datacamp

COURSE GLOSSARY

Intermediate Importing Data in Python

API (Application Programming Interface): A set of rules and endpoints that allow one software application to request services or data from another in a standardized way

BeautifulSoup: A Python library for parsing HTML and XML that converts messy "tag soup" into a navigable parse tree and provides methods to extract elements like titles, links, and text

CSV (Comma-Separated Values): A plain-text file format where each line represents a record and fields are separated by commas, commonly used for tabular data exchange

GET request: An HTTP method used to request and retrieve data from a specified resource on a server without causing side effects on the server

HTML (Hypertext Markup Language): The standard markup language for documents designed to be displayed in web browsers, composed of tags that define structure and content

HTTP (HyperText Transfer Protocol): The standard application-level protocol used for transferring resources (like web pages) across the World Wide Web

HTTPResponse: An object returned by lower-level URL fetching functions (e.g., urllib.urlopen) that represents the server's response and provides methods such as read() to access the response body

OAuth: An authorization protocol that enables applications to obtain limited access to user accounts on an HTTP service by using access tokens rather than sharing credentials

pandas DataFrame: A two-dimensional, labeled data structure in Python's pandas library that stores tabular data with rows and columns and supports powerful data manipulation operations

Parsing (HTML parsing): The act of analyzing raw HTML and converting it into a structured representation so that specific elements or data can be programmatically extracted

Query string: The portion of a URL that follows a question mark (?) and encodes parameter names and values used to pass data or filter requests to a web service or API

requests (Python package): A popular third-party Python library that simplifies making HTTP requests and returns Response objects with convenient methods like .text and .json()

Response object (requests): The object returned by requests methods (e.g., requests.get) that encapsulates server response data and provides attributes and methods like .status_code, .text, and .json()

REST (Representational State Transfer): An architectural style for designing networked applications that uses stateless operations over HTTP (typically via standard verbs like GET, POST, PUT, DELETE)

Structured data: Data that follows a predefined model or schema (e.g., tables, CSVs, or JSON objects), making it straightforward to query and analyze

tweepy: A Python library that provides a convenient interface for accessing Twitter's APIs, including streaming and REST endpoints for reading and writing tweets

Twitter Streaming API: A Twitter endpoint that delivers a continuous real-time stream of public tweets and other data, which can be filtered by keywords, users, or locations

Unstructured data: Data that lacks a predefined model or organization (e.g., free text in HTML) and therefore often requires parsing or transformation to become useful

URL (Uniform Resource Locator): A string that specifies the address of a resource on the internet, typically including a protocol (like http or https) and a resource name or path

urllib: A Python standard-library package that provides high-level interfaces for fetching data across the web, including functions like urlopen and urlretrieve

urlretrieve: A utility function (from urllib.request) that makes an HTTP GET request to a given URL and saves the retrieved content directly to a local file

Web scraping: The automated process of retrieving web pages and extracting useful information from their HTML or other content formats