41443 - Anuj Mutha

```
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)
```

```
Crawler(urls=['https://www.imdb.com/']).run()
```

```
2022-11-02 11:46:14,917 INFO:Crawling: https://www.imdb.com/
2022-11-02 11:46:17,881 INFO:Crawling: https://www.imdb.com/?ref_=nv_home
2022-11-02 11:46:19,505 INFO:Crawling: https://www.imdb.com/calendar/?ref_=nv_mv_cal
2022-11-02 11:46:23,611 INFO:Crawling: https://www.imdb.com/chart/top/?ref_=nv_mv_250
2022-11-02 11:46:25,549 INFO:Crawling: https://www.imdb.com/chart/moviemeter/?ref_=nv
2022-11-02 11:46:26,964 INFO:Crawling: https://www.imdb.com/feature/genre/?ref_=nv_ch
2022-11-02 11:46:27,830 INFO:Crawling: https://www.imdb.com/chart/boxoffice/?ref_=nv_
2022-11-02 11:46:28,919 INFO:Crawling: https://www.imdb.com/showtimes/?ref_=nv_mv_sh
2022-11-02 11:46:31,067 INFO:Crawling: https://www.imdb.com/news/movie/?ref_=nv_nw_mv
2022-11-02 11:46:32,242 INFO:Crawling: https://www.imdb.com/india/toprated/?ref_=nv_m
2022-11-02 11:46:33,509 INFO:Crawling: https://www.imdb.com/whats-on-tv/?ref_=nv_tv_c
2022-11-02 11:46:34,491 INFO:Crawling: https://www.imdb.com/chart/toptv/?ref_=nv_tvv_
2022-11-02 11:46:36,461 INFO:Crawling: https://www.imdb.com/chart/tvmeter/?ref_=nv_tv
2022-11-02 11:46:37,699 INFO:Crawling: https://www.imdb.com/feature/genre/
2022-11-02 11:46:38,941 INFO:Crawling: https://www.imdb.com/news/tv/?ref_=nv_nw_tv
2022-11-02 11:46:39,754 INFO:Crawling: https://www.imdb.com/india/tv?ref_=nv_tv_in
2022-11-02 11:46:41,054 INFO:Crawling: https://www.imdb.com/what-to-watch/?ref_=nv_wa
2022-11-02 11:46:42,409 INFO:Crawling: https://www.imdb.com/trailers/?ref_=nv_mv_tr
2022-11-02 11:46:43,474 INFO:Crawling: https://www.imdb.com/originals/?ref_=nv_sf_ori
2022-11-02 11:46:45,058 INFO:Crawling: https://www.imdb.com/imdbpicks/?ref_=nv_pi
2022-11-02 11:46:46,055 INFO:Crawling: https://www.imdb.com/podcasts/?ref_=nv_pod
2022-11-02 11:46:47,158 INFO:Crawling: https://www.imdb.com/oscars/?ref_=nv_ev_acd
2022-11-02 11:46:48,194 INFO:Crawling: https://m.imdb.com/feature/bestpicture/?ref_=r
2022-11-02 11:46:50,681 INFO:Crawling: https://www.imdb.com/search/title/?count=100&g
2022-11-02 11:46:53,112 INFO:Crawling: https://www.imdb.com/emmys/?ref_=nv_ev_rte
2022-11-02 11:46:54,174 INFO:Crawling: https://www.imdb.com/starmeterawards/?ref_=nv_
2022-11-02 11:46:55,484 INFO:Crawling: https://www.imdb.com/comic-con/?ref_=nv_ev_com
2022-11-02 11:46:56,740 INFO:Crawling: https://www.imdb.com/nycc/?ref_=nv_ev_nycc
2022-11-02 11:46:57,883 INFO:Crawling: https://www.imdb.com/sundance/?ref_=nv_ev_sun
2022-11-02 11:46:59,122 INFO:Crawling: https://www.imdb.com/toronto/?ref_=nv_ev_tor
2022-11-02 11:47:00,129 INFO:Crawling: https://www.imdb.com/awards-central/?ref_=nv_e
2022-11-02 11:47:01,159 INFO:Crawling: https://www.imdb.com/festival-central/?ref_=nv
2022-11-02 11:47:02,196 INFO:Crawling: https://www.imdb.com/event/all/?ref_=nv_ev_all
2022-11-02 11:47:03,044 INFO:Crawling: https://www.imdb.com/feature/bornondate/?ref_=
2022-11-02 11:47:04,778 INFO:Crawling: https://m.imdb.com/chart/starmeter/?ref_=nv_ce
2022-11-02 11:47:06,786 INFO:Crawling: https://www.imdb.com/search/name/?match_all=tr
2022-11-02 11:47:08,099 INFO:Crawling: https://www.imdb.com/news/celebrity/?ref_=nv_c
2022-11-02 11:47:09,273 INFO:Crawling: https://help.imdb.com/imdb?ref_=cons_nb_hlp
2022-11-02 11:47:10,600 INFO:Crawling: https://contribute.imdb.com/czone?ref_=nv_cm_c
2022-11-02 11:47:12,032 INFO:Crawling: https://www.imdb.com/poll/?ref_=nv_cm_pl
2022-11-02 11:47:13,294 INFO:Crawling: https://pro.imdb.com?ref_=cons_nb_hm&rf=cons_r
2022-11-02 11:47:14,476 INFO:Crawling: https://www.imdb.com/search/
2022-11-02 11:47:15,459 INFO:Crawling: https://pro.imdb.com/login/ap?u=/login/lwa&imd
2022-11-02 11:47:15,768 INFO:Crawling: https://www.imdb.com/list/watchlist?ref_=nv_us
2022-11-02 11:47:16,879 INFO:Crawling: https://www.imdb.com/registration/signin?ref=r
2022-11-02 11:47:17,800 INFO:Crawling: https://help.imdb.com/article/issues/G6TCMBKAA
2022-11-02 11:47:18,436 INFO:Crawling: https://www.imdb.com/list/ls053181649/videopla
2022-11-02 11:47:19,806 INFO:Crawling: https://www.imdb.com/list/ls053181649/videopla
```

```
2022-11-02 11:47:21,156 INFO:Crawling: https://www.imdb.com/list/ls053181649/videopla
2022-11-02 11:47:22,535 INFO:Crawling: https://www.imdb.com/list/ls053181649/videopla
2022-11-02 11:47:24,468 INFO:Crawling: https://www.imdb.com/list/ls053181649/videopla
2022-11-02 11:47:25,960 INFO:Crawling: https://www.imdb.com/list/ls053181649/videopla
2022-11-02 11:47:27,259 INFO:Crawling: https://www.imdb.com/trailers/?ref_=hm_hp_sm
2022-11-02 11:47:28,272 INFO:Crawling: https://www.imdb.com/list/ls563618650/mediavie
2022-11-02 11:47:29,152 INFO:Crawling: https://www.imdb.com/list/ls563618650/mediavie
2022-11-02 11:47:29,613 INFO:Crawling: https://www.imdb.com/superheroes/hollywood-pow
2022-11-02 11:47:30,370 INFO:Crawling: https://www.imdb.com/superheroes/hollywood-pow
```