**CSE205**

**Introduction to Networking**

**Project 2 report**

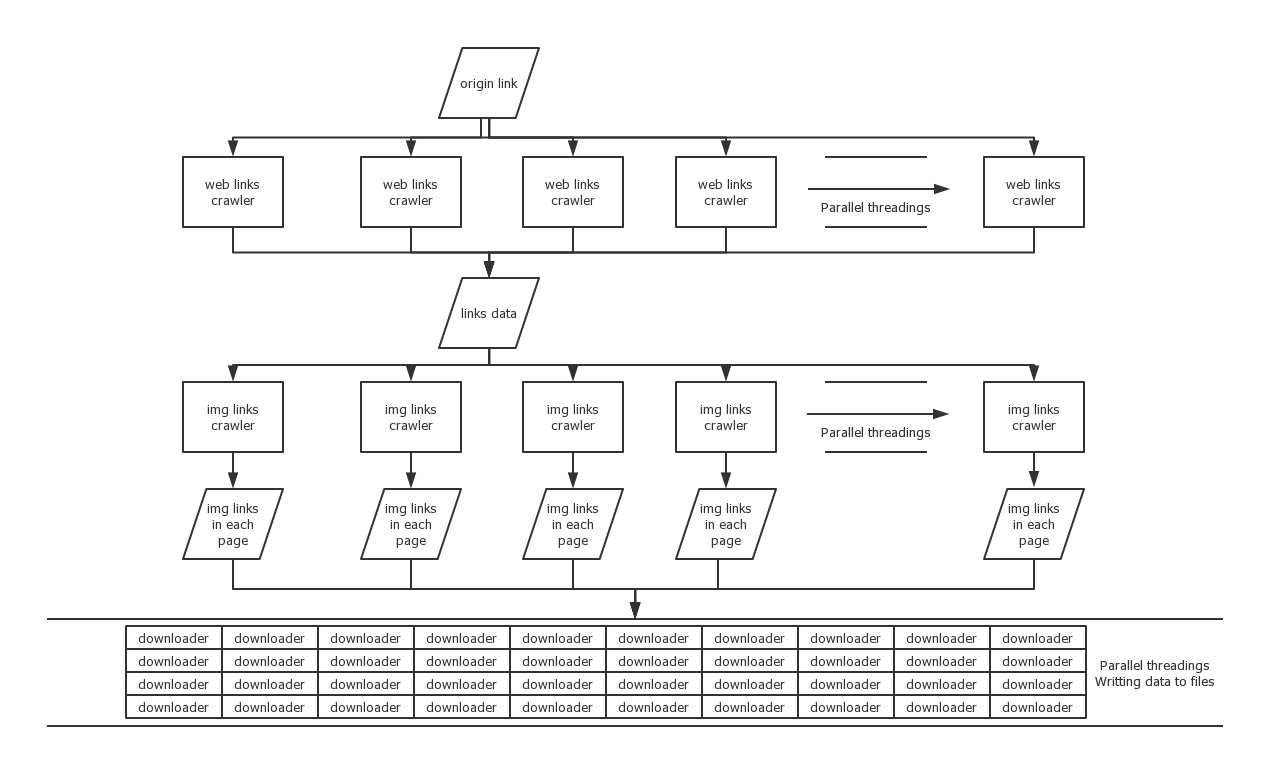
Zixuan.guo

1508672

**Features and implementations**

* HTTP GET request
  + Based on sockets programming
* HTTPs GET request
  + By using ssl library
* Download images in the pages
  + Based on html parser
* depth control based on breadth-first search
  + Attention: Origin page is depth 1
* Parallel multi-threading

**Instructions**



* Files: Crawler.py and forCSSE.py
* forCSSE.py is only for crawling the test page(<http://csse.xjtlu.edu.cn/classes/CSE205>)
* For general using, please run Crawler.py
  + You can change the link you want to crawl by changing crawl function in line 372
  + Make sure using high-end computer (The test environment use 12-thread CPU with 100m bandwidth network)
* Directory hierarchy
  + The folds’ names are their host name, instead of URLs, so there are only 2 levels folder hierarchy.

**Known bugs**

* Winerror 10054 and 10060
  + Caused by multiple reasons, but mostly caused by anti-crawler strategy.
* 'NoneType' object is not subscriptable
  + Happened when no data received.
* Errno 11004 getaddrinfo failed
  + Caused by wrong URL