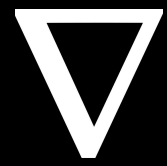
Shaik Barood Mohammed Umar Adnaa...
Usability 10.0 · 300 kB
1 File (CSV)

▲ 7

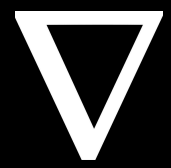




Exemplo de Dataset

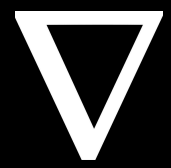
The background features a large, stylized 'V' shape composed of three overlapping triangles in light blue, light purple, and light pink. A blue line extends from the left, and a pink line extends from the right, both framing the central text.

Web Scrapping



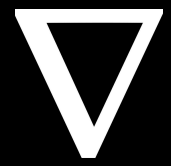
O que é Web Scraping?

- técnica de **extração automatizada de dados** de sites na web
- converte o conteúdo de páginas HTML em informações estruturadas para análise ou uso posterior
- maneira de se obter um **conjunto de dados** de uma página web

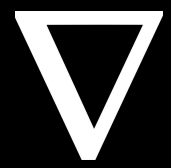


O que é uma página HTML?

- HTML = HyperText Markup Language
- linguagem usada para estruturar conteúdo na web
- Tags: `<p>` `<h1>` ``
- permite entender o formato de sites,
facilitando a busca do dados desejados

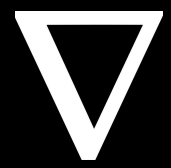


Montando uma página básica em HTML



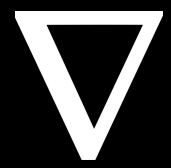
Como Web Scraping funciona

- **Acessa** páginas web como um navegador
- **Identifica e extrai** dados específicos (títulos, preços, imagens, valores)
- **Armazena** os dados em um formato útil, como CSV, JSON, ou bancos de dados



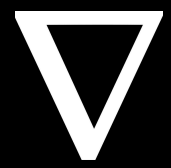
Beautiful Soup

- biblioteca em Python que **permite a extração de dados** que estão em HTML
- útil para **páginas estáticas**
- ela e **Requests** são as principais livrarias utilizadas em Web Scraping desse tipo de página



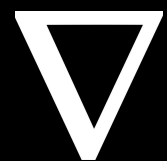
Selenium

- outra biblioteca disponível no Python, criada para lidar com sites feitos em JavaScript
- o programa torna-se capaz de **simular um navegador**
- permite ao programador o acesso aos dados de **páginas dinâmicas**



Considerações Legais

- **Termos de Serviço (ToS):**
 - verifique sempre os ToS do site para garantir que o scraping é permitido
- **Leis de Proteção de Dados:**
 - GDPR/LGPD: evite coletar dados pessoais sem consentimento
- **Direitos Autorais:**
 - conteúdos protegidos por direitos autorais não devem ser reutilizados sem permissão

The background features a large, stylized 'V' shape composed of three overlapping triangles in light blue, light purple, and light pink. A blue line starts from the left, extends horizontally, and then turns vertically downwards. A pink line starts from the right, extends horizontally, and then turns vertically upwards. These lines frame the central text.

Fim da aula