Extracting Text Data from CBC News Using Selenium

In [ ]:
```
## Importing the libraires
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import Select
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import (
    StaleElementReferenceException, TimeoutException,
    ElementClickInterceptedException, ElementNotInteractableException
)
from webdriver_manager.chrome import ChromeDriverManager
import re
import time

service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service)

# keywords for searching
keywords = [
    "Job Market", "Employment Trends", "Labor Market", "Hiring Trends",
    "Job Growth", "Workforce Development", "Labor Shortage", "Job Openings",
    "Unemployment Rate", "Recession and Jobs", "Economic Impact on Employment",
    "Layoffs", "Downsizing", "Underemployment", "Workforce Diversity",
    "Inclusive Hiring", "Diversity in Tech Jobs", "Canadian Job Market"
]

# scraped data stored in the articles_data
articles_data = []

try:
    # Iterating through each keyword
    for keyword in keywords:
        # Going to CBC news
        driver.get("https://www.cbc.ca/news")
        time.sleep(2)

        # Locate and clicking  the search button
        search_button = WebDriverWait(driver, 10).until(
            EC.element_to_be_clickable((By.ID, "searchButton"))
        )
        search_button.click()
        time.sleep(1)

        # Locating the search input field, enter keyword, and then  press Enter
        search_bar = WebDriverWait(driver, 10).until(
            EC.visibility_of_element_located((By.CLASS_NAME, "searchInput"))
        )
```

```
        search_bar.clear()
        search_bar.send_keys(keyword)
        search_bar.send_keys(Keys.RETURN)
        time.sleep(3)

        #  sorting order by date to get the newest first
        dropdown = WebDriverWait(driver, 10).until(
            EC.element_to_be_clickable((By.ID, "sortOrderSelect"))
        )
        select = Select(dropdown)
        select.select_by_value("date")
        time.sleep(2)

        # stop_searching flag
        stop_searching = False

        while True:
            # all articles in the search results
            articles = driver.find_elements(By.CLASS_NAME, "contentWrapper")

            for article in articles:
                try:
                    # Extracting the date text
                    date_text = article.find_element(By.CLASS_NAME, "timeStamp").te

                    # Checking if the article date is from 2024 or newer
                    year_match = re.search(r'\b(20\d{2})\b', date_text)
                    if year_match and int(year_match.group(0)) < 2024:
                        stop_searching = True
                        break
                    else:
                        # Extracting  the title and description
                        title = article.find_element(By.CLASS_NAME, "headline").tex
                        description = article.find_element(By.CLASS_NAME, "descript

                        # article data appended to the list as a dictionary
                        articles_data.append({
                            "Keyword": keyword,
                            "Title": title,
                            "Description": description,
                            "Date": date_text
                        })

                except Exception as e:
                    print(f"Could not extract data for an article with keyword '{ke

            # Stop loading more articles if older ones have been reached
            if stop_searching:
                print(f"Stopping search for keyword '{keyword}' as articles before
                break

            # loading  more articles
            try:
                load_more_button = WebDriverWait(driver, 15).until(
                    EC.element_to_be_clickable((By.XPATH, "//button[contains(text()
                )
```

```
                driver.execute_script("arguments[0].scrollIntoView(true);", load_mo
                time.sleep(1)

                # Clicking "Load More" button
                for attempt in range(3):  # Retry up to 3 times
                    try:
                        load_more_button.click()
                        print("Load More button found and clicked.")
                        time.sleep(3)  # time sleep added to  wait for new content
                        break
                    except ElementClickInterceptedException:
                        print("Element intercepted, retrying...")
                        time.sleep(1)  # Short delay before retrying

            except (StaleElementReferenceException, TimeoutException, ElementNotInt
                print(f"No more articles to load for keyword '{keyword}'.")
                break

finally:
    driver.quit()

# # Output the gathered data
# for article in articles_data:
#     print(article)
```

```
Element intercepted, retrying...
Element intercepted, retrying...
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Stopping search for keyword 'Job Market' as articles before 2024 are reached.
Stopping search for keyword 'Employment Trends' as articles before 2024 are reached.
Stopping search for keyword 'Labor Market' as articles before 2024 are reached.
Load More button found and clicked.
Load More button found and clicked.
Stopping search for keyword 'Hiring Trends' as articles before 2024 are reached.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
```

```
Stopping search for keyword 'Job Growth' as articles before 2024 are reached.
Stopping search for keyword 'Workforce Development' as articles before 2024 are reac
hed.
Stopping search for keyword 'Labor Shortage' as articles before 2024 are reached.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Stopping search for keyword 'Job Openings' as articles before 2024 are reached.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Stopping search for keyword 'Unemployment Rate' as articles before 2024 are reached.
Load More button found and clicked.
Stopping search for keyword 'Recession and Jobs' as articles before 2024 are reache
d.
Stopping search for keyword 'Economic Impact on Employment' as articles before 2024
are reached.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
```

```
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Stopping search for keyword 'Layoffs' as articles before 2024 are reached.
Stopping search for keyword 'Downsizing' as articles before 2024 are reached.
Stopping search for keyword 'Underemployment' as articles before 2024 are reached.
Stopping search for keyword 'Workforce Diversity' as articles before 2024 are reache
d.
Stopping search for keyword 'Inclusive Hiring' as articles before 2024 are reached.
Stopping search for keyword 'Diversity in Tech Jobs' as articles before 2024 are rea
ched.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Load More button found and clicked.
Stopping search for keyword 'Canadian Job Market' as articles before 2024 are reache
d.
```

In [8]:
```
df = pd.DataFrame(articles_data)
df.shape
```

Out[8]: (24780, 4)

In [ ]:
```
df.to_excel("data.xlsx",index= False)
```