# Documentation for WebCrawler Script

Latest documentation by Tanishk Deoghare

Latest update: 26 June 2024

## Overview

This script automates the process of crawling home pages of specified banks and all the sub-pages linked on the home page, extracting links from those pages, and searching for specific ESG (Environmental, Social, and Governance) terms within those links. The results are stored in Excel files for further analysis.

## Prerequisites

Required installations:

* Node.js
* npm (Node Package Manager)
* Required npm packages: puppeteer, xlsx, path, fs, slugify, html2plaintext

## File Structure

* esg\_terms.json: JSON file containing ESG terms to search for.
* banks-to-crawl.json: JSON file containing the list of bank URLs to crawl.
* links.xlsx: Excel file where links extracted from the home pages are stored.
* results.xlsx: Excel file where the results of the ESG terms search are stored.
* parsing\_test: Directory to store the plain text content of crawled pages.

## Script

### 1. Crawl the Home Pages

**Function:** crawlHomePages

* + Crawls the home pages of banks listed in banks-to-crawl.json, extracts all the links from these pages, and writes the links to links.xlsx.

**Function:** writeLinksToExcel

* + Writes the extracted links to links.xlsx.

### 2: Read Links from Excel File

**Function:** readLinksFromExcel

* + Reads the links from links.xlsx.

### 3: Crawl Links and Search for ESG Terms

**Function:** crawlLinksAndSearchESGTerms

* + Crawls each link from links.xlsx, extracts the plain text content, searches for ESG terms, and saves the results.

**Function:** searchWords

* + Searches for the specified ESG terms within the text.

### 4: Save Results to Excel File

**Function:** writeResultsToExcel

* + Writes the results of the ESG terms search to results.xlsx.

## How to Run

1. Install Node.js on the system
2. Install the necessary npm packages:

Run the following code in the working directory -> npm install puppeteer xlsx slugify html2plaintext

1. Ensure the esg\_terms.json and banks-to-crawl.json files are in the same directory as the script.
2. Add the list of banks to crawl and esg terms to search for in the respective files.
3. Create a directory named parsing\_test in the same location as the script to store the plain text content of the crawled pages.
4. Run the script using the following command -> node json\_testing\_links\_results2.js
5. Check the results in the results.xlsx file and remove any duplicate lines.