

Web scraping with Python project

The main objective of the project is ETL: Extract, Transform, Load

ETL (extract, transform, load) is the “general procedure of copying data from one or more sources into a destination system which represents the data differently from the source” ([Wikipedia](#)). That’s just a fancy way to say that ETL is the process of taking data from one place, massaging it a little, and saving it in another place.

Web scraping is one form of ETL: you *extract* data from a website, *transform* it to fit the format you want, and *load* it into a CSV file.

Furthermore, ETL is an essential process for businesses and organizations that rely on data analysis to make informed decisions.

The ETL process typically consists of the following stages:

1. **Extraction:** This stage involves identifying and extracting data from various sources, such as databases, flat files, web services, and APIs. The extracted data is then moved to a staging area for further processing.
2. **Transformation:** This stage involves cleaning, filtering, validating, and transforming the data into a format that is suitable for analysis. Common data transformation tasks include data mapping, aggregation, normalization, and enrichment. The transformed data is then loaded into a target database or data warehouse.
3. **Loading:** This stage involves loading the transformed data into the target database or data warehouse. The loading process can be done using various techniques, such as batch processing, real-time streaming, or incremental updates. The loaded data is then ready for analysis and reporting.

Overall, the ETL process helps organizations to consolidate and integrate data from multiple sources, ensuring data consistency, accuracy, and completeness. It enables analysts, data scientists, and decision-makers to access, analyze, and visualize data in a meaningful way, leading to better insights and outcomes.

The project will scrape data from the UK’s government official website: <https://www.gov.uk/search/news-and-communications> .

Web scraping with Python project

The main objective is to extract, transform and load all titles of the articles and their descriptions from the UK's Government news and communication page.