Report on Web Scraping of News Headlines from BBC News Website

Introduction:

Web scraping is the process of extracting data from websites and is commonly used in data science and analytics to gather information for research or analysis purposes.

Methodology:

1. Used Python requests library to send a request to the website and retrieve the HTML content of the page.
2. Used BeautifulSoup library to parse the HTML and extract the news headline data from the webpage.
3. Extracted the data by finding the HTML tags that contained the headlines and extracting the text within those tags.

Results:

After successfully extracting the news headline data, the cleaning of data was done to remove any unwanted characters or formatting, we removed newline characters and tabs from the headlines using string manipulation techniques in Python. We then stored the cleaned headlines in a Pandas dataframe and added a sequential numbering column to the dataframe. We then saved the cleaned data to a CSV file for further analysis. Additionally, we plotted a bar chart of the length of each headline to visualize the distribution of headline lengths in the data.

Challenges and Problems Faced:

During the web scraping process, we encountered several challenges and problems. Firstly, we faced issues with the HTML structure of the webpage, as the news headlines were located in different HTML tags on different parts of the page. We had to carefully examine the structure of the webpage to identify the correct HTML tags to extract the headline data.

We also encountered formatting issues with the extracted data, as there were newline characters and tabs present in the headlines. We had to use string manipulation techniques to clean the data and remove these unwanted characters.

Conclusion:

In conclusion, we successfully extracted news headline data from the BBC News website using Python web scraping techniques. We cleaned and saved the data to a CSV file for further analysis and visualization. The project presented several challenges and required careful examination of the HTML structure of the webpage and attention to data formatting issues. Overall, the project demonstrated the power of web scraping in gathering data for research and analysis purposes.