



Collecting Data from the Web

Siaterlis Konstantinos

Layout

1. Data Collection
2. Scrape with Python
 - a. HTML Webpages
 - b. JavaScript Generated Webpages
3. Other Challenges
4. What about a Website?
 - a. Web Crawler Architecture
 - b. A Custom Web Crawler
5. Contact Me

Data Collection



APIs

WIKIPEDIA
facebook



Scrape



Web Scraping



DO NOT scrape everything immediately

DO NOT overload websites

Ask for permission



DISCLAIMER

Scrape with Python

Python Libs

- Urllib2 + BeautifulSoup4
- Scrapy
- Dryscrape
- Selenium + BeautifulSoup4
- etc



HTML Webpages

Urllib2 + BeautifulSoup4

```
hdr = {'User-Agent': 'Mozilla/4.0 (compatible; MSIE 5.5; Windows NT)'}  
req = urllib2.Request(url, headers=hdr)  
page = urllib2.urlopen(req)  
soup = BeautifulSoup(page.read(), "lxml")  
  
pageContent = soup.find('div', {'class': 'entry-content'})
```

HTML Webpages

Urllib2 + BeautifulSoup4

EXAMPLE 1

HTML Webpages

Scrapy

```
scrapy startproject tutorial
```



```
tutorial/  
  scrapy.cfg          # deploy configuration file  
  
tutorial/  
  __init__.py         # project's Python module, you'll import your code from here  
  
  items.py            # project items definition file  
  
  pipelines.py        # project pipelines file  
  
  settings.py         # project settings file  
  
  spiders/  
    __init__.py       # a directory where you'll later put your spiders
```

Create File

HTML Webpages


Scrapy

```
import scrapy

class QuotesSpider(scrapy.Spider):
    name = "meetupTest"

    def start_requests(self):
        urls = [
            'https://mydataminingssite.com/2017/03/13/topic-modeling-with-python/',
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = 'temp-%s.html' % page
        with open(filename, 'wb') as f:
            f.write(response.body)
        self.log('Saved file %s' % filename)
```



```
scrapy crawl meetupTest
```

HTML Webpages

Scrapy

EXAMPLE 2

Javascript Generated Webpages

Selenium + BeautifulSoup4

```
driver = webdriver.Chrome()
driver.get(site)
time.sleep(1)
page = driver.page_source
driver.close()

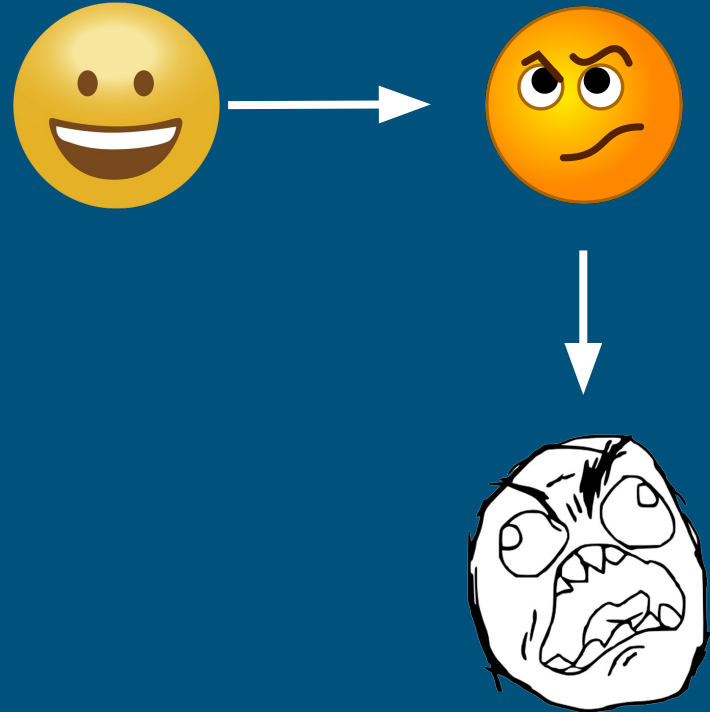
soup = BeautifulSoup(page, "lxml")
wrapper = soup.find('div', {'id': 'js-list'})
```

Javascript Generated Webpages

Selenium + Urllib2 + BeautifulSoup4

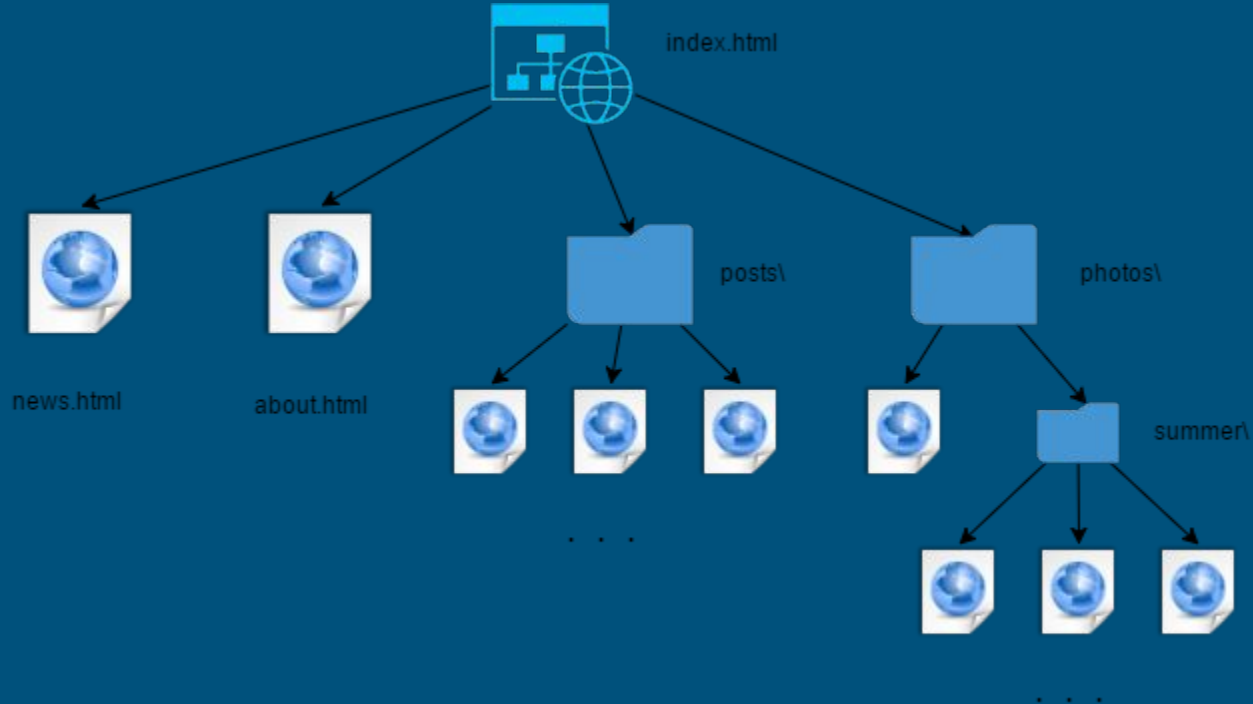
EXAMPLE 3

Challenges on Scraping



- Bad designed webpages
- Pop-ups
- Filtering

What about a Website

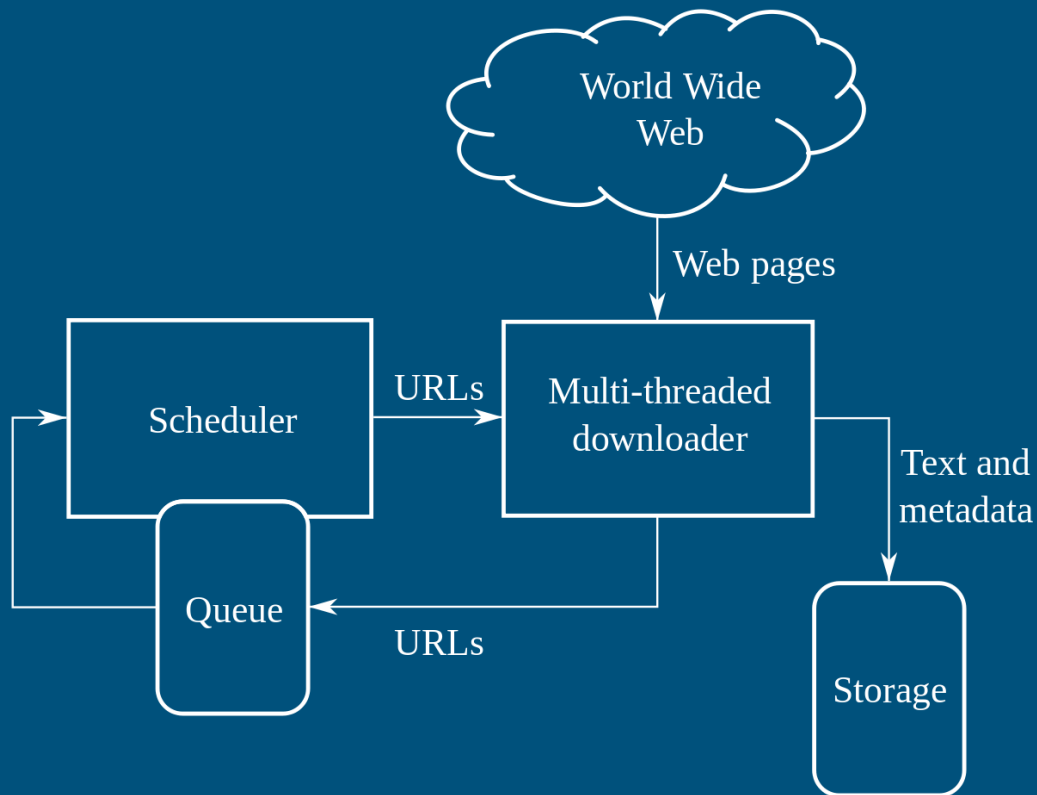


Web Crawlers

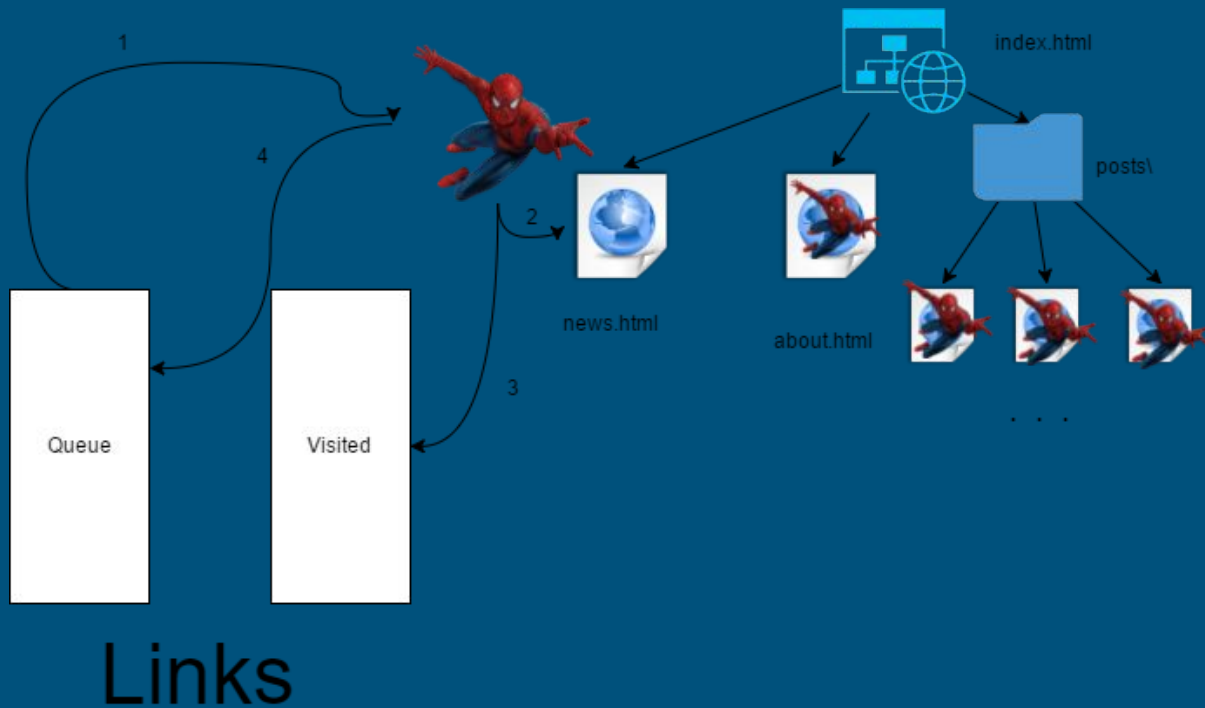
- Scrapy
- Mechanize
- Twill (based on Mechanize)
- Custom



Web Crawler Architecture



Custom Web Crawler



Custom Web Crawler in Python



Contact Me



@siaterliskonsta



<https://mydataminingsite.com/>



siaterliskonstantinos



siakon89



siaterliskonsta@gmail.com

