

Aim:-Implement a simple web crawler

Program Code

```
from bs4 import BeautifulSoup
import requests
pages_crawled = []
def crawler(url):
    page = requests.get(url)
    soup = BeautifulSoup(page.text, 'html.parser')
    links = soup.find_all('a')
    for link in links:
        if 'href' in link.attrs:
            if link['href'].startswith('/wiki') and '.' not in link['href']:
                if link['href'] not in pages_crawled:
                    new_link = f"https://en.wikipedia.org{link['href']}"
                    pages_crawled.append(link['href'])
                    try:
                        with open('data.csv', 'a') as file:
                            file.write(f'{soup.title.text}; {soup.h1.text}; {link["href"]}\n')
                        crawler(new_link)
                    except:
                        continue
crawler('https://en.wikipedia.org')
```

Output

```
The file was loaded in a wrong encoding: 'UTF-8'      Reload in 'windows-1252'  Set project encoding to 'windows-1252'  Reload in another encoding

1  Wikipedia, the free encyclopedia; Main Page; /wiki/Wikipedia
2  Wikipedia - Wikipedia; Wikipedia; /wiki/Main_Page
3  Wikipedia, the free encyclopedia; Main Page; /wiki/Free_content
4  Free content - Wikipedia; Free content; /wiki/Definition_of_Free_Cultural_Works
5  Definition of Free Cultural Works - Wikipedia; Definition of Free Cultural Works; /wiki/Free_content_movement
6  Free-culture movement - Wikipedia; Free-culture movement; /wiki/Free_culture_(disambiguation)
7  Free Culture - Wikipedia; Free Culture; /wiki/Free_Culture_(book)
8  Free Culture (book) - Wikipedia; Free Culture (book); /wiki/Lawrence_Lessig
9  Lawrence Lessig - Wikipedia; Lawrence Lessig; /wiki/Lawrence_Lessing
10 Lawrence Lessing - Wikipedia; Lawrence Lessing; /wiki/Science_writer
11 Science journalism - Wikipedia; Science journalism; /wiki/Scientific_journalism
12 Scientific journalism - Wikipedia; Scientific journalism; /wiki/Science_journalism
13 Science journalism - Wikipedia; Science journalism; /wiki/Scientific_writing
14 Scientific writing - Wikipedia; Scientific writing; /wiki/Science_writing
15 Science journalism - Wikipedia; Science journalism; /wiki/Science_communication
16 Science communication - Wikipedia; Science communication; /wiki/Science_publishing
17 Scientific literature - Wikipedia; Scientific literature; /wiki/Medical_literature
18 Medical literature - Wikipedia; Medical literature; /wiki/Edwin_Smith_Papyrus
19 Edwin Smith Papyrus - Wikipedia; Edwin Smith Papyrus; /wiki/New_York_Academy_of_Medicine
20 New York Academy of Medicine - Wikipedia; New York Academy of Medicine; /wiki/Eclecticism_in_architecture
21 Eclecticism in architecture - Wikipedia; Eclecticism in architecture; /wiki/Basilica
```