Name: Manav Mehta

Roll No.: 41445 **Batch:** R4 **Class:** BE-IV

Course: Information Retrieval

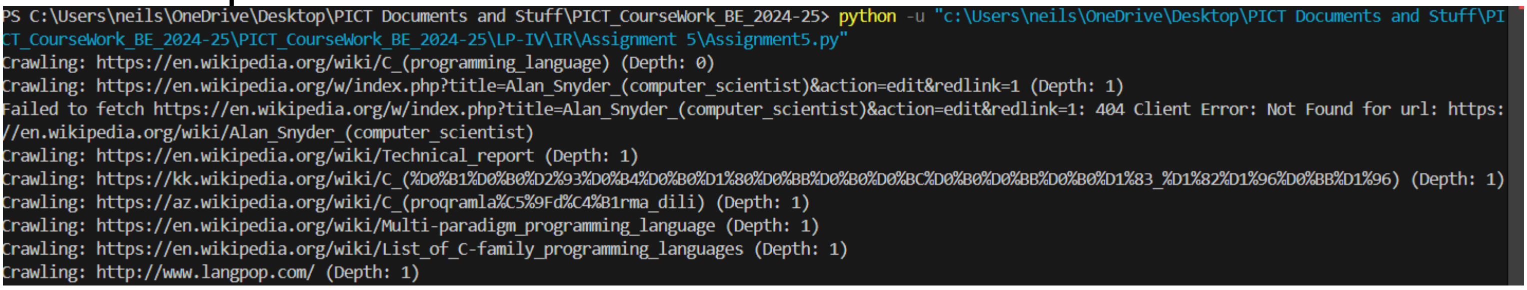
Assignment 5

i. **Code -**

```
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin, urlparse

def fetch_page(url):
    try:
        response = requests.get(url)
        response.raise_for_status()
        return BeautifulSoup(response.text, "html.parser")
    except requests.exceptions.RequestException as e:
        print(f"Failed to fetch {url}: {e}")
        return None

def extract_links(soup, base_url):
    links = set()
    for anchor in soup.find_all("a", href=True):
        href = anchor["href"]
        full_url = urljoin(base_url, href)
        if urlparse(full_url).scheme in ["http", "https"]:
            links.add(full_url)
    return links

def crawl(start_url, depth=2):
    visited = set()
    to_visit = [(start_url, 0)]

    while to_visit:
        current_url, current_depth = to_visit.pop(0)
        if current_url in visited or current_depth > depth:
            continue
        print(f"Crawling: {current_url} (Depth: {current_depth})")
        visited.add(current_url)
        soup = fetch_page(current_url)
        if soup is None:
            continue
        links = extract_links(soup, current_url)
        for link in links:
            if link not in visited:
                to_visit.append((link, current_depth + 1))

start_url = "https://en.wikipedia.org/wiki/C_(programming_language)"
```

```
crawl(start_url, depth=2)
```

ii. **Output -**



```
PS C:\Users\neils\OneDrive\Desktop\PICT Documents and Stuff\PICT_CourseWork_BE_2024-25> python -u "c:\Users\neils\OneDrive\Desktop\PICT Documents and Stuff\PI
CT_CourseWork_BE_2024-25\PICT_CourseWork_BE_2024-25\LP-IV\IR\Assignment 5\Assignment5.py"
Crawling: https://en.wikipedia.org/wiki/C_(programming_language) (Depth: 0)
Crawling: https://en.wikipedia.org/w/index.php?title=Alan_Snyder_(computer_scientist)&action=edit&redlink=1 (Depth: 1)
Failed to fetch https://en.wikipedia.org/w/index.php?title=Alan_Snyder_(computer_scientist)&action=edit&redlink=1: 404 Client Error: Not Found for url: https:
//en.wikipedia.org/wiki/Alan_Snyder_(computer_scientist)
Crawling: https://en.wikipedia.org/wiki/Technical_report (Depth: 1)
Crawling: https://kk.wikipedia.org/wiki/C_(%D0%B1%D0%B0%D2%93%D0%B4%D0%B0%D1%80%D0%BB%D0%B0%D0%BC%D0%B0%D0%BB%D0%B0%D1%83_%D1%82%D1%96%D0%BB%D1%96) (Depth: 1)
Crawling: https://az.wikipedia.org/wiki/C_(proqramla%C5%9Fd%C4%B1rma_dili) (Depth: 1)
Crawling: https://en.wikipedia.org/wiki/Multi-paradigm_programming_language (Depth: 1)
Crawling: https://en.wikipedia.org/wiki/List_of_C-family_programming_languages (Depth: 1)
Crawling: http://www.langpop.com/ (Depth: 1)
```