

Chapter 2

Covid Live Multi-Threaded Live COVID 19 Data Scraper

Parth Birthare

Vellore Institute of Technology, India

Maheswari Raja

Vellore Institute of Technology, India

Ganesan Ramachandran

Vellore Institute of Technology, India

Carol Hargreaves

 <https://orcid.org/0000-0002-5522-4058>
National University of Singapore, Singapore

Shreya Birthare

Vellore Institute of Technology, India, India

ABSTRACT

COVID-19 began in 2019, and by the advent of 2020, it had become widespread and adversely affected the world. In this work—Covid Live, COVID-19 data is scraped from an online website, which gives an overview of the status of the pandemic in the desired format. The authors built an application interface using a Python micro web Flask framework. The data scraping uses a multi-threading concept to reduce the program's runtime error, resulting in receiving the data quickly, and faster than existing web crawlers and scrapers. This paper focuses on dealing with storing scraped data in the desired format. It also provides options to hear the audio of the scraped data and to download the scraped data. The authors present visualizations of current trends with scraping period details and demonstrate an efficient application that does the data scraping quickly and efficiently.

DOI: 10.4018/978-1-6684-6523-3.ch002

INTRODUCTION

The COVID-19 pandemic has infected many people (Yang et al., 2020). Furthermore, in attempts to slow down the spread of the novel virus, nations imposed complete and partial lockdowns hoping to prevent further spread of the virus (Atalan, 2020). Many lives have been affected by the pandemic. Scientists and researchers are constantly analyzing the impact and devising methods to purge the spread of the virus further and predict possible variants. Therefore, positive case patterns, environmental and biological factors, and policies are crucial for COVID-19 research (Sha et al., 2021). Therefore, there should be a database for maintaining records or spatiotemporal COVID-19 records that countries publish from virus testing after the advent of 2020. As per the public data, most data come from a few international agencies, such as WHO, GHC, or CDC. The subcommittees within the organization make the data public after collecting information and producing the dataset (COVID Data Tracker, 2021).

Also, it seems that data is the new differentiation. It is the focus of market research and business strategy. Whether one has to start a new work or shake out a new strategy for an existing business, one needs to access and analyze a large amount of data for better results. Web scraping plays a role in easing up the process, and in the present scenario, large institutions are also collaborating to provide facilities to track COVID-19 in real-time. For example, Johns Hopkins University developed a COVID-19 dashboard that is regularly updated by extracting data from around eight non-government sources and publicly providing it as a single dataset (Dong, Du, & Gardner, 2020). 1Point3Acres provides a similar service, which aims to be transparent to the public about COVID-19 cases (1point3acres Global COVID-19 Tracker & Interactive Charts, 2021).

Web scraping is the process of mining data, usually unstructured data from any number of sources, efficiently and faster and storing it as structured data for further analysis (Sirisuriya, 2015). It is effortless because it does not involve visiting the web pages to copy-paste the extensive data. Data extraction can be done from any website, anywhere, no matter how large and complex the data is. Moreover, some websites may have the type of data that cannot be copied and pasted directly. For example, it can be in the form of CSV, image, or text. Web scraping can be copying, grabbing, pasting text, or parsing HTML (Sirisuriya, 2015). Further, many tools and methods are available for web scraping (Persson, Evaluating tools and techniques for web scraping, 2019; Saurkar, Pathare, & Gode, 2018). Not only in python but also in other programming languages (Easily harvest (scrape) web pages, n.d.).

27 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/covid-live-multi-threaded-live-covid-19-data-scraper/318550

Related Content

Fuzzy-based Gain Adaptive Scheme for Set-Point Modulated Model Reference Adaptive Controller

A. K. Pal, Indrajit Naskar and Sampa Paul (2018). *International Journal of Natural Computing Research* (pp. 1-19).

www.irma-international.org/article/fuzzy-based-gain-adaptive-scheme-for-set-point-modulated-model-reference-adaptive-controller/217020

Multi-Agent Systems Research and Social Science Theory Building

H. Verhagen (2007). *Handbook of Research on Nature-Inspired Computing for Economics and Management* (pp. 101-110).

www.irma-international.org/chapter/multi-agent-systems-research-social/21123

On Quasi Discrete Topological Spaces in Information Systems

Tutut Herawan (2012). *International Journal of Artificial Life Research* (pp. 38-52).

www.irma-international.org/article/quasi-discrete-topological-spaces-information/74335

Identifying Subtypes of Cancer Using Genomic Data by Applying Data Mining Techniques

Tejal Upadhyay and Samir Patel (2019). *International Journal of Natural Computing Research* (pp. 55-64).

www.irma-international.org/article/identifying-subtypes-of-cancer-using-genomic-data-by-applying-data-mining-techniques/231573

Quantum Automata with Open Time Evolution

Mika Hirvensalo (2012). *Nature-Inspired Computing Design, Development, and Applications* (pp. 74-89).

www.irma-international.org/chapter/quantum-automata-open-time-evolution1/66771