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()
```

```
ERROR:root:Failed to crawl: None
Traceback (most recent call last):
  File "<ipython-input-3-c6e7bf87d013>", line 32, in run
    self.crawl(url)
  File "<ipython-input-3-c6e7bf87d013>", line 23, in crawl
    html = self.download_url(url)
  File "<ipython-input-3-c6e7bf87d013>", line 8, in download_url
    return requests.get(url).text
  File "/usr/local/lib/python3.7/dist-packages/requests/api.py", line 76, in get
    return request('get', url, params=params, **kwargs)
  File "/usr/local/lib/python3.7/dist-packages/requests/api.py", line 61, in request
    return session.request(method=method, url=url, **kwargs)
  File "/usr/local/lib/python3.7/dist-packages/requests/sessions.py", line 516, in reque
    prep = self.prepare_request(req)
  File "/usr/local/lib/python3.7/dist-packages/requests/sessions.py", line 459, in prepa
    hooks=merge_hooks(request.hooks, self.hooks),
  File "/usr/local/lib/python3.7/dist-packages/requests/models.py", line 314, in prepare
    self.prepare_url(url, params)
  File "/usr/local/lib/python3.7/dist-packages/requests/models.py", line 388, in prepare
    raise MissingSchema(error)
requests.exceptions.MissingSchema: Invalid URL 'None': No schema supplied. Perhaps you
```

Executing (3m 1s) (> r > … > dow… > g > r… > r… > … > … > u… > _ma… > get… > … > _re… > r… > re… > r… •••