AIM

Program to implement a simple web crawler (ensure ethical conduct).

INSTALLATION CODE

pip install requests bs4

OUTPUT

Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (2.23.0)
Requirement already satisfied: bs4 in /usr/local/lib/python3.7/dist- packages (0.0.1)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests) (2021.10.8)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests) (1.24.3) Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests) (2.10)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.7/dist-packages (from bs4) (4.6.3)

Programming code:

```
import logging
from urllib.parse
import urljoin
import requests
from bs4 import BeautifulSoup
logging.basicConfig(
format='%(asctime)s %(levelname)s:%(message)s',
level=logging.INFO)
class Crawler:
def init (self,
urls=[]):
self.visited_urls
= []
self.urls_to_visit
= urls
```

```
def download_url(self,
    url): return
    requests.get(url).text

def get_linked_urls(self, url, html):
soup = BeautifulSoup(html,
'html.parser')   for link in
soup.find_all('a'):
path = link.get('href')
      if path and path.startswith('/'):
path = urljoin(url, path)
yield path
def add_url_to_visit(self, url):
if url not in self.visited_urls and url not in self.urls_to_visit:
self.urls_to_visit.append(url)
def crawl(self, url):
   html = self.download_url(url)
   for url in self.get_linked_urls(url, html):
    self.add_url_to_visit(url)

def run(self):
while self.urls_to_visit:
url = self.urls_to_visit.pop(0)
logging.info(f'Crawling: {url}')
try:
self.crawl(url)
except Exception:
logging.exception(f'Failed to crawl: {url}')
finally:
self.visited_urls.append(url)
if__name_____== '_main_':
Crawler
(url: =['https://www.imdb.com/']).run()
```

OUTPUT

2022-03-22 10:42:36,095 INFO:Crawling: https://www.imdb.com/
2022-03-22 10:42:36,931 INFO:Crawling:
https://www.imdb.com/?ref_=nv_home
2022-03-22 10:42:37,778 INFO:Crawling:
https://www.imdb.com/calendar/?ref_=nv_mv_cal
2022-03-22 10:42:38,164 INFO:Crawling:
https://www.imdb.com/list/ls016522954/?ref_=nv_tvv_dvd 2022-03-22 10:42:41,281 INFO:Crawling:
https://www.imdb.com/chart/top/?ref_=nv_mv_250
2022-03-22 10:42:42,869 INFO:Crawling:
https://www.imdb.com/chart/moviemeter/?ref_=nv_mv_mpm 2022-03-22 10:42:44,039 INFO:Crawling:

https://www.imdb.com/feature/genre/?ref_=nv_ch_gr
2022-03-22 10:42:44,413 INFO:Crawling:
https://www.imdb.com/chart/boxoffice/?ref_=nv_ch_cht 2022-

03-22 10:42:44,718 INFO:Crawling: https://www.imdb.com/showtimes/?ref_=nv_mv_sh
2022-03-22 10:42:45,305 INFO:Crawling: https://www.imdb.com/movies-in-theaters/?ref_=nv_mv_inth
2022-03-22 10:42:45,727 INFO:Crawling: https://www.imdb.com/coming-soon/?ref_=nv_mv_cs
2022-03-22 10:42:46,672 INFO:Crawling: https://www.imdb.com/news/movie/?ref_=nv_nw_mv
2022-03-22 10:42:47,212 INFO:Crawling: https://www.imdb.com/india/toprated/?ref_=nv_mv_in
2022-03-22 10:42:47,904 INFO:Crawling: https://www.imdb.com/whats-on-tv/?ref_=nv_tv_ontv
2022-03-22 10:42:48,300 INFO:Crawling: https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250
2022-03-22 10:42:49,114 INFO:Crawling: https://www.imdb.com/chart/tvmeter/?ref_=nv_tvv_mptv
2022-03-22 10:42:49,763 INFO:Crawling: https://www.imdb.com/feature/genre/
2022-03-22 10:42:50,141 INFO:Crawling: https://www.imdb.com/news/tv/?ref_=nv_nw_tv
2022-03-22 10:42:50,478 INFO:Crawling: https://www.imdb.com/india/tv?ref_=nv_tv_in
2022-03-22 10:42:50,898 INFO:Crawling: https://www.imdb.com/what-to-watch/?ref_=nv_watch
2022-03-22 10:42:51,572 INFO:Crawling: https://www.imdb.com/trailers/?ref_=nv_mv_tr
2022-03-22 10:42:52,003 INFO:Crawling: https://www.imdb.com/originals/?ref_=nv_sf_ori
2022-03-22 10:42:52,225 INFO:Crawling: https://www.imdb.com/imdbpicks/?ref_=nv_pi
2022-03-22 10:42:52,567 INFO:Crawling: https://www.imdb.com/podcasts/?ref_=nv_pod
2022-03-22 10:42:52,861 INFO:Crawling: https://www.imdb.com/oscars/?ref_=nv_ev_acd
2022-03-22 10:42:53,254 INFO:Crawling: https://m.imdb.com/feature/bestpicture/?ref_=nv_ch_osc 2022-03-22 10:42:53,893 INFO:Crawling: https://www.imdb.com/search/title/?count=100&groups=oscar_best_picture_winners&sort=year%2Cdesc&ref_=nv_ch_osc
2022-03-22 10:42:54,908 INFO:Crawling: https://www.imdb.com/emmys/?ref_=nv_ev_rte
2022-03-22 10:42:55,171 INFO:Crawling: https://www.imdb.com/imdbpicks/womenshistorymonth/?ref_=nv_ev_whm 2022-03-22 10:42:55,686 INFO:Crawling: https://www.imdb.com/starmeterawards/?ref_=nv_ev_sma
2022-03-22 10:42:56,004 INFO:Crawling: https://www.imdb.com/comic-con/?ref_=nv_ev_comic
2022-03-22 10:42:56,444 INFO:Crawling: https://www.imdb.com/nycc/?ref_=nv_ev_nycc
2022-03-22 10:42:56,790 INFO:Crawling: https://www.imdb.com/sundance/?ref_=nv_ev_sun