Turning The Web Into Knowledge

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

Abstract

The Web [1] was designed to store, update and locate documents in an organised and accessible manner. It quickly became a key element in the development of the Information Age [2] and is now the main tool for billions of people who interact daily with the vast digital universe. With the ever-increasing growth of the Web, what initially served only to share data has evolved into the largest *Mouseion* [3] ever created. Every page, every post, every comment or every transaction produces data that can be collected, analysed and transformed into valuable insights. This project seeks to understand the full potential of Web Scraping, demonstrating that this technology is transversal to all areas and that data, when used well, can generate multiple advantages and new forms of value.

1. Introduction

Commonly, accessing data is not simple. With each passing day, more restrictions and barriers to accessing truly valuable information arise. It is in this scenario that Web Scraping has gained - and continues to gain - relevance. Web scraping consists of extracting data available online, whether on web pages, social networks, or virtually any space on the vast web. It is a powerful tool that allows you to transform scattered information into usable data, enabling deeper analysis and the creation of structured knowledge. This project was born with the purpose of creating a beneficial and powerful tool that would allow the collection of large volumes of sports statistics in a structured, simple, and efficient way. It also aims to demonstrate that anyone can create their own **oil** [4] and manipulate it as they see fit, whether for analysis, study, or knowledge creation.

2. Data Scientist

In the pursuit of knowledge, data are discrete values that provide information and can describe quantities, qualities, facts, or statistics, among other things. However, not all data convey value or are ready to be used [5]. Having access to clean, easily interpretable data is a highly valued skill in today's world, and this field is called Data Science. Data Science [6] is an interdisciplinary field that combines statistics and computer science to extract value from data, transforming it into knowledge that can be used to make decisions. This science deals with various types of data, including noisy, nebulous, structured, and unstructured data. It is, therefore, an area dedicated to the study and organised analysis of scientific, market, financial, social, geographical, historical, biological, psychological data, among many others. The goal is to extract knowledge, detect patterns, or obtain useful information for decision-making. Data Science as a field has existed for 30 years but has gained more prominence in recent years due to several factors, such as the emergence and popularisation of large databases and the development of areas such as Machine Learning.

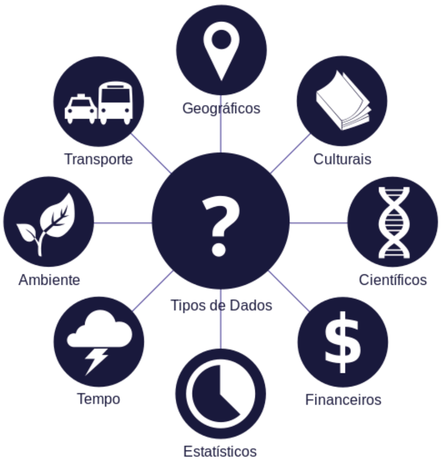


Figure 1

3. Data that changed the game

In 2016, researchers at Harvard Medical School and Beth Israel Deaconess Medical Centre demonstrated that Machine Learning models could achieve very high performance in detecting breast cancer, approaching the accuracy of radiologists. Subsequent studies confirmed these results, demonstrating that Machine Learning models can help reduce the number of misinterpreted mammograms and support more informed clinical decisions [7]. To achieve this level of performance, the models had to be trained with enormous amounts of data (Big Data), which clearly illustrates the importance of the availability of quality data in the development of early detection and diagnostic support technologies. Nowadays, data volumes that were once considered enormous no longer fall into the Big Data category. Currently, sets of this size are even used in schools to teach the principles of Machine Learning, almost as teaching material. This clearly illustrates how quickly data evolves in scale and complexity over time.

A modern example of this are Large Language Models (LLMs) [8] such as ChatGPT, Gemini, Grok, and Mistral. Based on the Transformer architecture [9], these models are trained with vast amounts of data — often in the order of terabytes (1000000000000 bytes) — during the pre-training phase in order to achieve excellent levels of performance and comprehension. Most of these models are trained with data from the Internet, the largest source of information available. In this process, Common Crawl [10] stands out, a non-profit organisation that continuously “scans” the Web, using techniques similar to web scraping to create a giant public repository of data from the Internet. It is no exaggeration to say that its goal is to create a Library of Alexandria [11]. This data cannot be directly inserted into a pre-training process. Intensive cleaning, filtering, and deduplication (the foundations of data cleaning) are required, as well as additional steps specific to training AI models. More recently, the Hugging Face platform [12] launched a dataset derived from Common Crawl, called FineWeb, which is a carefully filtered, cleaned and structured dataset designed precisely to pre-train LLMs more efficiently and with higher quality data. The name ‘FineWeb’ reflects this idea: a refined version of the web, ready to feed large-scale models.

I could give many other examples, but these two clearly illustrate what can be done with data and show that this revolution is only just beginning.

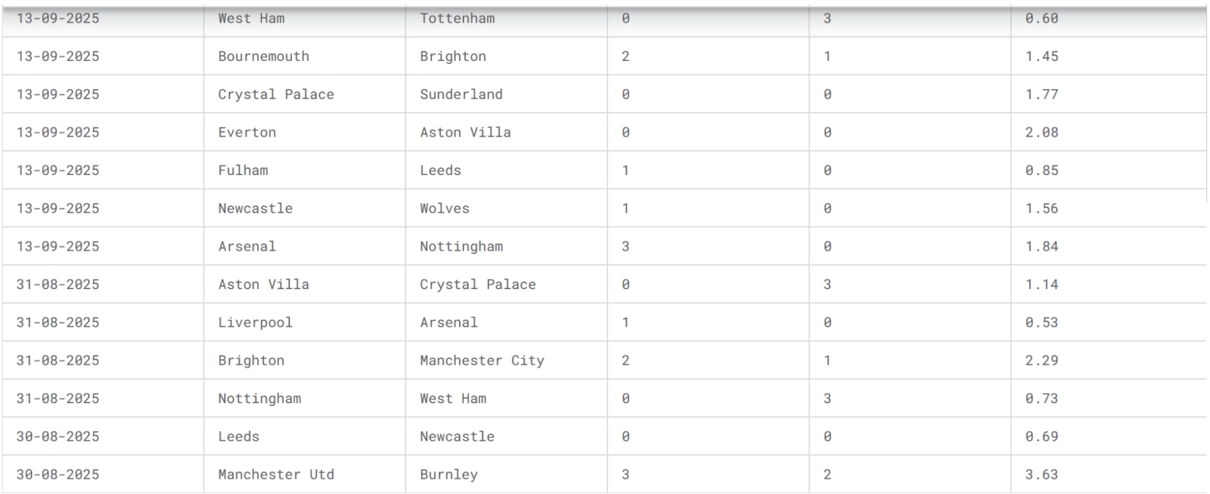
4. Project architecture

I have always had a deep interest in football, especially its statistical and tactical aspects. I greatly value the work of technical teams, particularly those involved in player observation and analysis, which often consist of masters in Data Science. It was precisely this interest that persuaded me to want to analyse large volumes of football data. Kaggle [13], a platform well known for providing data sets, has numerous statistical data sets on football, but I wanted something different: I wanted data collected by me, recent, reliable and, above all, scalable. With this idea in mind, I began to study which Python libraries I could use to bring my project to life. One of them was obvious: *Pandas* [14], the most famous and versatile library for data manipulation. After some additional research, I also selected *Playwright* [15], a web automation library that allowed me to navigate a website without human intervention, and Beautiful Soup [16], which made it easier for me to analyse the DOM (Document Object Model) structure of the page and allowed me to extract information simply and effectively. With these tools, I developed a first .py file capable of automatically extracting data and storing it in a clear and organised dataset. However, after completing this first phase, a new question arose:

‘What if I already have data from several journeys saved and now want to add only the most recent ones?’

Would it be necessary to repeat the entire data collection process? The answer is simple: **no**.

In addition to being impractical, redoing the entire data extraction would be a waste of computational resources. Thus, the second .py file was created, designed precisely to detect the last journey in the dataset and extract only the subsequent data, allowing the information set to be updated efficiently and automatically.



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

Figure 2 - Some of the statistics

5. Term

After completing this project, I finally understood the enormous potential that lies before us in the information age, where data is increasingly abundant, accessible, and of better quality. This document, which accompanies the work developed, aims to demonstrate this potential and serve as a beacon for all those who wish to take their first steps in the world of Data Science, an essential area for anyone wishing to enter the highly renowned fields of Machine Learning or Artificial Intelligence.