

# Crawler Documentation

## Overview

This is a web crawler built using Scrapy to extract unique links from a specified website. It respects the robots.txt file by default but can be configured to ignore it if necessary.

## Requirements

- Python 3.6+
- Scrapy 2.11.2

## Usage

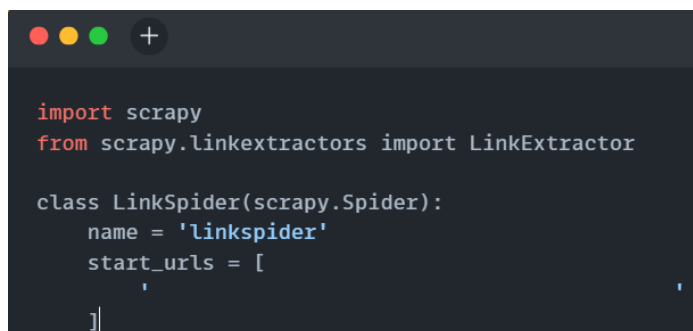
1. Navigate to the “**spiders**” directory using the following command :-

```
cd scrapper\crawler\webcrawler\webcrawler\spiders
```

2. Run the spider using the following command:

```
scrapy crawl linkspider
```

## Code Explanation

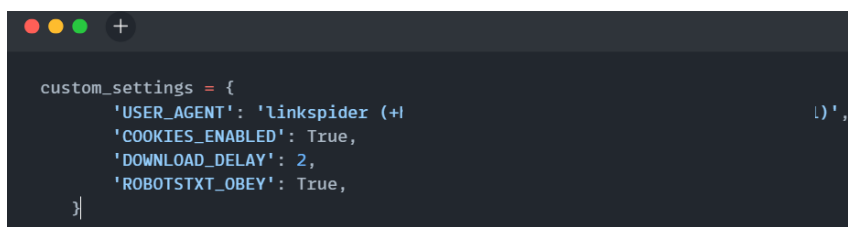


```
import scrapy
from scrapy.linkextractors import LinkExtractor

class LinkSpider(scrapy.Spider):
    name = 'linkspider'
    start_urls = [
        'http://www.example.com',
    ]
```

- Defines the “linkspider” class and the starting URL for the spider.

## Custom Settings



```
custom_settings = {
    'USER_AGENT': 'linkspider (+)',
    'COOKIES_ENABLED': True,
    'DOWNLOAD_DELAY': 2,
    'ROBOTSTXT_OBEY': True,
}
```

- Configures custom settings for the spider.

## Initialization

```
def __init__(self, *args, **kwargs):
    super(LinkSpider, self).__init__(*args, **kwargs)
    self.visited_urls = set()
    self.allowed_domains = ['']
```

- Initializes the spider and creates a set to store visited URLs.

## Parsing Method

```
def parse(self, response):
    links = LinkExtractor(allow_domains=self.allowed_domains).extract_links(response)

    extracted_links = set()
    for link in links:
        if link.url not in self.visited_urls:
            self.visited_urls.add(link.url)
            extracted_links.add(link.url)
            yield scrapy.Request(link.url, callback=self.parse)
```

- Extracts links from the response and recursively follows unvisited links. Also defines a particular domain, to avoid following rogue links.

## Saving Links

```
with open('links1.py', 'a') as f:
    f.write('unique_links = [\n')
    for link in extracted_links:
        f.write(f'    "{link}",\n')
    f.write(']\n')

self.log(f'Removed duplicates. Saved {len(extracted_links)} unique links to links1.py')
```

- Writes the unique links to a file and logs the result.

## Notes

- Modify the base url , start url and allowed domain as per target website.