Country Heads Web Scraping

This project is dedicated to extracting and organizing comprehensive data about country heads and key government officials across various nations. The data was gathered using web scraping techniques in Python and has been structured to ensure accuracy and consistency.

Below is a summary of the fields included in the final dataset.

Fields in the final dataset

Field	Description
ISO2CODE	The ISO 3166-1 alpha-2 code for the country
Country_Head_Name_EN	The name of the country in English
Country_Head_Name_AR	The name of the country in Arabic
Designation_EN	The official title or position (e.g., President, King, Interior Minister, Defense Minister) in English.
Person_Name_EN	The name of the individual holding the position in English
Person_Name_AR	The name of the individual holding the position in Arabic.
Assumed_Office_Date	The date on which the individual assumed office.
Image_URL	The URL linking to the image of the individual
ID	A unique identifier for each country
Designation_ID	A unique identifier for each official title or position
Person_ID	A unique identifier for each individual holding a position

Notes:

- Some of these fields were collected from Wikipedia pages, while others were created as unique identifiers to ensure data integrity.
- Not all Arabic names are included due to the absence of Arabic versions on Wikipedia for some entries.

Libraries Used

This project utilizes several Python libraries to handle various tasks such as web scraping, data manipulation, and URL encoding. Below is a brief overview of each library and its role in the project:

Library	Description
urllib.parse	Specifically, the `quote` function is used to encode URLs by converting special characters into percent-encoded formats, ensuring that the URLs are properly formatted for web requests.
Pandas (`**pd**`)	A powerful data manipulation library used to create and manage DataFrames. It plays a crucial role in organizing, analyzing, and exporting the collected data.
re	The regular expression library is used for string manipulation, allowing for pattern matching and text extraction tasks essential for cleaning and processing data.
requests	A simple and elegant HTTP library for making web requests. It is used to send requests to web pages and retrieve the HTML content needed for data extraction.
BeautifulSoup from `**bs4**`	A library for parsing HTML and XML documents. It is used to navigate and extract data from the HTML content retrieved by `requests`, making the web scraping process more manageable.

Main Functions Used

This section outlines the key functions used in the project, including their inputs and outputs. All URLs used are from Wikipedia:

Function	Input	Output
get_country_data	A Wikipedia URL containing country ISO codes.	Retrieves the country's ISO2CODE, the name of the country in English (`Country_Head_Name_EN`), and the URL for the country (`Country_URL`).
get_arabic_wikipedia_link	The English Wikipedia link for a person or a country.	Obtains the corresponding Arabic Wikipedia link.
get_country_name_arabic	The Arabic Wikipedia URL for a person or country.	Retrieves the Arabic name of the country or person.
fetch_government_data	The Wikipedia English URL for the country.	Collects government-related data, including the official title in English (`Designation_EN`) and the name of the person in English (`Person_Name_EN`).
get_person_data_main	The Wikipedia URL for the person’s page.	Retrieves the assumed office date (`Assumed_Office_Date`) and the image URL (`Image_URL`) for the person.
extract_defense_ministers_from_table	The Wikipedia link for defense ministers and the index of the relevant table.	Extracts data about defense ministers, including the country name (`State`), the name of the defense minister (`Defense_Minister_Name`), and the URL for the defense minister page (`Defense_Minister_url`).
get_person_minister_data	The Wikipedia URL for the defense/interior minister’s page.	Retrieves the image link for the defense/interior minister.
extract_interior_ministers_from_table	The Wikipedia link for interior ministers and the index of the relevant table.	Extracts data about interior ministers, including the country name (`State`), the name of the interior minister (`Interior_Minister_Name`), and the URL for the interior minister (`Interior_Minister_url`).

Transformations Functions Used

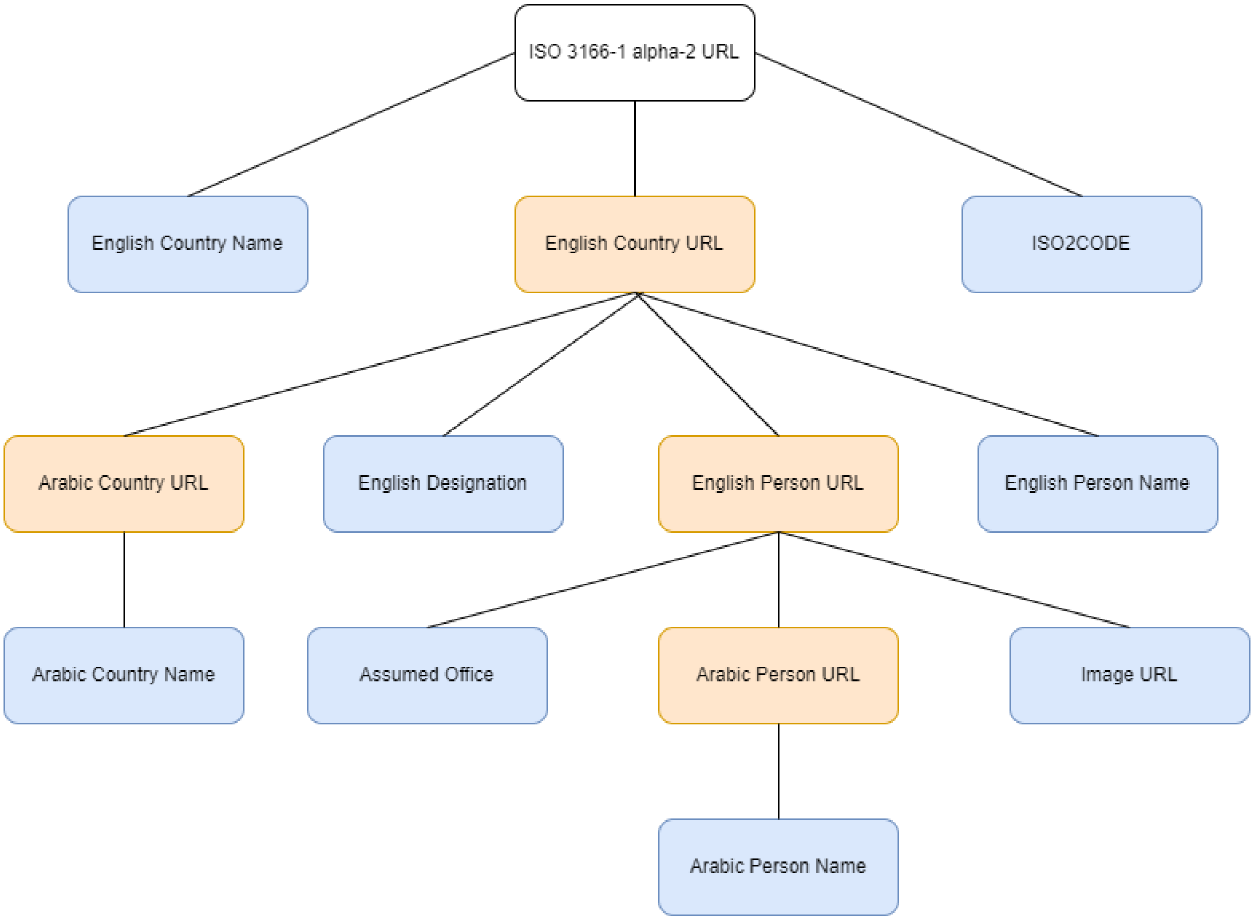
This section describes the auxiliary functions used in the project, which assist in data processing and manipulation:

Function	Purpose
assign_country_url	Updates certain country URLs in the dataset as required by the team.
expand_data	Appends the `Person_Name_EN` list and `Person_URL` list to the DataFrame, expanding it with additional rows for each person.
assign_designation_id	Adds a `Designation_ID` column to the DataFrame, assigning unique identifiers to each official designation.
has_date	Checks if the `Assumed_Office_Date` is a valid date, ensuring that the date data is properly formatted and accurate.
aggregate_designations	Merges duplicated names into a single row. For example, if a person holds multiple titles (e.g., "President: X, Defense Minister: X"), they will be combined into one record (e.g., "President/Defense Minister: X").
add_defense_ministers	Iterates over the defense ministers' data and appends matching records to the DataFrame. Specifically: It loops through the data of defense ministers and checks for any matches between the `State` (from the first DataFrame) and the `Country_Head_Name_EN` (from the second DataFrame). If a match is found, the corresponding defense minister data is appended to the new DataFrame. The second DataFrame is then combined with the appended data from the new. The combined DataFrame is sorted by `ID` and `Designation_ID`, and finally saved to an Excel sheet.

Workflow Diagram

The diagram illustrates the workflow of the project, depicting how data is processed and integrated. Each step is represented by a node, with arrows showing the flow of data between them.

- Blue: Indicates fields that are included in the final data.
- Orange: Represents fields that are excluded from the final data.



Deployment and Schedule

- The project is scheduled to run daily, automating data extraction and updates.
- The ETL team is responsible for maintaining and executing the script regularly.
- The script is set to update the data daily to keep the dataset current.
- The ETL team manages the updates and ensures the data remains accurate and consistent.