```
!pip install scrapy
!pip install pandas
!pip install boto3
```

```
Collecting scrapy
  Downloading Scrapy-2.11.2-py2.py3-none-any.whl.metadata (5.3 kB)
Collecting Twisted>=18.9.0 (from scrapy)
  Downloading twisted-24.7.0-py3-none-any.whl.metadata (18 kB)
Requirement already satisfied: cryptography>=36.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (43.0.1)
Collecting cssselect>=0.9.1 (from scrapy)
  Downloading cssselect-1.2.0-py2.py3-none-any.whl.metadata (2.2 kB)
Collecting itemloaders>=1.0.1 (from scrapy)
  Downloading itemloaders-1.3.1-py3-none-any.whl.metadata (3.9 kB)
Collecting parsel>=1.5.0 (from scrapy)
  Downloading parsel-1.9.1-py2.py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: pyOpenSSL>=21.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.2.1)
Collecting queuelib>=1.4.2 (from scrapy)
  Downloading queuelib-1.7.0-py2.py3-none-any.whl.metadata (5.7 kB)
Collecting service-identity>=18.1.0 (from scrapy)
  Downloading service_identity-24.1.0-py3-none-any.whl.metadata (4.8
kB)
Collecting w3lib>=1.17.0 (from scrapy)
  Downloading w3lib-2.2.1-py3-none-any.whl.metadata (2.1 kB)
Collecting zope.interface>=5.1.0 (from scrapy)
  Downloading zope.interface-7.0.3-cp310-cp310-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux
2014_x86_64.whl.metadata (43 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.6/43.6 kB 521.6 kB/s eta
0:00:00
 scrapy)
  Downloading Protego-0.3.1-py2.py3-none-any.whl.metadata (5.9 kB)
Collecting itemadapter>=0.1.0 (from scrapy)
  Downloading itemadapter-0.9.0-py3-none-any.whl.metadata (17 kB)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from scrapy) (71.0.4)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.1)
Collecting tldextract (from scrapy)
  Downloading tldextract-5.1.2-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: lxml>=4.4.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (4.9.4)
Requirement already satisfied: defusedxml>=0.7.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.7.1)
Collecting PyDispatcher>=2.0.5 (from scrapy)
  Downloading PyDispatcher-2.0.7-py3-none-any.whl.metadata (2.4 kB)
Requirement already satisfied: cffi>=1.12 in
/usr/local/lib/python3.10/dist-packages (from cryptography>=36.0.0-
>scrapy) (1.17.1)
```

```
Collecting jmespath>=0.9.5 (from itemloaders>=1.0.1->scrapy)
  Downloading jmespath-1.0.1-py3-none-any.whl.metadata (7.6 kB)
Requirement already satisfied: attrs>=19.1.0 in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (24.2.0)
Requirement already satisfied: pyasn1 in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (0.6.0)
Requirement already satisfied: pyasn1-modules in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (0.4.0)
Collecting automat>=0.8.0 (from Twisted>=18.9.0->scrapy)
  Downloading Automat-24.8.1-py3-none-any.whl.metadata (8.4 kB)
Collecting constantly>=15.1 (from Twisted>=18.9.0->scrapy)
  Downloading constantly-23.10.4-py3-none-any.whl.metadata (1.8 kB)
Collecting hyperlink>=17.1.1 (from Twisted>=18.9.0->scrapy)
  Downloading hyperlink-21.0.0-py2.py3-none-any.whl.metadata (1.5 kB)
Collecting incremental>=24.7.0 (from Twisted>=18.9.0->scrapy)
  Downloading incremental-24.7.2-py3-none-any.whl.metadata (8.1 kB)
Requirement already satisfied: typing-extensions>=4.2.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(4.12.2)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-
packages (from tldextract->scrapy) (3.8)
Requirement already satisfied: requests>=2.1.0 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(2.32.3)
Collecting requests-file>=1.4 (from tldextract->scrapy)
  Downloading requests_file-2.1.0-py2.py3-none-any.whl.metadata (1.7
kB)
Requirement already satisfied: filelock>=3.0.8 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(3.16.0)
Requirement already satisfied: pycparser in
/usr/local/lib/python3.10/dist-packages (from cffi>=1.12-
>cryptography>=36.0.0->scrapy) (2.22)
Requirement already satisfied: tomli in
/usr/local/lib/python3.10/dist-packages (from incremental>=24.7.0-
>Twisted>=18.9.0->scrapy) (2.0.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (2024.8.30)
Downloading Scrapy-2.11.2-py2.py3-none-any.whl (290 kB)
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 290.1/290.1 kB 8.2 MB/s eta
0:00:00
adapter-0.9.0-py3-none-any.whl (11 kB)
Downloading itemloaders-1.3.1-py3-none-any.whl (12 kB)
Downloading parsel-1.9.1-py2.py3-none-any.whl (17 kB)
Downloading Protego-0.3.1-py2.py3-none-any.whl (8.5 kB)
Downloading PyDispatcher-2.0.7-py3-none-any.whl (12 kB)
Downloading queuelib-1.7.0-py2.py3-none-any.whl (13 kB)
Downloading service_identity-24.1.0-py3-none-any.whl (12 kB)
Downloading twisted-24.7.0-py3-none-any.whl (3.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.2/3.2 MB 42.6 MB/s eta
0:00:00
anylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2
014_x86_64.whl (254 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 254.1/254.1 kB 14.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 97.6/97.6 kB 6.5 MB/s eta
0:00:00
at-24.8.1-py3-none-any.whl (42 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 42.6/42.6 kB 2.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 74.6/74.6 kB 4.3 MB/s eta
0:00:00
ental-24.7.2-py3-none-any.whl (20 kB)
Downloading jmespath-1.0.1-py3-none-any.whl (20 kB)
Downloading requests_file-2.1.0-py2.py3-none-any.whl (4.2 kB)
Installing collected packages: PyDispatcher, zope.interface, w3lib,
queuelib, protego, jmespath, itemadapter, incremental, hyperlink,
cssselect, constantly, automat, Twisted, requests-file, parsel,
tldextract, service-identity, itemloaders, scrapy
Successfully installed PyDispatcher-2.0.7 Twisted-24.7.0 automat-
24.8.1 constantly-23.10.4 cssselect-1.2.0 hyperlink-21.0.0
incremental-24.7.2 itemadapter-0.9.0 itemloaders-1.3.1 jmespath-1.0.1
parsel-1.9.1 protego-0.3.1 queuelib-1.7.0 requests-file-2.1.0 scrapy-
2.11.2 service-identity-24.1.0 tldextract-5.1.2 w3lib-2.2.1
zope.interface-7.0.3
Requirement already satisfied: pandas in
/usr/local/lib/python3.10/dist-packages (2.1.4)
Requirement already satisfied: numpy<2,>=1.22.4 in
/usr/local/lib/python3.10/dist-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.1 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2024.1)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2-
>pandas) (1.16.0)
```

```
Collecting boto3
  Downloading boto3-1.35.17-py3-none-any.whl.metadata (6.6 kB)
Collecting botocore<1.36.0,>=1.35.17 (from boto3)
  Downloading botocore-1.35.17-py3-none-any.whl.metadata (5.7 kB)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in
/usr/local/lib/python3.10/dist-packages (from boto3) (1.0.1)
Collecting s3transfer<0.11.0,>=0.10.0 (from boto3)
  Downloading s3transfer-0.10.2-py3-none-any.whl.metadata (1.7 kB)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in
/usr/local/lib/python3.10/dist-packages (from
botocore<1.36.0,>=1.35.17->boto3) (2.8.2)
Requirement already satisfied: urllib3!=2.2.0,<3,>=1.25.4 in
/usr/local/lib/python3.10/dist-packages (from
botocore<1.36.0,>=1.35.17->boto3) (2.0.7)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-
dateutil<3.0.0,>=2.1->botocore<1.36.0,>=1.35.17->boto3) (1.16.0)
Downloading boto3-1.35.17-py3-none-any.whl (139 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 139.2/139.2 kB 3.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.5/12.5 MB 55.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 82.7/82.7 kB 5.3 MB/s eta
0:00:00
```

```
%%bash
scrapy startproject news_scraper
cd news_scraper
scrapy genspider livemint www.livemint.com
scrapy genspider economics www.economics.com
scrapy genspider telegraf www.telegraf.com
scrapy genspider inc42 www.inc42.com
scrapy genspider digitalterminal www.digitalterminal.com
```

```
New Scrapy project 'news_scraper', using template directory
'/usr/local/lib/python3.10/dist-packages/scrapy/templates/project',
created in:
    /content/news_scraper

You can start your first spider with:
    cd news_scraper
    scrapy genspider example example.com
Created spider 'livemint' using template 'basic' in module:
  news_scraper.spiders.livemint
Created spider 'economics' using template 'basic' in module:
  news_scraper.spiders.economics
Created spider 'telegraf' using template 'basic' in module:
  news_scraper.spiders.telegraf
Created spider 'inc42' using template 'basic' in module:
  news_scraper.spiders.inc42
```

```
Created spider 'digitalterminal' using template 'basic' in module:
  news_scraper.spiders.digitalterminal
```

```
!pip install scrapy boto3
```

```
Requirement already satisfied: scrapy in
/usr/local/lib/python3.10/dist-packages (2.11.2)
Requirement already satisfied: boto3 in
/usr/local/lib/python3.10/dist-packages (1.35.17)
Requirement already satisfied: Twisted>=18.9.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.7.0)
Requirement already satisfied: cryptography>=36.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (43.0.1)
Requirement already satisfied: cssselect>=0.9.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.2.0)
Requirement already satisfied: itemloaders>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.3.1)
Requirement already satisfied: parsel>=1.5.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.9.1)
Requirement already satisfied: pyOpenSSL>=21.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.2.1)
Requirement already satisfied: queuelib>=1.4.2 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.7.0)
Requirement already satisfied: service-identity>=18.1.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.1.0)
Requirement already satisfied: w3lib>=1.17.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (2.2.1)
Requirement already satisfied: zope.interface>=5.1.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (7.0.3)
Requirement already satisfied: protego>=0.1.15 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.3.1)
Requirement already satisfied: itemadapter>=0.1.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.9.0)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from scrapy) (71.0.4)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.1)
Requirement already satisfied: tldextract in
/usr/local/lib/python3.10/dist-packages (from scrapy) (5.1.2)
Requirement already satisfied: lxml>=4.4.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (4.9.4)
Requirement already satisfied: defusedxml>=0.7.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.7.1)
Requirement already satisfied: PyDispatcher>=2.0.5 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (2.0.7)
Requirement already satisfied: botocore<1.36.0,>=1.35.17 in
/usr/local/lib/python3.10/dist-packages (from boto3) (1.35.17)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in
/usr/local/lib/python3.10/dist-packages (from boto3) (1.0.1)
Requirement already satisfied: s3transfer<0.11.0,>=0.10.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from boto3) (0.10.2)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in
/usr/local/lib/python3.10/dist-packages (from
botocore<1.36.0,>=1.35.17->boto3) (2.8.2)
Requirement already satisfied: urllib3!=2.2.0,<3,>=1.25.4 in
/usr/local/lib/python3.10/dist-packages (from
botocore<1.36.0,>=1.35.17->boto3) (2.0.7)
Requirement already satisfied: cffi>=1.12 in
/usr/local/lib/python3.10/dist-packages (from cryptography>=36.0.0-
>scrapy) (1.17.1)
Requirement already satisfied: attrs>=19.1.0 in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (24.2.0)
Requirement already satisfied: pyasn1 in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (0.6.0)
Requirement already satisfied: pyasn1-modules in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (0.4.0)
Requirement already satisfied: automat>=0.8.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(24.8.1)
Requirement already satisfied: constantly>=15.1 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(23.10.4)
Requirement already satisfied: hyperlink>=17.1.1 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(21.0.0)
Requirement already satisfied: incremental>=24.7.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(24.7.2)
Requirement already satisfied: typing-extensions>=4.2.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(4.12.2)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-
packages (from tldextract->scrapy) (3.8)
Requirement already satisfied: requests>=2.1.0 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(2.32.3)
Requirement already satisfied: requests-file>=1.4 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(2.1.0)
Requirement already satisfied: filelock>=3.0.8 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(3.16.0)
Requirement already satisfied: pycparser in
/usr/local/lib/python3.10/dist-packages (from cffi>=1.12-
>cryptography>=36.0.0->scrapy) (2.22)
Requirement already satisfied: tomli in
```

```
/usr/local/lib/python3.10/dist-packages (from incremental>=24.7.0-
>Twisted>=18.9.0->scrapy) (2.0.1)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-
dateutil<3.0.0,>=2.1->botocore<1.36.0,>=1.35.17->boto3) (1.16.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (3.3.2)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (2024.8.30)
```

```python
import scrapy
from scrapy.crawler import CrawlerProcess
import json

# Define the structure of the article item
class NewsArticle(scrapy.Item):
    article_url = scrapy.Field()
    title = scrapy.Field()
    author_name = scrapy.Field()
    author_url = scrapy.Field()
    article_content = scrapy.Field()
    published_date = scrapy.Field()

# Spider to scrape articles from Livemint
class LivemintSpider(scrapy.Spider):
    name = 'livemint'
    allowed_domains = ['www.livemint.com']
    start_urls = ['https://www.livemint.com/']

    def parse(self, response):
        # Extract links to all category pages
        category_links = response.css('nav.nav
a::attr(href)').getall()
        for link in category_links:
            yield response.follow(link, self.parse_category)

    def parse_category(self, response):
        # Extract links to individual articles
        article_links = response.css('h2.headline
a::attr(href)').getall()
        for link in article_links:
            yield response.follow(link, self.parse_article)

    def parse_article(self, response):
        # Extract article details
        article = NewsArticle()
        article['article_url'] = response.url
        article['title'] = response.css('h1::text').get()
        article['author_name'] = response.css('span.authorName
a::text').get()
        article['author_url'] = response.css('span.authorName
a::attr(href)').get()
        article['article_content'] = '
'.join(response.css('div.contentSec p::text').getall())
        article['published_date'] =
response.css('span.pubtime::text').get()

        yield article

# Pipeline to write the scraped data to a local JSON file
class JsonWriterPipeline:
    def open_spider(self, spider):
        self.file = open('articles.json', 'w')

    def close_spider(self, spider):
        self.file.close()

    def process_item(self, item, spider):
        line = json.dumps(dict(item)) + "\n"
        self.file.write(line)
        return item

# Configure the crawler process
process = CrawlerProcess(settings={
    "FEEDS": {
        "articles.json": {"format": "json"},  # This saves the output
to a local JSON file
    },
    "ITEM_PIPELINES": {
        '__main__.JsonWriterPipeline': 300,
    },
    "USER_AGENT": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36",
})

# Run the spider
process.crawl(LivemintSpider)
process.start()
```

```
INFO:scrapy.utils.log:Scrapy 2.11.2 started (bot: scrapybot)
2024-09-12 15:06:16 [scrapy.utils.log] INFO: Scrapy 2.11.2 started
(bot: scrapybot)
INFO:scrapy.utils.log:Versions: lxml 4.9.4.0, libxml2 2.10.3,
cssselect 1.2.0, parsel 1.9.1, w3lib 2.2.1, Twisted 24.7.0, Python
3.10.12 (main, Jul 29 2024, 16:56:48) [GCC 11.4.0], pyOpenSSL 24.2.1
(OpenSSL 3.3.2 3 Sep 2024), cryptography 43.0.1, Platform Linux-
6.1.85+-x86_64-with-glibc2.35
```

```
2024-09-12 15:06:16 [scrapy.utils.log] INFO: Versions: lxml 4.9.4.0,
libxml2 2.10.3, cssselect 1.2.0, parsel 1.9.1, w3lib 2.2.1, Twisted
24.7.0, Python 3.10.12 (main, Jul 29 2024, 16:56:48) [GCC 11.4.0],
pyOpenSSL 24.2.1 (OpenSSL 3.3.2 3 Sep 2024), cryptography 43.0.1,
Platform Linux-6.1.85+-x86_64-with-glibc2.35
INFO:scrapy.addons:Enabled addons:
[]
2024-09-12 15:06:16 [scrapy.addons] INFO: Enabled addons:
[]
/usr/local/lib/python3.10/dist-packages/scrapy/utils/request.py:254:
ScrapyDeprecationWarning: '2.6' is a deprecated value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting.

It is also the default value. In other words, it is normal to get this
warning if you have not defined a value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting. This is so for
backward compatibility reasons, but it will change in a future version
of Scrapy.

See the documentation of the 'REQUEST_FINGERPRINTER_IMPLEMENTATION'
setting for information on how to handle this deprecation.
  return cls(crawler)
DEBUG:scrapy.utils.log:Using reactor:
twisted.internet.epollreactor.EPollReactor
2024-09-12 15:06:16 [scrapy.utils.log] DEBUG: Using reactor:
twisted.internet.epollreactor.EPollReactor
INFO:scrapy.extensions.telnet:Telnet Password: c53d0dde6055a246
2024-09-12 15:06:16 [scrapy.extensions.telnet] INFO: Telnet Password:
c53d0dde6055a246
INFO:scrapy.middleware:Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2024-09-12 15:06:16 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
INFO:scrapy.crawler:Overridden settings:
{'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 '
               '(KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36'}
2024-09-12 15:06:16 [scrapy.crawler] INFO: Overridden settings:
{'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 '
```

```
               '(KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36'}
INFO:scrapy.middleware:Enabled downloader middlewares:
['scrapy.downloadermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',

'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddlewar
e',

'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware'
,
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',

'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddlewar
e',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2024-09-12 15:06:16 [scrapy.middleware] INFO: Enabled downloader
middlewares:
['scrapy.downloadermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',

'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddlewar
e',

'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware'
,
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',

'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddlewar
e',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
INFO:scrapy.middleware:Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2024-09-12 15:06:16 [scrapy.middleware] INFO: Enabled spider
middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
INFO:scrapy.middleware:Enabled item pipelines:
['__main__.JsonWriterPipeline']
2024-09-12 15:06:16 [scrapy.middleware] INFO: Enabled item pipelines:
['__main__.JsonWriterPipeline']
INFO:scrapy.core.engine:Spider opened
2024-09-12 15:06:16 [scrapy.core.engine] INFO: Spider opened
INFO:scrapy.extensions.logstats:Crawled 0 pages (at 0 pages/min),
scraped 0 items (at 0 items/min)
2024-09-12 15:06:16 [scrapy.extensions.logstats] INFO: Crawled 0 pages
(at 0 pages/min), scraped 0 items (at 0 items/min)
INFO:scrapy.extensions.telnet:Telnet console listening on
127.0.0.1:6023
2024-09-12 15:06:16 [scrapy.extensions.telnet] INFO: Telnet console
listening on 127.0.0.1:6023
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1):
publicsuffix.org:443
2024-09-12 15:06:17 [urllib3.connectionpool] DEBUG: Starting new HTTPS
connection (1): publicsuffix.org:443
DEBUG:urllib3.connectionpool:https://publicsuffix.org:443 "GET
/list/public_suffix_list.dat HTTP/1.1" 200 86760
2024-09-12 15:06:17 [urllib3.connectionpool] DEBUG:
https://publicsuffix.org:443 "GET /list/public_suffix_list.dat
HTTP/1.1" 200 86760
DEBUG:scrapy.core.engine:Crawled (200) <GET https://www.livemint.com/>
(referer: None)
2024-09-12 15:06:17 [scrapy.core.engine] DEBUG: Crawled (200) <GET
https://www.livemint.com/> (referer: None)
INFO:scrapy.core.engine:Closing spider (finished)
2024-09-12 15:06:17 [scrapy.core.engine] INFO: Closing spider
(finished)
INFO:scrapy.extensions.feedexport:Stored json feed (0 items) in:
articles.json
2024-09-12 15:06:17 [scrapy.extensions.feedexport] INFO: Stored json
feed (0 items) in: articles.json
INFO:scrapy.statscollectors:Dumping Scrapy stats:
{'downloader/request_bytes': 297,
 'downloader/request_count': 1,
 'downloader/request_method_count/GET': 1,
 'downloader/response_bytes': 200941,
 'downloader/response_count': 1,
 'downloader/response_status_count/200': 1,
 'elapsed_time_seconds': 0.747348,
 'feedexport/success_count/FileFeedStorage': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2024, 9, 12, 15, 6, 17, 481740,
tzinfo=datetime.timezone.utc),
 'httpcompression/response_bytes': 1183700,
 'httpcompression/response_count': 1,
 'log_count/DEBUG': 4,
 'log_count/INFO': 11,
 'memusage/max': 171704320,
 'memusage/startup': 171704320,
 'response_received_count': 1,
 'scheduler/dequeued': 1,
 'scheduler/dequeued/memory': 1,
 'scheduler/enqueued': 1,
 'scheduler/enqueued/memory': 1,
 'start_time': datetime.datetime(2024, 9, 12, 15, 6, 16, 734392,
tzinfo=datetime.timezone.utc)}
2024-09-12 15:06:17 [scrapy.statscollectors] INFO: Dumping Scrapy
stats:
{'downloader/request_bytes': 297,
 'downloader/request_count': 1,
 'downloader/request_method_count/GET': 1,
 'downloader/response_bytes': 200941,
 'downloader/response_count': 1,
 'downloader/response_status_count/200': 1,
 'elapsed_time_seconds': 0.747348,
 'feedexport/success_count/FileFeedStorage': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2024, 9, 12, 15, 6, 17, 481740,
tzinfo=datetime.timezone.utc),
 'httpcompression/response_bytes': 1183700,
 'httpcompression/response_count': 1,
 'log_count/DEBUG': 4,
 'log_count/INFO': 11,
 'memusage/max': 171704320,
 'memusage/startup': 171704320,
 'response_received_count': 1,
 'scheduler/dequeued': 1,
 'scheduler/dequeued/memory': 1,
 'scheduler/enqueued': 1,
 'scheduler/enqueued/memory': 1,
 'start_time': datetime.datetime(2024, 9, 12, 15, 6, 16, 734392,
tzinfo=datetime.timezone.utc)}
INFO:scrapy.core.engine:Spider closed (finished)
2024-09-12 15:06:17 [scrapy.core.engine] INFO: Spider closed
(finished)
```

```python
import scrapy
from scrapy.crawler import CrawlerProcess

class NewsArticle(scrapy.Item):
    article_url = scrapy.Field()
    title = scrapy.Field()
    author_name = scrapy.Field()
    author_url = scrapy.Field()
    article_content = scrapy.Field()
    published_date = scrapy.Field()

class LivemintSpider(scrapy.Spider):
    name = 'livemint'
    allowed_domains = ['livemint.com']
    start_urls = ['https://www.livemint.com/']

    def parse(self, response):
        # Extract category links (e.g., sections like Business,
Technology)
        category_links = response.css('nav.nav
a::attr(href)').getall()

        # Follow category links
        for link in category_links:
            yield response.follow(link, self.parse_category)

    def parse_category(self, response):
        # Extract article links from the category page
        article_links = response.css('h2.headline
a::attr(href)').getall()

        # Log for debugging
        print(f"Found article links: {article_links}")

        # Follow each article link
        for link in article_links:
            yield response.follow(link, self.parse_article)

    def parse_article(self, response):
        # Extract article details
        article = NewsArticle()
        article['article_url'] = response.url
        article['title'] = response.css('h1::text').get()
        article['author_name'] = response.css('span.authorName
a::text').get()
        article['author_url'] = response.css('span.authorName
a::attr(href)').get()
        article['article_content'] = '
'.join(response.css('div.contentSec p::text').getall())
        article['published_date'] =
response.css('span.pubtime::text').get()

        yield article

# Configure and run the spider
process = CrawlerProcess(settings={
    "FEEDS": {
        "articles.json": {"format": "json"},
    },
    "USER_AGENT": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36",
    "DOWNLOAD_DELAY": 1,  # Adding a delay to avoid overloading the
server
})

if not process.crawlers:
    process.crawl(LivemintSpider)
    process.start()
```

```
INFO:scrapy.utils.log:Scrapy 2.11.2 started (bot: scrapybot)
2024-09-12 15:13:38 [scrapy.utils.log] INFO: Scrapy 2.11.2 started
(bot: scrapybot)
INFO:scrapy.utils.log:Versions: lxml 4.9.4.0, libxml2 2.10.3,
cssselect 1.2.0, parsel 1.9.1, w3lib 2.2.1, Twisted 24.7.0, Python
3.10.12 (main, Jul 29 2024, 16:56:48) [GCC 11.4.0], pyOpenSSL 24.2.1
(OpenSSL 3.3.2 3 Sep 2024), cryptography 43.0.1, Platform Linux-
6.1.85+-x86_64-with-glibc2.35
2024-09-12 15:13:38 [scrapy.utils.log] INFO: Versions: lxml 4.9.4.0,
libxml2 2.10.3, cssselect 1.2.0, parsel 1.9.1, w3lib 2.2.1, Twisted
24.7.0, Python 3.10.12 (main, Jul 29 2024, 16:56:48) [GCC 11.4.0],
pyOpenSSL 24.2.1 (OpenSSL 3.3.2 3 Sep 2024), cryptography 43.0.1,
Platform Linux-6.1.85+-x86_64-with-glibc2.35
INFO:scrapy.addons:Enabled addons:
[]
2024-09-12 15:13:38 [scrapy.addons] INFO: Enabled addons:
[]
/usr/local/lib/python3.10/dist-packages/scrapy/utils/request.py:254:
ScrapyDeprecationWarning: '2.6' is a deprecated value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting.

It is also the default value. In other words, it is normal to get this
warning if you have not defined a value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting. This is so for
backward compatibility reasons, but it will change in a future version
of Scrapy.

See the documentation of the 'REQUEST_FINGERPRINTER_IMPLEMENTATION'
setting for information on how to handle this deprecation.
  return cls(crawler)
DEBUG:scrapy.utils.log:Using reactor:
twisted.internet.epollreactor.EPollReactor
2024-09-12 15:13:38 [scrapy.utils.log] DEBUG: Using reactor:
twisted.internet.epollreactor.EPollReactor
INFO:scrapy.extensions.telnet:Telnet Password: 9419ea00ee139f5d
2024-09-12 15:13:38 [scrapy.extensions.telnet] INFO: Telnet Password:
9419ea00ee139f5d
INFO:scrapy.middleware:Enabled extensions:
```

```
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2024-09-12 15:13:38 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
INFO:scrapy.crawler:Overridden settings:
{'DOWNLOAD_DELAY': 1,
 'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 '
               '(KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36'}
2024-09-12 15:13:38 [scrapy.crawler] INFO: Overridden settings:
{'DOWNLOAD_DELAY': 1,
 'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 '
               '(KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36'}
INFO:scrapy.middleware:Enabled downloader middlewares:
['scrapy.downloadermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',

'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddlewar
e',

'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware'
,
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',

'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddlewar
e',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2024-09-12 15:13:38 [scrapy.middleware] INFO: Enabled downloader
middlewares:
['scrapy.downloadermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',

'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddlewar
e',
'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware'
,
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',

'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddlewar
e',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
INFO:scrapy.middleware:Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2024-09-12 15:13:38 [scrapy.middleware] INFO: Enabled spider
middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
INFO:scrapy.middleware:Enabled item pipelines:
[]
2024-09-12 15:13:38 [scrapy.middleware] INFO: Enabled item pipelines:
[]
INFO:scrapy.core.engine:Spider opened
2024-09-12 15:13:38 [scrapy.core.engine] INFO: Spider opened
INFO:scrapy.extensions.logstats:Crawled 0 pages (at 0 pages/min),
scraped 0 items (at 0 items/min)
2024-09-12 15:13:38 [scrapy.extensions.logstats] INFO: Crawled 0 pages
(at 0 pages/min), scraped 0 items (at 0 items/min)
INFO:scrapy.extensions.telnet:Telnet console listening on
127.0.0.1:6024
2024-09-12 15:13:38 [scrapy.extensions.telnet] INFO: Telnet console
listening on 127.0.0.1:6024

----------------------------------------------------------------------
-----
ReactorNotRestartable                     Traceback (most recent call
last)
<ipython-input-6-7a5f48a5f896> in <cell line: 57>()
     57 if not process.crawlers:
     58     process.crawl(LivemintSpider)
---> 59     process.start()

/usr/local/lib/python3.10/dist-packages/scrapy/crawler.py in
start(self, stop_after_crawl, install_signal_handlers)
    427                 "after", "startup", install_shutdown_handlers,
self._signal_shutdown
    428             )
--> 429
reactor.run(installSignalHandlers=install_signal_handlers)  # blocking
call
    430
    431     def _graceful_stop_reactor(self) -> Deferred:

/usr/local/lib/python3.10/dist-packages/twisted/internet/base.py in
run(self, installSignalHandlers)
    697
    698     def run(self, installSignalHandlers: bool = True) -> None:
--> 699
self.startRunning(installSignalHandlers=installSignalHandlers)
    700         try:
    701             self.mainLoop()

/usr/local/lib/python3.10/dist-packages/twisted/internet/base.py in
startRunning(self, installSignalHandlers)
    928             raise error.ReactorAlreadyRunning()
    929         if self._startedBefore:
--> 930             raise error.ReactorNotRestartable()
    931
    932         self._signals.uninstall()

ReactorNotRestartable:
```

```
!pip install scrapy
```

```
Requirement already satisfied: scrapy in
/usr/local/lib/python3.10/dist-packages (2.11.2)
Requirement already satisfied: Twisted>=18.9.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.7.0)
Requirement already satisfied: cryptography>=36.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (43.0.1)
Requirement already satisfied: cssselect>=0.9.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.2.0)
Requirement already satisfied: itemloaders>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.3.1)
Requirement already satisfied: parsel>=1.5.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.9.1)
Requirement already satisfied: pyOpenSSL>=21.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.2.1)
Requirement already satisfied: queuelib>=1.4.2 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (1.7.0)
Requirement already satisfied: service-identity>=18.1.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.1.0)
Requirement already satisfied: w3lib>=1.17.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (2.2.1)
Requirement already satisfied: zope.interface>=5.1.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from scrapy) (7.0.3)
Requirement already satisfied: protego>=0.1.15 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.3.1)
Requirement already satisfied: itemadapter>=0.1.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.9.0)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from scrapy) (71.0.4)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.1)
Requirement already satisfied: tldextract in
/usr/local/lib/python3.10/dist-packages (from scrapy) (5.1.2)
Requirement already satisfied: lxml>=4.4.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (4.9.4)
Requirement already satisfied: defusedxml>=0.7.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (0.7.1)
Requirement already satisfied: PyDispatcher>=2.0.5 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (2.0.7)
Requirement already satisfied: cffi>=1.12 in
/usr/local/lib/python3.10/dist-packages (from cryptography>=36.0.0-
>scrapy) (1.17.1)
Requirement already satisfied: jmespath>=0.9.5 in
/usr/local/lib/python3.10/dist-packages (from itemloaders>=1.0.1-
>scrapy) (1.0.1)
Requirement already satisfied: attrs>=19.1.0 in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (24.2.0)
Requirement already satisfied: pyasn1 in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (0.6.0)
Requirement already satisfied: pyasn1-modules in
/usr/local/lib/python3.10/dist-packages (from service-
identity>=18.1.0->scrapy) (0.4.0)
Requirement already satisfied: automat>=0.8.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(24.8.1)
Requirement already satisfied: constantly>=15.1 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(23.10.4)
Requirement already satisfied: hyperlink>=17.1.1 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(21.0.0)
Requirement already satisfied: incremental>=24.7.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(24.7.2)
Requirement already satisfied: typing-extensions>=4.2.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy)
(4.12.2)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-
packages (from tldextract->scrapy) (3.8)
```

```
Requirement already satisfied: requests>=2.1.0 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(2.32.3)
Requirement already satisfied: requests-file>=1.4 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(2.1.0)
Requirement already satisfied: filelock>=3.0.8 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy)
(3.16.0)
Requirement already satisfied: pycparser in
/usr/local/lib/python3.10/dist-packages (from cffi>=1.12-
>cryptography>=36.0.0->scrapy) (2.22)
Requirement already satisfied: tomli in
/usr/local/lib/python3.10/dist-packages (from incremental>=24.7.0-
>Twisted>=18.9.0->scrapy) (2.0.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.1.0-
>tldextract->scrapy) (2024.8.30)
```

```python
import scrapy
from scrapy.crawler import CrawlerProcess

# Define your Scrapy item
class NewsArticle(scrapy.Item):
    article_url = scrapy.Field()
    title = scrapy.Field()
    author_name = scrapy.Field()
    author_url = scrapy.Field()
    article_content = scrapy.Field()
    published_date = scrapy.Field()

# Define your spider
class LivemintSpider(scrapy.Spider):
    name = 'livemint'
    allowed_domains = ['livemint.com']
    start_urls = ['https://www.livemint.com/']

    def parse(self, response):
        # Extract category links (e.g., sections like Business,
Technology)
        category_links = response.css('nav.nav
a::attr(href)').getall()

        # Follow category links
        for link in category_links:
            yield response.follow(link, self.parse_category)

    def parse_category(self, response):
        # Extract article links from the category page
        article_links = response.css('h2.headline
a::attr(href)').getall()

        # Follow each article link
        for link in article_links:
            yield response.follow(link, self.parse_article)

    def parse_article(self, response):
        # Extract article details
        article = NewsArticle()
        article['article_url'] = response.url
        article['title'] = response.css('h1::text').get()
        article['author_name'] = response.css('span.authorName
a::text').get()
        article['author_url'] = response.css('span.authorName
a::attr(href)').get()
        article['article_content'] = '
'.join(response.css('div.contentSec p::text').getall())
        article['published_date'] =
response.css('span.pubtime::text').get()

        yield article

# Run the spider
from twisted.internet import reactor
from scrapy.crawler import CrawlerRunner

runner = CrawlerRunner({
    "FEEDS": {
        "articles.json": {"format": "json"},
    },
    "USER_AGENT": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36",
    "DOWNLOAD_DELAY": 1,  # Delay to avoid overloading the server
})

# Run spider within the reactor's event loop
def run_spider():
    deferred = runner.crawl(LivemintSpider)
    deferred.addBoth(lambda _: reactor.stop())

run_spider()
```

```python
# Start the Twisted reactor to run Scrapy
reactor.run()
```

```
/usr/local/lib/python3.10/dist-packages/scrapy/utils/request.py:254:
ScrapyDeprecationWarning: '2.6' is a deprecated value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting.

It is also the default value. In other words, it is normal to get this
warning if you have not defined a value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting. This is so for
backward compatibility reasons, but it will change in a future version
of Scrapy.

See the documentation of the 'REQUEST_FINGERPRINTER_IMPLEMENTATION'
setting for information on how to handle this deprecation.
  return cls(crawler)
```

```python
import json

# Load the JSON file
with open('articles.json') as f:
    articles = json.load(f)

# Print the first few articles
for article in articles[:5]:
    print(json.dumps(article, indent=2))
```