Web Scraper with Sentiment Analysis

This project is a C++ web scraper that fetches search results from DuckDuckGo and performs sentiment analysis on the content of each website. It uses libcurl for web requests, libxml2 for HTML parsing, cpr for HTTP requests, and jsoncpp for JSON handling.

Prerequisites

Before you can build and run the web scraper, you need to have the following libraries and dependencies installed on your system:

Ubuntu

```
sudo apt update
sudo apt install libcurl4-openssl-dev libxml2-dev pkg-config
```

Arch Linux

```
sudo pacman -Syu
sudo pacman -S curl libxml2 pkgconf
```

CentOS

```
sudo yum install epel-release
sudo yum install curl-devel libxml2-devel pkgconfig
```

Building the Project

Follow the steps below to build the project:

1. Clone the project repository:

   ```
   git clone https://github.com/piyushS3V3N/cpp_scrapper-ver-0.1-alpha.git
   cd web-scraper
   ```

2. Create a build directory and navigate to it:

   ```
   mkdir build
   cd build
   ```

3. Generate the Makefile using CMake:

   ```
   cmake ..
   ```

4. Build the project:

   ```
   cmake --build .
   ```

Running the Web Scraper

To run the web scraper, execute the built binary:

```
./web_scraper
```

The program will fetch search results for the query "terrorist attack in india" from DuckDuckGo, extract the text content from each website, and perform sentiment analysis on the extracted text.

Please note that web scraping may violate the terms of service of some websites. Ensure you have permission to scrape the websites before using this tool on a large scale.

To-Do List

- Improve error handling for web requests and parsing.
- Implement support for fetching more search result pages.
- Add unit tests for critical functions.
- Implement user input for search queries.

Additional Notes

- The web scraper will automatically download the VADER lexicon for sentiment analysis. VADER is a lexicon and rule-based sentiment analysis tool that is commonly used for social media text analysis.
- The program uses the `std::optional` type to handle the result of sentiment analysis. This requires C++17 support. If your compiler does not support C++17, you may need to update your compiler or modify the code to use a different approach for handling optional values.
- The web scraper is designed for educational purposes and may require modifications to be used in a production environment.

Remember to be respectful when scraping websites and follow their terms of service. Unauthorized scraping may lead to legal issues or getting banned from the website.