House Price Prediction Project

1. What does web scraping mean?

Nowadays, access to information represents an essential source for innovation and development. However, the necessary knowledge is not always easy to find or accessible. To facilitate this, the technique of web scraping comes into play, an efficient method for extracting data from web pages, providing significant resources for analysis, market studies, or process automation.

Web scraping is an efficient method for extracting data from web pages and transforming it into a structured format. This technique opens a gateway to a vast world of information available online, allowing users to automate the data collection process and integrate it into their applications or analyses.

We will explore the fundamental concepts of web scraping, associated technologies, and best practices to ensure efficient and ethical data collection.

1. How does a web scraper work?

Initially, the scraper will receive one or more URLs to download (using a request) locally before starting the scraping process. It will load the HTML page or, if more advanced, even JavaScript and CSS files. Then, the scraper will select useful information found on the HTML page using tokens (special characters/words/character combinations that delimit the desired information). After completing this process, the scraper saves the selected information in a file, usually CSV, but there are various options such as text or even JSON.

1. House Price Prediction Project

This project involves building and using a web scraper to collect the necessary data for creating a file containing information about houses currently on the market: their size, land size, number of rooms, number of bathrooms, utilities, proximity to parks and transportation, etc. Thus, a valuable dataset is formed for predicting house prices based on their information and amenities.

The data provided by the scraper is fed into a regression model based on machine learning (it is supervised machine learning, as our dataset contains pre-set labels and can provide feedback to the model, having the house price within the dataset). Thus, the model is trained on concrete data, ensuring the optimization of the prediction algorithm.

1. Possible resources for use

The preferred programming language for such projects is Python because it contains a multitude of useful libraries for both web scraping and machine learning.

For web scraping, the most useful libraries are Beautiful Soup, Scrapy, Pandas, and Selenium.

Beautiful Soup and Scrapy are used for the efficient extraction of information from HTML and XML files, preserving them according to user preferences.

At the same time, Pandas is very useful for data visualization, analysis, and manipulation. It provides structures and operations for the numerical manipulation of data tables like Excel or CSV.

Additionally, another commonly used package in a Python web scraper is re. It provides support for regular expressions, commonly encountered patterns for extracting desired data from a site.

For machine learning, useful libraries include NumPy, Scikit-Learn, and TensorFlow.

NumPy is popular for processing vectors and multidimensional matrices because it can be used to perform a wide variety of mathematical operations. Its ability to handle linear algebra, Fourier transformation, and many others makes NumPy ideal for machine learning and AI projects, allowing users to easily manipulate matrices to effortlessly improve the performance of machine learning algorithms. NumPy is faster and easier to use than most other Python libraries.

Scikit Learn is a very popular machine learning library built on NumPy and SciPy. It contains most classic supervised and unsupervised machine learning algorithms and can be used for data extraction, modeling, and analysis. The simple design of Scikit Learn offers a user-friendly library for beginners in machine learning.

The open-source TensorFlow library specializes in what is called differentiable programming, which means it can automatically calculate the derivatives of a function in a high-level language. Both machine learning and deep learning models are easily developed and evaluated with the flexible architecture and framework of TensorFlow. TensorFlow can be used to visualize machine learning models on both desktop and mobile platforms.