Project authors

Agnieszka Dogiel
Album No. 411947

Oksana Chubatenko
Album No. 456299

Natalia Roszczypała
Album No. 387477

Project description

The project was designed to conduct web scraping from the CIA World Factbook website, specifically targeting countries in Africa and Europe with Beautiful Soup, Scrapy and Selenium scrapers. The purpose was to extract relevant information, store it in a structured format, namely a CSV file, and compare the scrapers performance.

The chosen website is accessible at: https://www.cia.gov/the-world-factbook/?fbclid=IwAR0kAuWnnhqHmTQsIvdBbOs6DDCYva4AFVIExbSsnmTSdK27boinqNbnZIo (Accessed on June 10, 2023).
Description provided by website:

The World Factbook provides basic intelligence on the history, people, government, economy, energy, geography, environment, communications, transportation, military, terrorism, and transnational issues for 266 world entities.

The information obtained through web scraping included demographic and economic data. By gathering these data points, we aimed to perform a simple analysis of the collected information. In our analysis, we examined various factors such as population, age structure, population growth, birth and death rate, life expectancy, sex ratio, GDP, inflation rate, and unemployment rate for all of the countries on selected continents. By organizing and analyzing this data, we aimed to gain insights into the socioeconomic characteristics of the countries in question, which are further presented in this report.

Description of scraper mechanism

Each of the scrapers we used is performing following steps:

1. The code is organized in methods for scraping information and storing it in a CSV file.
2. The code begins by importing the necessary modules.
3. We provide starting URL's for two continents, namely Africa and Europe, and for each of this website the scraper is looking for countries links or countries names in case of Selenium.
4. Scraper is looping through provided country links (or country names) and for every country it searches chosen information and collects it.
5. Information the scraper is looking for is defined in the code. Each factor we want to scrap has its separate lines of code. We specify text to look for in HTML to find the data and specify the format of scraped information to have clean results. We also added the condition that in case of missing value on the website the code leaves this space blank. Scrapped information for each country is saved in a data frame, which is appended after each loop iteration.
6. All the collected data is yielded as a dictionary, representing a single row of data. The keys of the dictionary correspond to the names of scraped variables, and the values contain the extracted information.
7. Yielded information is then saved to the CSV file by command or function appropriate for the scraper.
8. In every code we added boolean parameter at the beginning of the code which limits the number of scraped countries websites to 100 if == True or scrapes all information if == False
9. For every scraper we estimated the running time to be able to compare the performance. The total scraping time is printed in the terminal.

Description of output

We generate the output in a CSV file containing rows with column names (which can be used as a header) and rows with data. Each row represents a country and each column represents a factor. First column contains country name and following columns are: Population, Total Median Age, Male Median Age, Female Median Age, Age 0-14, Age

15-64, Age 65 and more, Population growth rate, Birth rate, Death rate, Total life expectancy, Male life expectancy, Female life expectancy, Total sex ratio, GDP 2021 (bn), GDP 2020 (bn), GDP 2019 (bn), GDP growth rate 2021 (%), GDP growth rate 2020 (%), GDP growth rate 2019 (%), GDP per capita 2021 (bn), GDP per capita 2020 (bn), GDP per capita 2019 (bn), Inflation rate 2021 (%), Inflation rate 2020 (%), Inflation rate 2019 (%), Unemployment rate 2021 (%), Unemployment rate 2020 (%), Unemployment rate 2019 (%). In total we get results for 110 countries, which can be limited to 100 using the boolean parameter.

Data analysis

Based on the scraped information, some basic statistics were computed separately for African and European countries. The analysis proved that the average population in an African country is significantly higher than in European country, i.e., 25,410,591.31 and 11,059,263.42, respectively, with the maximum population value equal to 230,842,743 in Africa and 84,220,184 in Europe.. Furthermore, the average population growth rate for Africa equals 2.19% whereas this value in Europe equals only 0.09%. This relation is also reflected in the average birth and death rates which for Africa are equal to 28.82 and 7.27 per 1,000 population, and to 8.97 and 9.38 per 1,000 population in Europe. Not only this data shows that population value is increasing significantly faster in African countries, but also that in Europe this increase is negative. Nevertheless, the average life expectancy in Africa is slightly lower than in Europe, i.e., 66.62 versus 72.98, with relatively higher life expectancy for women on both continents.

In regards to the age structure, women are dominating on both continents, especially in Europe where the ratio of men to women is equal to 0.88. The value of the respective ratio in Africa equals 0.99 indicating a more balanced population. The median age for an African country amounts to 21.51, with the highest value being equal to 43.2. In Europe, the median age for the average country is equal to 38.25, with the highest median age being 55.4 indicating a significantly older population than on the African continent.

The economic conditions are also relatively different on both continents. In Africa, the average GDP value in 2021 was equal to USD 76.19 billion with a maximum of USD 790.63 billion, whereas these values in Europe amounted to USD 158.50 billion and USD 992.68 billion, respectively. Nevertheless, the average GDP growth rate in 2021 was only slightly higher for Europe, i.e., 4.36% for Africa and 5.43% for Europe. In regards to inflation and unemployment, these values are significantly higher in African countries. In

2021, the average inflation rate was equal to 6.35% and the average unemployment rate was equal to 10.04% whereas these rates in Europe achieved values of 2.41% and 5.68%.

Scrapers performance

The execution time for the Beautiful Soup scraper is equal to 74 seconds, for the Scrapy scraper the running time is 5 seconds, whereas for the Selenium scraper this value amounts to 883 seconds. Therefore, the most efficient is the scraper written with the use of the Scrapy package.

Division of work

Oksana - Beautiful Soup scraper
Agnieszka - Scrapy scraper
Natalia - Selenium scraper