

```
pip install requests bs4
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (2.28.1)
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 (3.0.2)
Requirement already satisfied: certifi<2017.4.17 in /usr/local/lib/python3.7/dist-packages (2017.4.17)
Requirement already satisfied: urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (1.25.1)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (2.10)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.7/dist-packages (4.12.3)
```

```
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:
```

```

except Exception:
    logging.exception(f'Failed to crawl: {url}')
finally:
    self.visited_urls.append(url)

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

```

 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](https://www.imdb.com/?ref=nv_home)  
 2022-03-22 10:42:37,778 INFO:Crawling: [https://www.imdb.com/calendar/?ref=nv\\_mv\\_c](https://www.imdb.com/calendar/?ref=nv_mv_c)  
 2022-03-22 10:42:38,164 INFO:Crawling: <https://www.imdb.com/list/ls016522954/?ref>  
 2022-03-22 10:42:41,281 INFO:Crawling: [https://www.imdb.com/chart/top/?ref=nv\\_mv](https://www.imdb.com/chart/top/?ref=nv_mv)  
 2022-03-22 10:42:42,869 INFO:Crawling: <https://www.imdb.com/chart/moviemeter/?ref>  
 2022-03-22 10:42:44,039 INFO:Crawling: <https://www.imdb.com/feature/genre/?ref=nv>  
 2022-03-22 10:42:44,413 INFO:Crawling: <https://www.imdb.com/chart/boxoffice/?ref=>  
 2022-03-22 10:42:44,718 INFO:Crawling: [https://www.imdb.com/showtimes/?ref=nv\\_mv](https://www.imdb.com/showtimes/?ref=nv_mv)  
 2022-03-22 10:42:45,305 INFO:Crawling: <https://www.imdb.com/movies-in-theaters/?re>  
 2022-03-22 10:42:45,727 INFO:Crawling: [https://www.imdb.com/coming-soon/?ref=nv\\_r](https://www.imdb.com/coming-soon/?ref=nv_r)  
 2022-03-22 10:42:46,672 INFO:Crawling: [https://www.imdb.com/news/movie/?ref=nv\\_nw](https://www.imdb.com/news/movie/?ref=nv_nw)  
 2022-03-22 10:42:47,212 INFO:Crawling: <https://www.imdb.com/india/toprated/?ref=r>  
 2022-03-22 10:42:47,904 INFO:Crawling: [https://www.imdb.com/whats-on-tv/?ref=nv\\_t](https://www.imdb.com/whats-on-tv/?ref=nv_t)  
 2022-03-22 10:42:48,300 INFO:Crawling: [https://www.imdb.com/chart/toptv/?ref=nv\\_t](https://www.imdb.com/chart/toptv/?ref=nv_t)  
 2022-03-22 10:42:49,114 INFO:Crawling: <https://www.imdb.com/chart/tvmeter/?ref=nv>  
 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](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\\_ir](https://www.imdb.com/india/tv?ref=nv_tv_ir)  
 2022-03-22 10:42:50,898 INFO:Crawling: <https://www.imdb.com/what-to-watch/?ref=nv>  
 2022-03-22 10:42:51,572 INFO:Crawling: [https://www.imdb.com/trailers/?ref=nv\\_mv\\_t](https://www.imdb.com/trailers/?ref=nv_mv_t)  
 2022-03-22 10:42:52,003 INFO:Crawling: [https://www.imdb.com/originals/?ref=nv\\_sf](https://www.imdb.com/originals/?ref=nv_sf)  
 2022-03-22 10:42:52,225 INFO:Crawling: [https://www.imdb.com/imdbpicks/?ref=nv\\_pi](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](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\\_acc](https://www.imdb.com/oscars/?ref=nv_ev_acc)  
 2022-03-22 10:42:53,254 INFO:Crawling: <https://m.imdb.com/feature/bestpicture/?ref>  
 2022-03-22 10:42:53,893 INFO:Crawling: <https://www.imdb.com/search/title/?count=10>  
 2022-03-22 10:42:54,908 INFO:Crawling: [https://www.imdb.com/emmys/?ref=nv\\_ev\\_rte](https://www.imdb.com/emmys/?ref=nv_ev_rte)  
 2022-03-22 10:42:55,171 INFO:Crawling: <https://www.imdb.com/imdbpicks/womenshistor>  
 2022-03-22 10:42:55,686 INFO:Crawling: <https://www.imdb.com/starmeterawards/?ref=>  
 2022-03-22 10:42:56,004 INFO:Crawling: [https://www.imdb.com/comic-con/?ref=nv\\_ev](https://www.imdb.com/comic-con/?ref=nv_ev)  
 2022-03-22 10:42:56,444 INFO:Crawling: [https://www.imdb.com/nycc/?ref=nv\\_ev\\_nycc](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\\_s](https://www.imdb.com/sundance/?ref=nv_ev_s)  
 2022-03-22 10:42:57,177 INFO:Crawling: [https://www.imdb.com/toronto/?ref=nv\\_ev\\_tc](https://www.imdb.com/toronto/?ref=nv_ev_tc)  
 2022-03-22 10:42:57,489 INFO:Crawling: <https://www.imdb.com/awards-central/?ref=r>  
 2022-03-22 10:42:57,798 INFO:Crawling: <https://www.imdb.com/festival-central/?ref>  
 2022-03-22 10:42:58,136 INFO:Crawling: [https://www.imdb.com/event/all/?ref=nv\\_ev](https://www.imdb.com/event/all/?ref=nv_ev)  
 2022-03-22 10:42:58,454 INFO:Crawling: <https://www.imdb.com/feature/bornondate/?re>  
 2022-03-22 10:42:59,033 INFO:Crawling: <https://m.imdb.com/chart/starmeter/?ref=nv>  
 2022-03-22 10:42:59,555 INFO:Crawling: [https://www.imdb.com/search/name/?match\\_all](https://www.imdb.com/search/name/?match_all)  
 2022-03-22 10:43:00,325 INFO:Crawling: <https://www.imdb.com/news/celebrity/?ref=r>  
 2022-03-22 10:43:00,750 INFO:Crawling: [https://help.imdb.com/imdb?ref=cons\\_nb\\_hlp](https://help.imdb.com/imdb?ref=cons_nb_hlp)  
 2022-03-22 10:43:01,248 INFO:Crawling: [https://contribute.imdb.com/czone?ref=nv\\_c](https://contribute.imdb.com/czone?ref=nv_c)  
 2022-03-22 10:43:01,432 INFO:Crawling: [https://www.imdb.com/poll/?ref=nv\\_cm\\_pl](https://www.imdb.com/poll/?ref=nv_cm_pl)  
 2022-03-22 10:43:02,059 INFO:Crawling: [https://pro.imdb.com?ref=cons\\_nb\\_hm&rf=cor](https://pro.imdb.com?ref=cons_nb_hm&rf=cor)  
 2022-03-22 10:43:02,394 INFO:Crawling: <https://www.imdb.com/search/>  
 2022-03-22 10:43:02,660 INFO:Crawling: <https://pro.imdb.com/login/ap?u=/login/lwa8>  
 2022-03-22 10:43:02,759 INFO:Crawling: <https://www.imdb.com/list/watchlist?ref=nv>  
 2022-03-22 10:43:03,086 INFO:Crawling: <https://www.imdb.com/registration/signin?re>  
 2022-03-22 10:43:03,361 INFO:Crawling: <https://help.imdb.com/article/issues/G6TCME>  
 2022-03-22 10:43:03,746 INFO:Crawling: <https://www.imdb.com/list/ls053181649/videc>

```
2022-03-22 10:43:04,270 INFO:Crawling: https://www.imdb.com/list/ls053181649/video
2022-03-22 10:43:04,762 INFO:Crawling: https://www.imdb.com/list/ls025720609/video
2022-03-22 10:43:05,231 INFO:Crawling: https://www.imdb.com/list/ls053181649/video
2022-03-22 10:43:05,673 INFO:Crawling: https://www.imdb.com/list/ls025720609/video
2022-03-22 10:43:06,115 INFO:Crawling: https://www.imdb.com/list/ls053181649/video
2022-03-22 10:43:06,584 INFO:Crawling: https://www.imdb.com/trailers/?ref\_=hm\_hp\_s
```