Turning The Web Into Knowledge

Eliezer Carvalho
eliezer.carvalho.11.8a@gmail.com

Abstracto

A Web [1] foi concebida para armazenar, atualizar e localizar documentos de forma organizada e acessível, rapidamente tornou-se um elemento fulcral no desenvolvimento da Era da Informação [2] e é hoje a principal ferramenta de milhares de milhões de pessoas que interagem diariamente com o vasto universo digital. Com o crescimento cada vez mais acentuado da Web, aquilo que inicialmente servia apenas para partilhar dados evoluiu para o maior *Mouseion* [3] alguma vez criado. Cada página, cada *post*, cada comentário ou cada transação produz dados que podem ser recolhidos, analisados e transformados em *insights* valiosos. Este projeto procura entender todo o potencial do ***Web Scraping*** na sua plenitude, demonstrando que esta tecnologia é transversal a todas as áreas e que os dados quando bem utilizados, podem gerar múltiplas vantagens e novas formas de valor.

1. Introdução

Comumente, o acesso aos dados não é simples. A cada dia que passa, surgem mais restrições e barreiras ao acesso à informação verdadeiramente valiosa. É neste cenário que o *Web Scraping* ganhou - e continua a ganhar - relevância. O *Web Scraping* consiste na extração de dados disponíveis online, seja em páginas da internet, redes sociais ou praticamente qualquer espaço da vastidão da web. Trata-se de uma ferramenta poderosa que permite transformar informação dispersa em dados utilizáveis, possibilitando análises mais profundas e a criação de conhecimento estruturado. Este projeto nasceu com o propósito de criar uma ferramenta benéfica e poderosa que permitisse a recolha de grandes volumes de dados estatísticos desportivos de forma estruturada, simples e eficiente. Pretende-se também demonstrar que qualquer pessoa pode criar o seu próprio **óleo** [4] e manipular da forma que entender seja para análise, estudo ou criação de conhecimento.

2. Ciência de Dados

Na procura de conhecimento, os dados são valores discretos que fornecem informações, podendo descrever quantidades, qualidades, factos ou estatísticas, entre outros. Todavia, nem todos os dados transmitem valor nem estão prontos para serem utilizados [5]. Ter acesso a dados limpos e de fácil interpretação é uma competência muito valorizada no mundo atual, essa área chama-se Ciência de Dados. A Ciência de Dados [6] é uma área interdisciplinar que combina estatística e a ciência da computação para extrair valor dos dados, transformando-os em conhecimento que pode ser utilizado para tomar decisões. Esta ciência lida com dados de vários tipos, incluindo ruidosos, nebulosos, estruturados ou não estruturados. Trata-se, portanto, de uma área dedicada ao estudo e à análise organizada de dados científicos, de mercado, financeiros, sociais, geográficos, históricos, biológicos, psicológicos, entre muitos outros. O objetivo é a extração de conhecimento, a deteção de padrões ou a obtenção de informações úteis para a tomada de decisões. A Ciência de Dados enquanto campo existe há 30 anos mas ganhou mais destaque nos últimos anos devido a vários fatores, como o surgimento e a popularização de grandes bancos de dados e o desenvolvimento de áreas como *Machine Learning*.

Figura 1

3. Dados que mudaram o jogo

Em 2016, investigadores da *Escola de Medicina de Harvard* e do *Centro Médico Beth Israel Deaconess* demonstraram que modelos de *Machine Learning* podiam alcançar um desempenho bastante elevado na deteção do cancro da mama, aproximando-se da precisão dos radiologistas. Estudos posteriores confirmaram estes resultados, demonstrando que os modelos de *Machine Learning* podem ajudar a reduzir o número de mamografias com interpretações erradas e a apoiar decisões clínicas mais informadas [7]. Para atingir este nível de desempenho, os modelos tiveram de ser treinados com quantidades enormes de dados (*Big*

Data), o que ilustra claramente a importância da disponibilidade de dados de qualidade no desenvolvimento de tecnologias de deteção precoce e apoio ao diagnóstico. Nos tempos que correm, volumes de dados que outrora eram considerados enormes já não se enquadram na categoria de *Big Data*. Atualmente, conjuntos desse tamanho são usados até em contexto escolar para ensinar os princípios de *Machine Learning*, quase como material didático. Esta veracidade ilustra de forma clara a rapidez com que os dados evoluem em escala e complexidade ao longo do tempo.

Modernamente, um excelente exemplo são os *Large Language Models* (LLMs) [8] como o ChatGPT, o Gemini, o Grok ou o Mistral. Baseados na arquitetura *Transformer* [9], estes modelos são treinados com uma imensidão de quantidades de dados — frequentemente na ordem dos terabytes (1000000000000 de bytes) — durante a fase de pré-treino, de modo a alcançarem níveis de desempenho e compreensão excelentes. A maioria destes modelos é treinada com dados provenientes da Internet, a maior fonte de informação disponível. Neste processo, destaca-se a *Common Crawl* [10], uma organização sem fins lucrativos que 'varre' continuamente a Web, utilizando técnicas semelhantes às do web scraping para criar um repositório público gigante com dados provenientes da internet. Não é exagero dizer que o seu objetivo é criar uma Biblioteca de Alexandria [11]. Estes dados não podem ser inseridos diretamente num processo de pré-treino. É necessário um trabalho intensivo de limpeza, filtragem e deduplicação (alicerces do *Data Cleaning*), bem como etapas adicionais específicas para o treino de modelos de IA. Mais recentemente, a plataforma *Hugging Face* [12] lançou um conjunto de dados derivado da *Common Crawl*, denominado *FineWeb*, que é um conjunto de dados cuidadosamente filtrado, limpo e estruturado, concebido precisamente para pré-treinar LLMs de forma mais eficiente e com dados de maior qualidade. O nome "FineWeb" reflete essa ideia: uma versão refinada da web, pronta para alimentar modelos de grande escala.

Poderia dar muitos outros exemplos, mas estes dois já ilustram claramente o que é possível fazer com dados e mostram que esta revolução está apenas a começar.

4. Arquitetura do projeto

Sempre tive um profundo interesse pelo futebol, em especial pela sua vertente estatística e tática. Valorizo profundamente o trabalho das equipas técnicas, em particular o das equipas de observação e análise de jogadores que muitas das vezes são constituídas por mestres em Ciência de Dados. Foi exatamente este interesse que me persuadiu a querer analisar grandes volumes de dados de futebol. *Kaggle* [13], uma plataforma bastante conhecida por disponibilizar conjuntos de dados, existem inúmeros *datasets* estatísticos sobre futebol porém queria algo diferente: queria dados recolhidos por mim, recentes, fiáveis e, sobretudo, escaláveis. Com essa ideia em mente, comecei a estudar que bibliotecas *Python* poderia utilizar para trazer vida ao meu projeto. Uma delas era óbvia: o *Pandas* [14], a biblioteca mais famosa e versátil para a manipulação de dados. Após alguma pesquisa adicional, selecionei também o *Playwright* [15], uma biblioteca de automação web que me permitiu

navegar num sítio web sem intervenção humana, e a *Beautiful Soup* [16], que me facilitou a análise da estrutura DOM (*Document Object Model*) da página e me permitiria extrair informações de forma simples e eficaz. Com estas ferramentas, desenvolvi um primeiro ficheiro .py capaz de extrair dados automaticamente e de os armazenar num conjunto de dados de forma clara e organizada. Contudo, após concluir esta primeira fase, surgiu uma nova questão:

"E se eu já tiver dados de várias jornadas guardados e quiser agora adicionar apenas os mais recentes?"

Seria necessário repetir o processo de recolha de dados na totalidade? A resposta é simples: **não**.

Além de não ser prático, refazer toda a extração de dados seria um desperdício de recursos computacionais e assim nasceu o segundo ficheiro .py, concebido precisamente para detetar a última jornada existente no conjunto de dados e extrair apenas os dados posteriores, permitindo atualizar o conjunto de informação de forma eficiente e automática.

13-09-2025	West Ham	Tottenham	0	3	0.60
13-09-2025	Bournemouth	Brighton	2	1	1.45
13-09-2025	Crystal Palace	Sunderland	0	0	1.77
13-09-2025	Everton	Aston Villa	0	0	2.08
13-09-2025	Fulham	Leeds	1	0	0.85
13-09-2025	Newcastle	Wolves	1	0	1.56
13-09-2025	Arsenal	Nottingham	3	0	1.84
31-08-2025	Aston Villa	Crystal Palace	0	3	1.14
31-08-2025	Liverpool	Arsenal	1	0	0.53
31-08-2025	Brighton	Manchester City	2	1	2.29
31-08-2025	Nottingham	West Ham	0	3	0.73
30-08-2025	Leeds	Newcastle	0	0	0.69
30-08-2025	Manchester Utd	Burnley	3	2	3.63

Figura 2 - Algumas das estatísticas recolhidas

5. Término

Após concluir este projeto, consegui finalmente compreender o enorme potencial que se revela diante de nós na era da informação em que os dados são cada vez mais abundantes, acessíveis e de melhor qualidade. O presente documento, que acompanha o trabalho desenvolvido, tem como objetivo demonstrar esse potencial e servir de farol para todos aqueles que desejem dar os primeiros passos no mundo da Ciência de Dados, uma área essencial para quem pretenda ingressar nos domínios famosíssimos do Machine Learning ou da Inteligência Artificial.