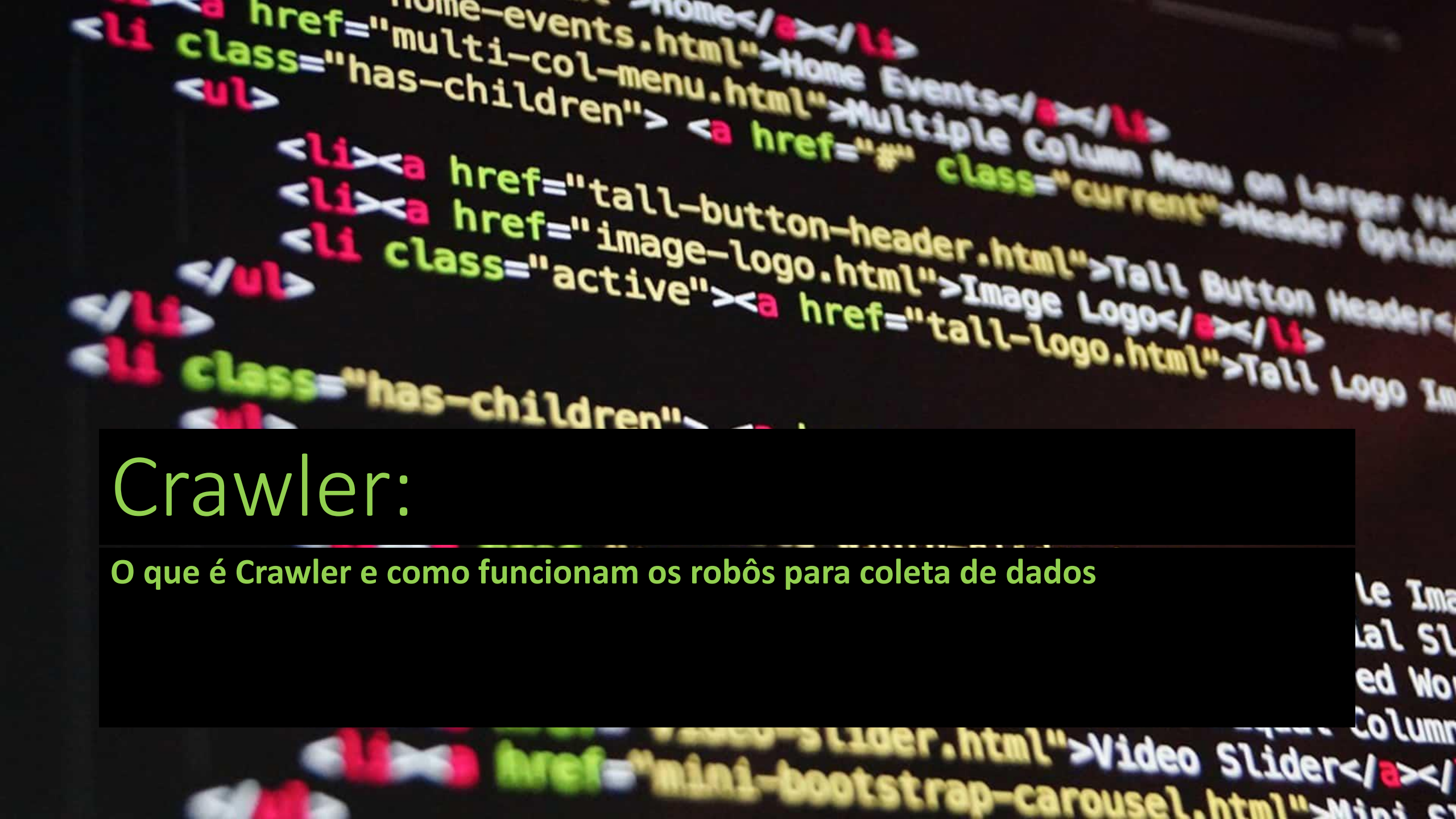
Classificação e Pesquisa de Dados

Cristiano Santos

cristiano.santos@amf.edu.br



Crawler:

O que é Crawler e como funcionam os robôs para coleta de dados

Crawler: Contexto dos dados

- Em toda a história da humanidade, **nunca antes produzimos e compartilhamos tanta informação.**

- Na era **big data**, a **cada segundo, circulam milhões de dados em rede** que, em sua maioria, **encontram-se não estruturados**, isto é, sem uma lógica de organização.

Crawler: Contexto dos dados

- Por si só, **dados isolados apresentam pouca ou nenhuma relevância.**

Crawler: Contexto dos dados

- Por si só, **dados isolados apresentam pouca ou nenhuma relevância.**

- Para que possam **adquirir significado**, um dos principais e mais complexos desafios está relacionado à **pesquisa, organização e análise de dados em escala.**

Crawler: Contexto dos dados

- Por si só, **dados isolados apresentam pouca ou nenhuma relevância.**
- Para que possam **adquirir significado**, um dos principais e mais complexos desafios está relacionado à **pesquisa, organização e análise de dados em escala.**
- Nesse cenário, surgem os **crawlers, robôs automatizados para fazer uma varredura e são capazes de agregar, classificar e entregar dados já estruturados.**

Crawler: O que é?

- **Crawler** ou **web crawler** são termos comuns utilizados para designar os **algoritmos criados para a coleta de dados**, também conhecidos por **spider ou scraper**.

Crawler: O que é?

- **Crawler** ou **web crawler** são termos comuns utilizados para designar os **algoritmos criados para a coleta de dados**, também conhecidos por **spider ou scraper**.

- Em uma explicação resumida, **crawlers** são **robôs rastreadores ou bots** que cumprem a **função de realizar a varredura em sites** ou em bancos de dados digitais.

Crawler: O que é?

- Crawlers são **robôs automatizados** que **fazem a pesquisa e extração de grande volume de dados em tempo real.**

- Principal recurso para os motores de busca na internet, esse tipo de automação também pode ser aplicado a estratégias de **data analysis em empresas.**

Crawler: Exemplo

- Os **mecanismos utilizados por buscadores como o Google** são o **principal exemplo prático** de como funciona um crawler:

o algoritmo, por meio de bots**, faz a busca em tempo real de links na internet** e promove a varredura completa das páginas, **a fim de entregá-las nos resultados de pesquisa aos usuários**, desde que tenham relevância para o tema de interesse na busca.

Crawler: Exemplo

- Trata-se de uma operação completa de data mining ou mineração de dados

Crawler: Como funcionam?

O processo de funcionamento de um web crawler envolve várias etapas:

- **Inicialização com URLs sementes**: O crawler começa com uma lista inicial de URLs, chamadas de sementes.

Crawler: Como funcionam?

O processo de funcionamento de um web crawler envolve várias etapas:

- **Inicialização com URLs sementes**: O crawler começa com uma lista inicial de URLs, chamadas de sementes.
- **Requisição e Download**: O crawler **faz uma requisição HTTP** GET **para baixar o conteúdo** da URL.

Crawler: Como funcionam?

O processo de funcionamento de um web crawler envolve várias etapas:

- **Inicialização com URLs sementes**: O crawler começa com uma lista inicial de URLs, chamadas de sementes.
- **Requisição e Download**: O crawler **faz uma requisição HTTP** GET **para baixar o conteúdo** da URL.
- **Parsing de Conteúdo**: O **conteúdo da página é analisado** para **extrair texto,** links, dados e metadados.

Crawler: Como funcionam?

O processo de funcionamento de um web crawler envolve várias etapas:

- **Armazenamento de Dados**: As **informações extraídas são armazenadas** em um banco de dados **ou índice de busca**.

Crawler: Como funcionam?

O processo de funcionamento de um web crawler envolve várias etapas:

- **Armazenamento de Dados**: As **informações extraídas são armazenadas** em um banco de dados **ou índice de busca**.
- **Rastreamento de Links**: O crawler segue os links encontrados na página para descobrir novas URLs, adicionando-as à lista de URLs a serem visitadas.

Crawler: Como funcionam?

O processo de funcionamento de um web crawler envolve várias etapas:

- **Armazenamento de Dados**: As **informações extraídas são armazenadas** em um banco de dados **ou índice de busca**.
- **Rastreamento de Links**: O crawler segue os links encontrados na página para descobrir novas URLs, adicionando-as à lista de URLs a serem visitadas.
- **Repetição do Processo**: Este processo se repete para cada nova URL descoberta, permitindo que o crawler percorra toda a web.

Crawler: Aplicações

Inteligência de mercado: Monitoramento de preços e concorrentes.

Compliance (Conformidade): Coleta de certidões e informações legais.

Validação de veículos: Verificação cadastral de veículos.

Imobiliário: Mapeamento de **anúncios e preços** de imóveis.

Backoffice: Automatização de tarefas administrativas.

Crawler: Benefícios

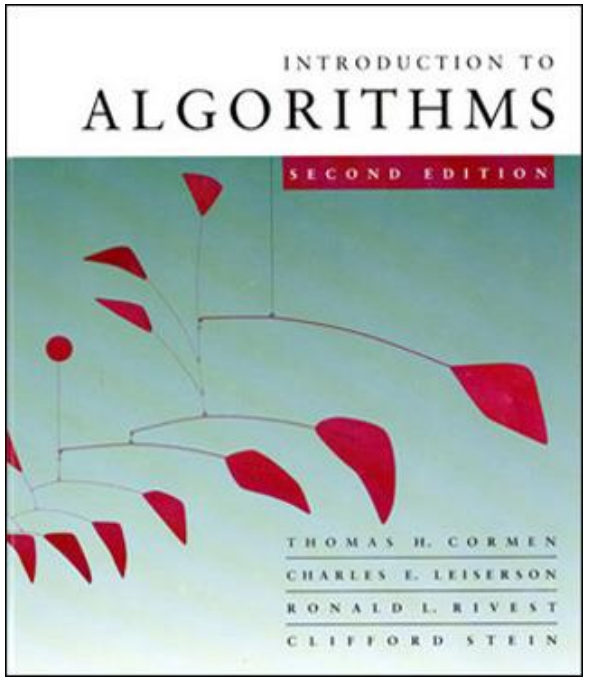
- **Produtividade:** Reduz tempo gasto em tarefas manuais.
- **Recursos:** Otimiza uso de recursos tecnológicos e humanos.
- **Custos:** Diminui custos operacionais.
- **Inteligência:** Melhora a qualidade dos dados para análise.
- **Decisões:** Proporciona dados em tempo real para decisões ágeis.

Bibliografia

- https://www.crawly.com.br/blog/o-que-e-crawler-robos-para-coleta-de-dados

Leitura importante

livro “Algorithms” de Cormen et al.



Classificação e Pesquisa de Dados

Cristiano Santos

cristiano.santos@amf.edu.br