UPES

UNIVERSITY OF THE FUTURE

Test Automation

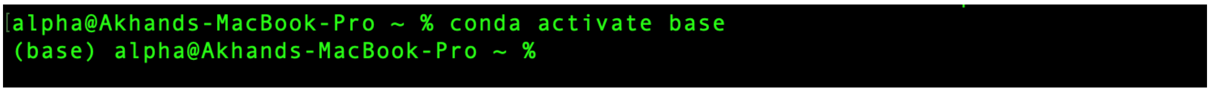
Experiment:

Scraping Images through scrapy

Submitted by :
Akhand Pratap Singh

Submitted to :
Mr. Alind

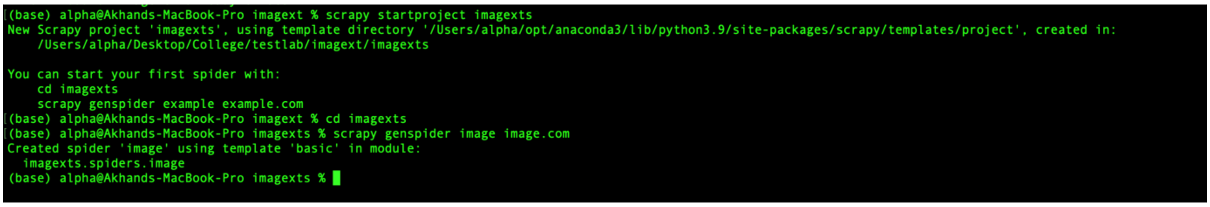
Step 1: Activate environment:



```
alpha@Akhands-MacBook-Pro ~ % conda activate base
(base) alpha@Akhands-MacBook-Pro ~ %
```

Step 2: Create a project

Command :-   scrapy startproject project_name



```
(base) alpha@Akhands-MacBook-Pro imagext % scrapy startproject imagexts
New Scrapy project 'imagexts', using template directory '/Users/alpha/opt/anaconda3/lib/python3.9/site-packages/scrapy/templates/project', created in:
    /Users/alpha/Desktop/College/testlab/imagext/imagexts

You can start your first spider with:
    cd imagexts
    scrapy genspider example example.com
(base) alpha@Akhands-MacBook-Pro imagext % cd imagexts
(base) alpha@Akhands-MacBook-Pro imagexts % scrapy genspider image image.com
Created spider 'image' using template 'basic' in module:
  imagexts.spiders.image
(base) alpha@Akhands-MacBook-Pro imagexts %
```

Step 3: RUN Command

scrapy genspider image image.com

Step 4: Write required code for Scraping Images through scrapy

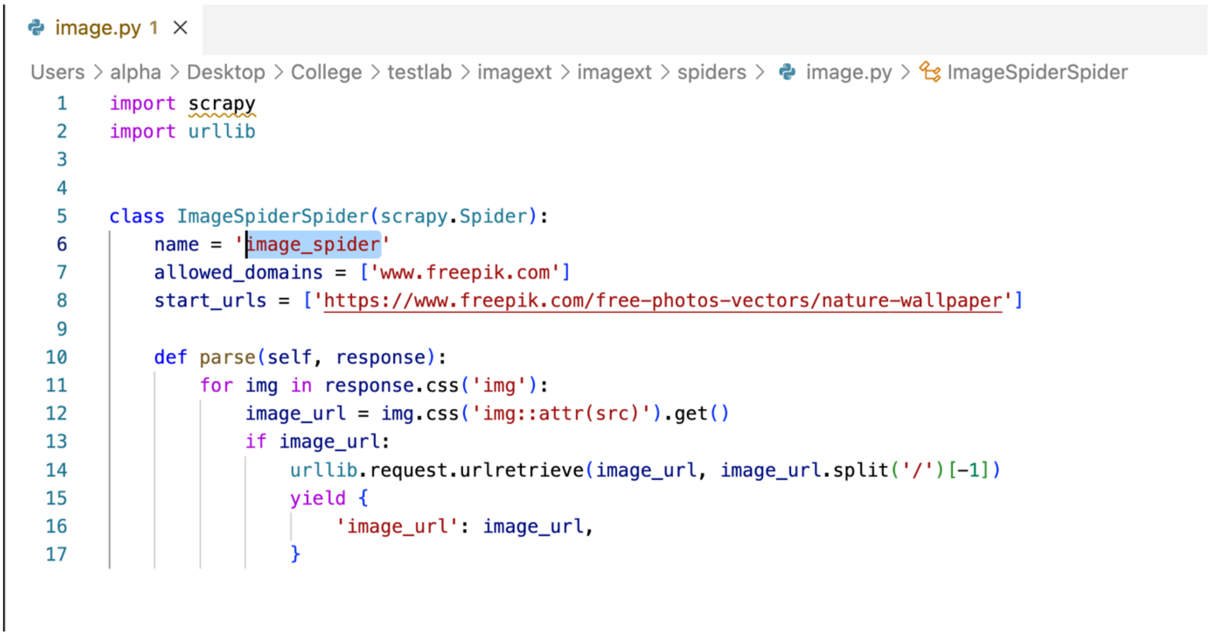
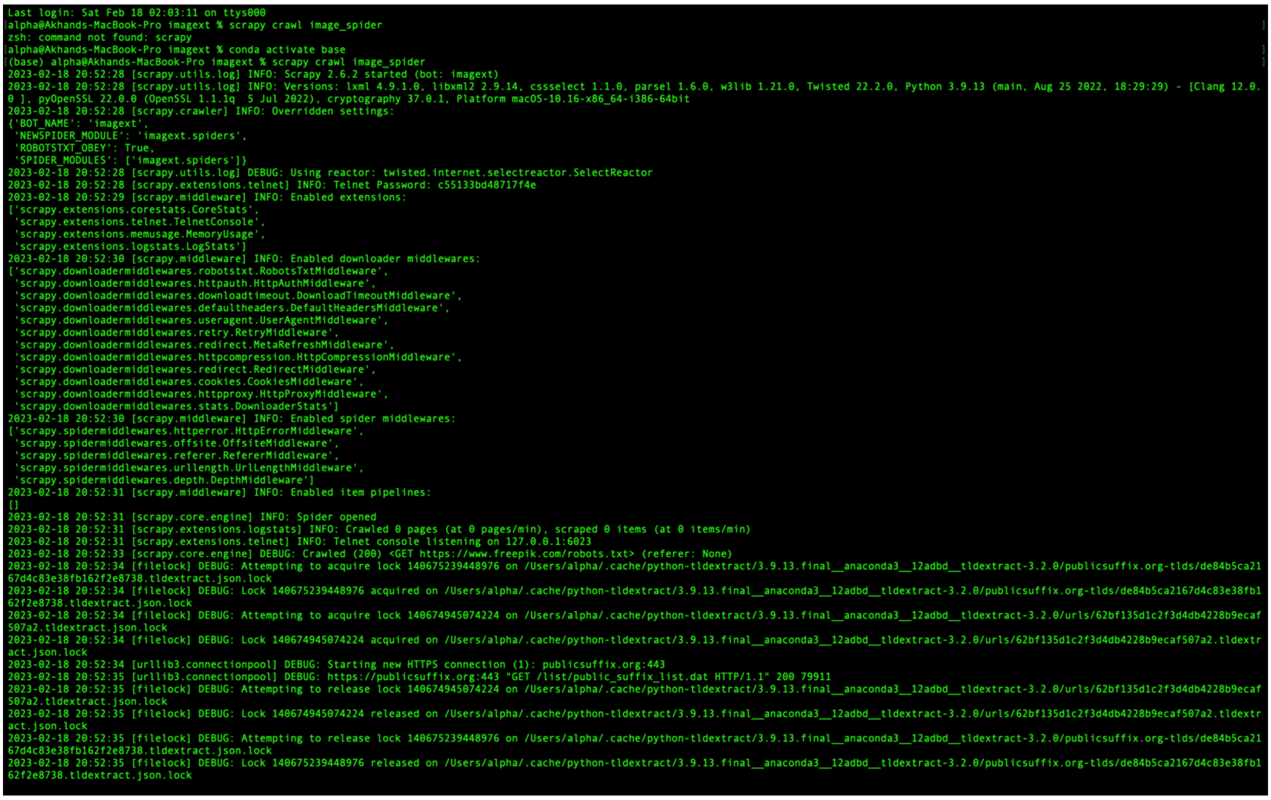


image.py 1 ×
Users > alpha > Desktop > College > testlab > imagext > imagext > spiders > image.py > ImageSpiderSpider

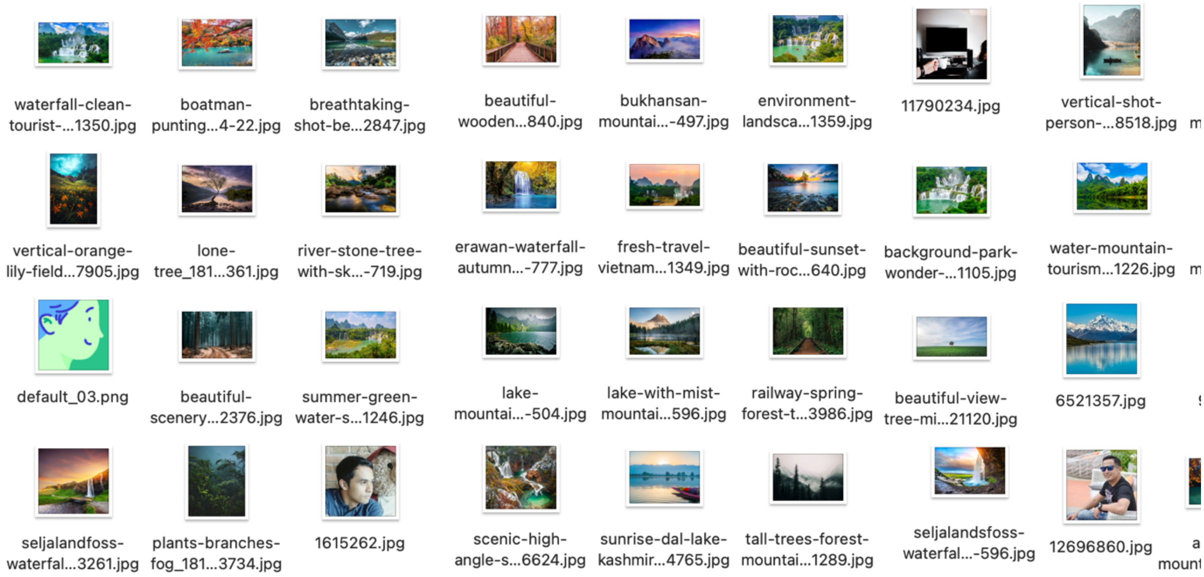
```python
import scrapy
import urllib


class ImageSpiderSpider(scrapy.Spider):
    name = 'image_spider'
    allowed_domains = ['www.freepik.com']
    start_urls = ['https://www.freepik.com/free-photos-vectors/nature-wallpaper']

    def parse(self, response):
        for img in response.css('img'):
            image_url = img.css('img::attr(src)').get()
            if image_url:
                urllib.request.urlretrieve(image_url, image_url.split('/')[-1])
                yield {
                    'image_url': image_url,
                }
```

Step 5: Run the script
Command : --   crapy crawl image_spider



```
Last login: Sat Feb 18 02:03:11 on ttys000
[alpha@Akhands-MacBook-Pro imagext % scrapy crawl image_spider
zsh: command not found: scrapy
[alpha@Akhands-MacBook-Pro imagext % conda activate base
[(base) alpha@Akhands-MacBook-Pro imagext % scrapy crawl image_spider
2023-02-18 20:52:28 [scrapy.utils.log] INFO: Scrapy 2.6.2 started (bot: imagext)
2023-02-18 20:52:28 [scrapy.utils.log] INFO: Versions: lxml 4.9.1.0, libxml2 2.9.14, cssselect 1.1.0, parsel 1.6.0, w3lib 1.21.0, Twisted 22.2.0, Python 3.9.13 (main, Aug 25 2022, 18:29:29) - [Clang 12.0.
0 ], pyOpenSSL 22.0.0 (OpenSSL 1.1.1q  5 Jul 2022), cryptography 37.0.1, Platform macOS-10.16-x86_64-i386-64bit
2023-02-18 20:52:28 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'imagext',
 'NEWSPIDER_MODULE': 'imagext.spiders',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['imagext.spiders']}
2023-02-18 20:52:28 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.selectreactor.SelectReactor
2023-02-18 20:52:28 [scrapy.extensions.telnet] INFO: Telnet Password: c55133bd48717f4e
2023-02-18 20:52:29 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.logstats.LogStats']
2023-02-18 20:52:30 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2023-02-18 20:52:30 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2023-02-18 20:52:31 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2023-02-18 20:52:31 [scrapy.core.engine] INFO: Spider opened
2023-02-18 20:52:31 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2023-02-18 20:52:31 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2023-02-18 20:52:33 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.freepik.com/robots.txt> (referer: None)
2023-02-18 20:52:34 [filelock] DEBUG: Attempting to acquire lock 140675239448976 on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/publicsuffix.org-tlds/de84b5ca21
67d4c83e38fb162f2e8738.tldextract.json.lock
2023-02-18 20:52:34 [filelock] DEBUG: Lock 140675239448976 acquired on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/publicsuffix.org-tlds/de84b5ca2167d4c83e38fb1
62f2e8738.tldextract.json.lock
2023-02-18 20:52:34 [filelock] DEBUG: Attempting to acquire lock 140674945074224 on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/urls/62bf135d1c2f3d4db4228b9ecaf
507a2.tldextract.json.lock
2023-02-18 20:52:34 [filelock] DEBUG: Lock 140674945074224 acquired on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/urls/62bf135d1c2f3d4db4228b9ecaf507a2.tldextr
act.json.lock
2023-02-18 20:52:34 [urllib3.connectionpool] DEBUG: Starting new HTTPS connection (1): publicsuffix.org:443
2023-02-18 20:52:35 [urllib3.connectionpool] DEBUG: https://publicsuffix.org:443 "GET /list/public_suffix_list.dat HTTP/1.1" 200 79911
2023-02-18 20:52:35 [filelock] DEBUG: Attempting to release lock 140674945074224 on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/urls/62bf135d1c2f3d4db4228b9ecaf
507a2.tldextract.json.lock
2023-02-18 20:52:35 [filelock] DEBUG: Lock 140674945074224 released on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/urls/62bf135d1c2f3d4db4228b9ecaf507a2.tldextr
act.json.lock
2023-02-18 20:52:35 [filelock] DEBUG: Attempting to release lock 140675239448976 on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/publicsuffix.org-tlds/de84b5ca21
67d4c83e38fb162f2e8738.tldextract.json.lock
2023-02-18 20:52:35 [filelock] DEBUG: Lock 140675239448976 released on /Users/alpha/.cache/python-tldextract/3.9.13.final__anaconda3__12adbd__tldextract-3.2.0/publicsuffix.org-tlds/de84b5ca2167d4c83e38fb1
62f2e8738.tldextract.json.lock
```

Downloaded image :


waterfall-clean-tourist-...1350.jpg
boatman-punting...4-22.jpg
breathtaking-shot-be...2847.jpg
beautiful-wooden...840.jpg
bukhansan-mountai...-497.jpg
environment-landsca...1359.jpg
11790234.jpg
vertical-shot-person-...8518.jpg
m
vertical-orange-lily-field...7905.jpg
lone-tree_181...361.jpg
river-stone-tree-with-sk...-719.jpg
erawan-waterfall-autumn...-777.jpg
fresh-travel-vietnam...1349.jpg
beautiful-sunset-with-roc...640.jpg
background-park-wonder-...1105.jpg
water-mountain-tourism...1226.jpg
m
default_03.png
beautiful-scenery...2376.jpg
summer-green-water-s...1246.jpg
lake-mountai...-504.jpg
lake-with-mist-mountai...596.jpg
railway-spring-forest-t...3986.jpg
beautiful-view-tree-mi...21120.jpg
6521357.jpg
seljalandfoss-waterfal...3261.jpg
plants-branches-fog_181...3734.jpg
1615262.jpg
scenic-high-angle-s...6624.jpg
sunrise-dal-lake-kashmir...4765.jpg
tall-trees-forest-mountai...1289.jpg
seljalandsfoss-waterfal...-596.jpg
12696860.jpg
a
mount