Project 4: The New Times Sports

You’ve just landed a role as a junior data analyst at a dynamic media company eager to unlock the stories hidden within sports news. Your mission: dive into the sports section of the New Times Rwanda website, scraping and extracting the rich text content of news articles.

Beyond simply gathering data, you’ll clean and preprocess this text, transforming chaotic information into clear, meaningful insights. Finally, you’ll bring the stories to life by visualizing key trends and patterns that reveal what’s capturing the nation’s sports attention.

Here is the workflow you should follow:

Part 1: Data Scraping

Step 1: Scrape Article Page Links

- Send a GET request to **https://www.newtimes.co.rw/sports**.
- Use **BeautifulSoup** to parse the HTML content.
- Extract all hyperlinks.
- **Filter** these links to keep only those containing the word **"sports"** in the URL.
- Store the filtered URLs in a list.

Step 2: Scrape Text Content from Each Link

- Loop through the list.
- Send a request to each URL.
- Use **BeautifulSoup** to extract all visible text content.
- Store the raw text from each page in another list.

Part 2: Text Cleaning & Preprocessing

Step 1: Clean the Text

For each text entry:

- Remove non-alphanumeric characters.
- Remove punctuation.
- Remove single characters surrounded by spaces.
- Replace multiple spaces with a single space.
- Convert text to lowercase.
- Store the cleaned text in a new list.

Step 2: Store Data in DataFrames

- Create df1 with a column **original_data** containing the raw text.
- Create df2 with a column **cleaned_data** containing the cleaned text.
- Optionally merge them into a combined dataframe df.

Part 3: Text Analysis & Visualization

With your data cleaned, it's time to **explore its distributions and relationships**. You are free to take this in any direction you deem fit. Here are possible things you can try:

- Concatenate all cleaned text into a single string and generate a word cloud of the most frequent words.
- Perform frequency distribution of top 20 words.
- See which two- or three-word phrases appear most often (like “premier league” or “world cup”) using a bigram/trigram.