🦀

Crawlab

https://github.com/crawlab-team/crawlab

installation is done with docker

```
FROM crawlabteam/crawlab:latest

# Install system dependencies
RUN apt-get update && apt-get install -y \
    curl \
    gnupg2 \
    git \
    unzip \
    libgconf-2-4 \
    libnss3 \
    libatk1.0-0 \
    libatk-bridge2.0-0 \
    libcups2 \
    libdrm2 \
    libxkbcommon0 \
    libxcomposite1 \
    libxdamage1 \
    libxfixes3 \
    libxrandr2 \
    libgbm1 \
    libasound2 \
    libpangocairo-1.0-0 \
    libgtk-3-0 \
    fonts-liberation \
    fonts-noto-color-emoji

# Install Google Chrome
RUN curl -sSL https://dl.google.com/linux/linux_signing_key.pub | apt-key add - \
    && echo "deb [arch=amd64] http://dl.google.com/linux/chrome/deb/ stable ma
    && apt-get update \
    && apt-get install -y google-chrome-stable \
    && rm -rf /var/lib/apt/lists/*

# Install Python dependencies
RUN pip install --no-cache-dir \
    scrapy-playwright \
    scrapegraphai \
    playwright

# Install Playwright browsers
RUN playwright install --with-deps chromium
```

```
version: '3.3'

services:
  master:
    build: ./crawlab  # Path to directory with custom Dockerfile
    container_name: crawlab_master
    restart: always
    environment:
      CRAWLAB_NODE_MASTER: "Y"
      CRAWLAB_MONGO_HOST: "mongo"
      CRAWLAB_MONGO_PORT: "27017"
      CRAWLAB_MONGO_DB: "crawlab"
      CRAWLAB_MONGO_USERNAME: "username"
      CRAWLAB_MONGO_PASSWORD: "password"
      CRAWLAB_MONGO_AUTHSOURCE: "admin"
      OLLAMA_HOST: "http://host.docker.internal:11434"
      PLAYWRIGHT_BROWSER_PATH: "/usr/bin/google-chrome-stable"
    volumes:
      - "/opt/.crawlab/master:/root/.crawlab"
      - "/opt/crawlab/master:/data"
      - "/var/crawlab/log:/var/log/crawlab"
  #     - "/usr/bin/google-chrome-stable:/usr/bin/google-chrome-stable"  # Chrom
    ports:
      - "8080:8080"
    depends_on:
      - mongo
    extra_hosts:
      - "host.docker.internal:host-gateway"

  mongo:
    image: mongo:4.2
    restart: always
    environment:
      MONGO_INITDB_ROOT_USERNAME: "username"
      MONGO_INITDB_ROOT_PASSWORD: "password"
    volumes:
      - "/opt/crawlab/mongo/data/db:/data/db"
    ports:
      - "27017:27017"
```

docker-compose up -d

okay now that this is installed

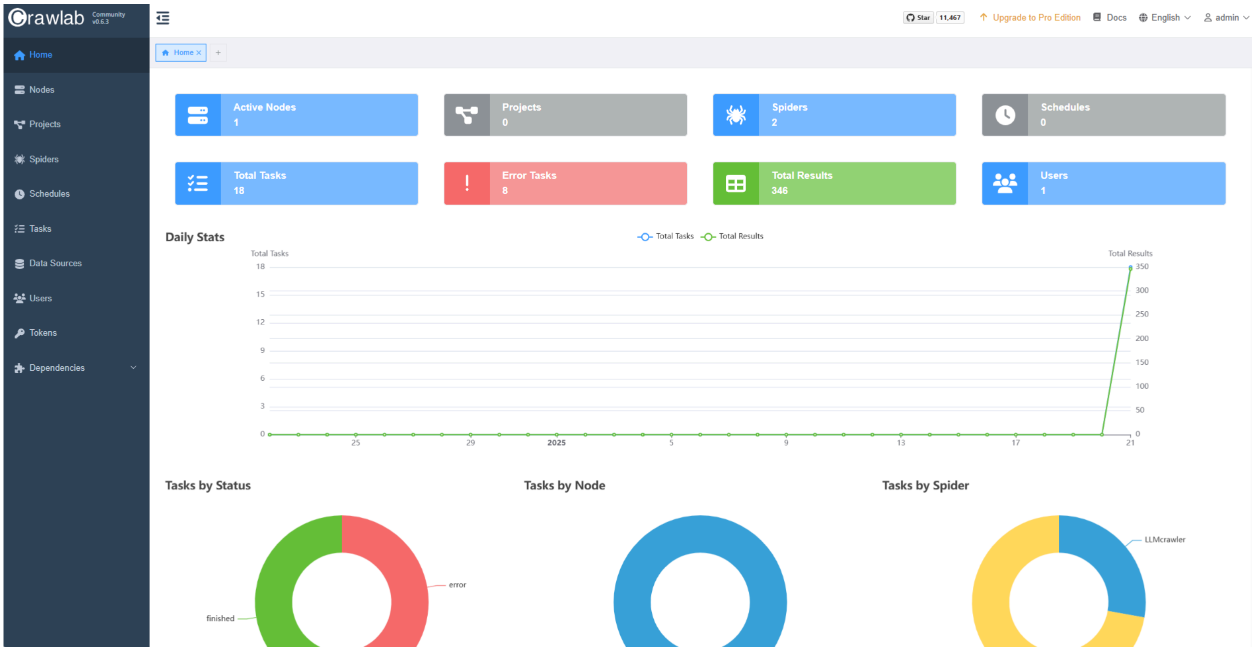
you make the scrapy project locally

and then you make a spider in crawlab

and then you upload all your local project files to crawlab

and then you run it on crawlab

important note - you have to give a command “scrapy crawl web” and web should be the name of your spider



Active Nodes 1

Projects 0

Spiders 2

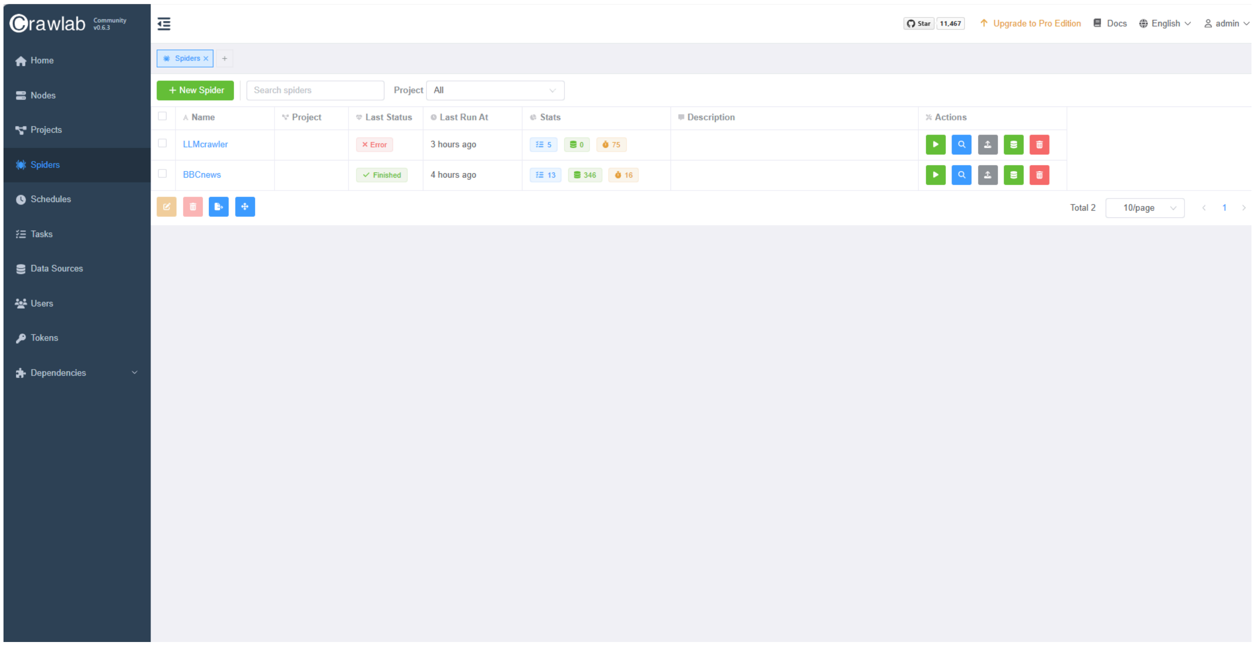
Schedules 0

Total Tasks 18

Error Tasks 8

Total Results 346

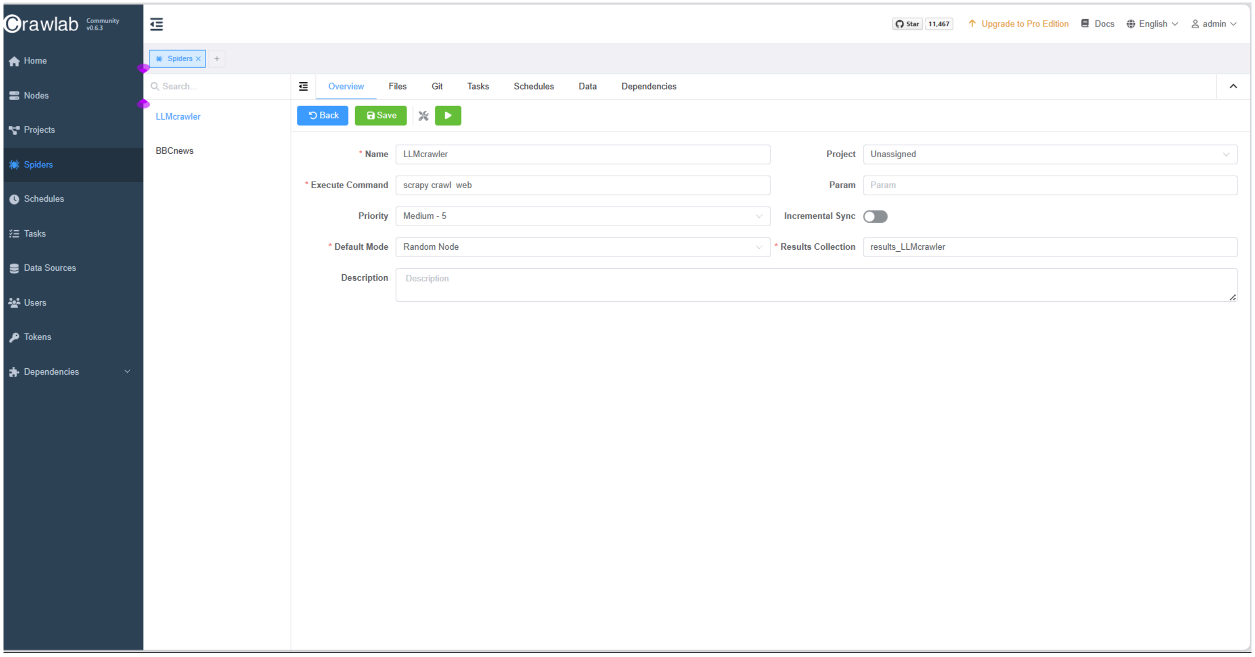
Users 1



Project All

□	Name	Project	Last Status	Last Run At	Stats	Description	Actions
□	LLMcrawler		✕ Error	3 hours ago	5 0 75		
□	BBCnews		✓ Finished	4 hours ago	13 346 16		

Total 2 10/page 1



* Name: LLMcrawler

* Execute Command: scrapy crawl web

Priority: Medium - 5

* Default Mode: Random Node

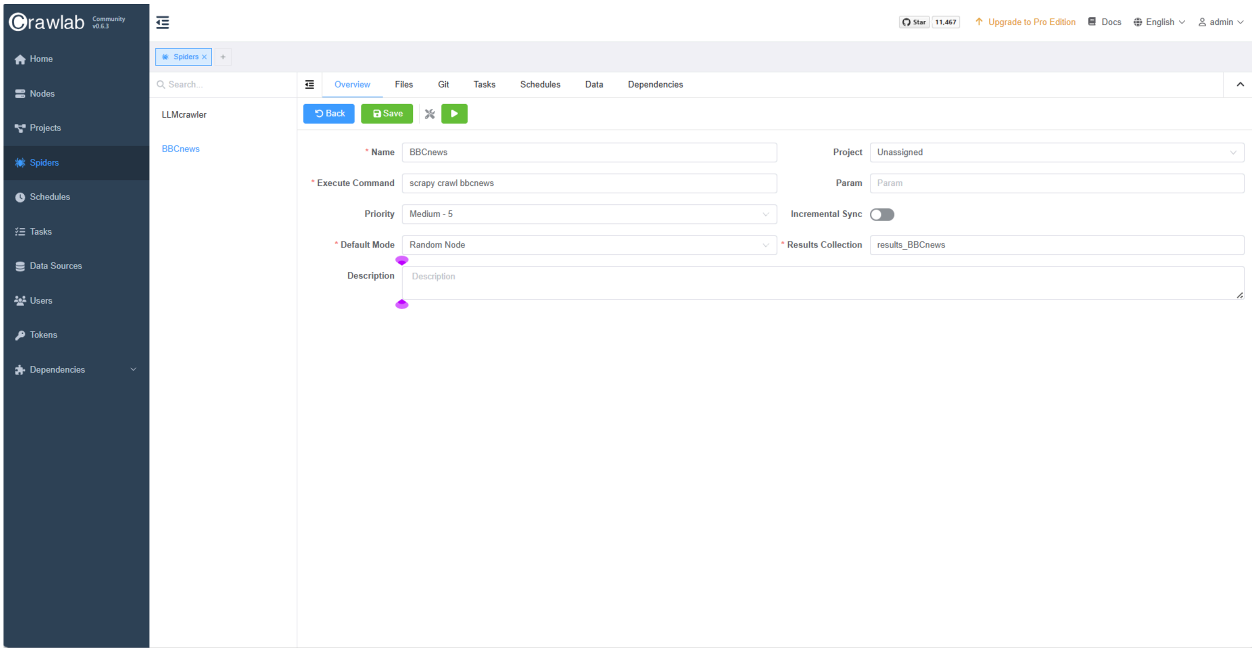
Description: Description

Project: Unassigned

Param: Param

Incremental Sync: ○

* Results Collection: results_LLMcrawler



* Name: BBCnews

* Execute Command: scrapy crawl bbcnews

Priority: Medium - 5

* Default Mode: Random Node

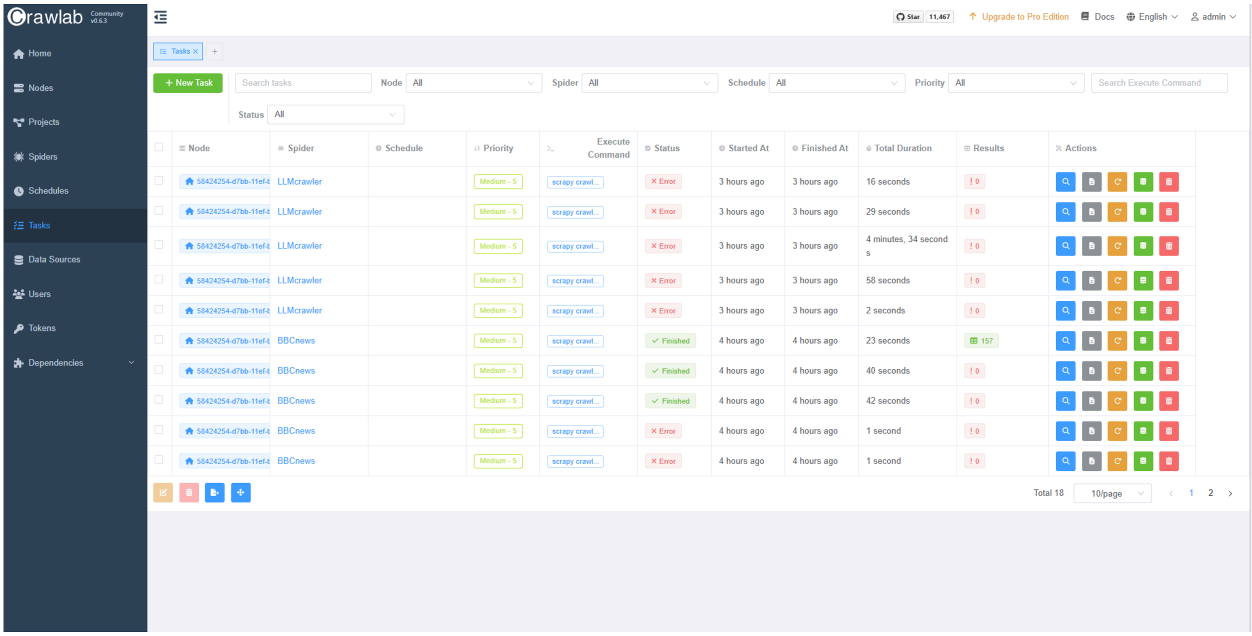
Description: Description

Project: Unassigned

Param: Param

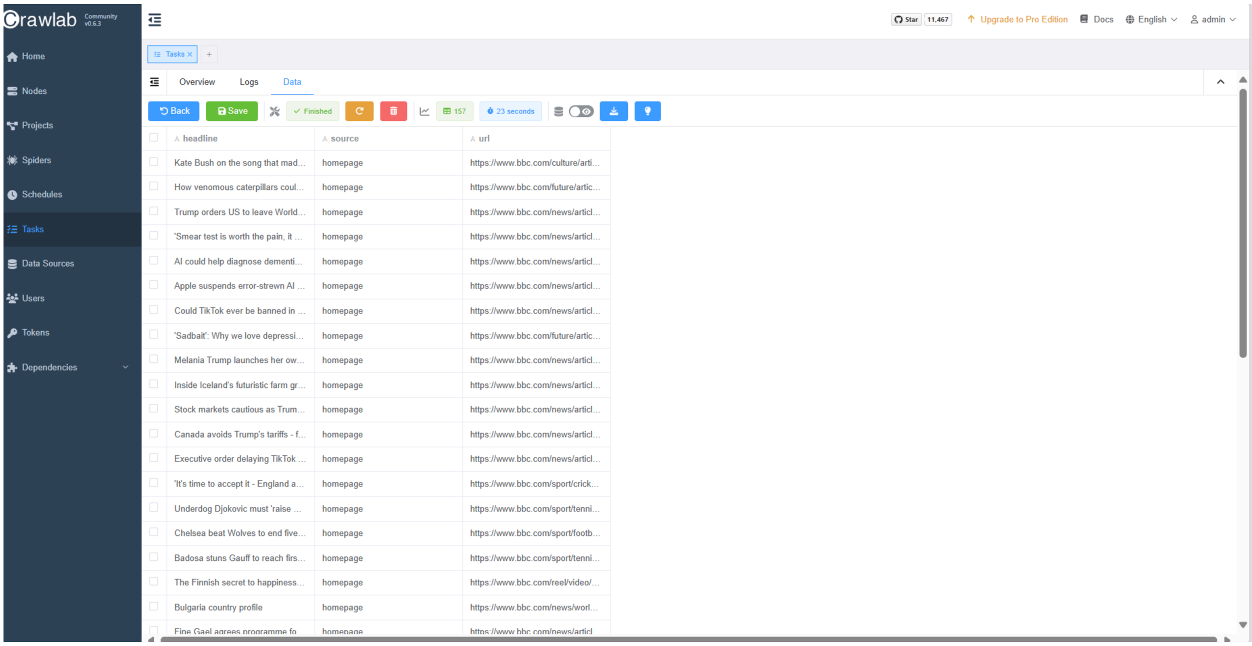
Incremental Sync: ○

* Results Collection: results_BBCnews

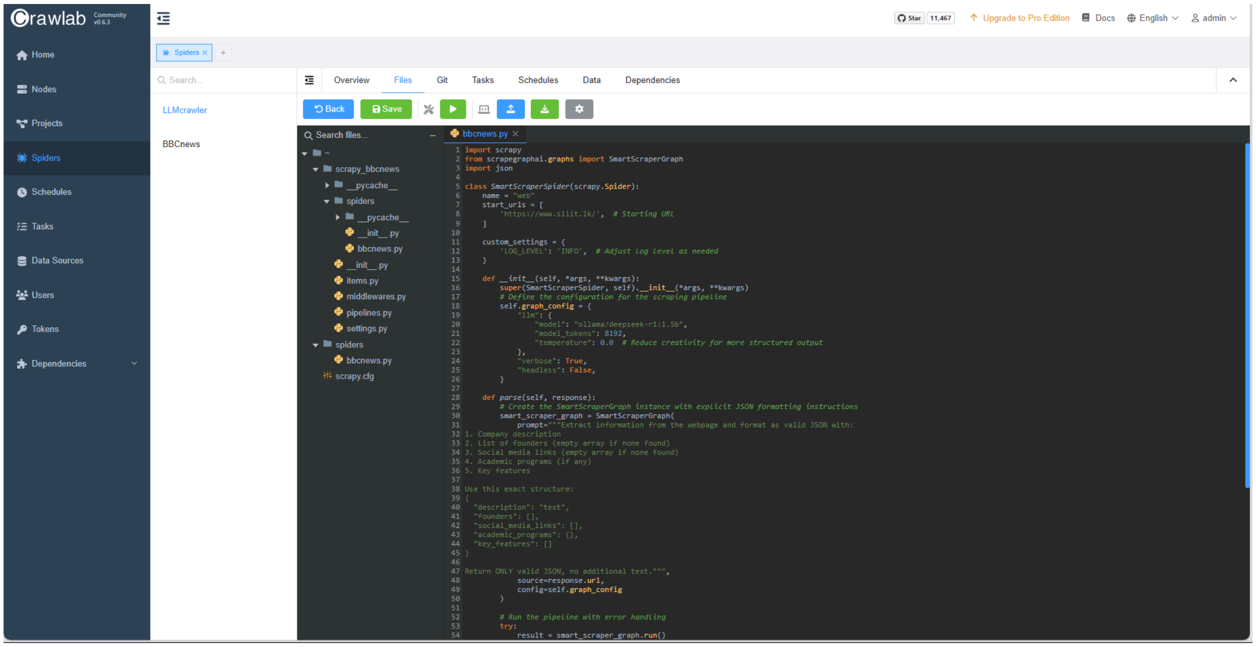


Node	Spider	Schedule	Priority	Execute Command	Status	Started At	Finished At	Total Duration	Results	Actions
58424254-d7bb-11ef-t	LLMcrawler		Medium - 5	scrapy crawl...	✕ Error	3 hours ago	3 hours ago	16 seconds	! 0	
58424254-d7bb-11ef-t	LLMcrawler		Medium - 5	scrapy crawl...	✕ Error	3 hours ago	3 hours ago	29 seconds	! 0	
58424254-d7bb-11ef-t	LLMcrawler		Medium - 5	scrapy crawl...	✕ Error	3 hours ago	3 hours ago	4 minutes, 34 seconds	! 0	
58424254-d7bb-11ef-t	LLMcrawler		Medium - 5	scrapy crawl...	✕ Error	3 hours ago	3 hours ago	58 seconds	! 0	
58424254-d7bb-11ef-t	LLMcrawler		Medium - 5	scrapy crawl...	✕ Error	3 hours ago	3 hours ago	2 seconds	! 0	
58424254-d7bb-11ef-t	BBCnews		Medium - 5	scrapy crawl...	✓ Finished	4 hours ago	4 hours ago	23 seconds	157	
58424254-d7bb-11ef-t	BBCnews		Medium - 5	scrapy crawl...	✓ Finished	4 hours ago	4 hours ago	40 seconds	! 0	
58424254-d7bb-11ef-t	BBCnews		Medium - 5	scrapy crawl...	✓ Finished	4 hours ago	4 hours ago	42 seconds	! 0	
58424254-d7bb-11ef-t	BBCnews		Medium - 5	scrapy crawl...	✕ Error	4 hours ago	4 hours ago	1 second	! 0	
58424254-d7bb-11ef-t	BBCnews		Medium - 5	scrapy crawl...	✕ Error	4 hours ago	4 hours ago	1 second	! 0	

Total 18 10/page ‹ 1 2 ›



headline	source	url
Kate Bush on the song that mad...	homepage	https://www.bbc.com/culture/arti...
How venomous caterpillars coul...	homepage	https://www.bbc.com/future/artic...
Trump orders US to leave World...	homepage	https://www.bbc.com/news/articl...
'Smear test is worth the pain, it ...	homepage	https://www.bbc.com/news/articl...
AI could help diagnose dementi...	homepage	https://www.bbc.com/news/articl...
Apple suspends error-strewn AI ...	homepage	https://www.bbc.com/news/articl...
Could TikTok ever be banned in ...	homepage	https://www.bbc.com/news/articl...
'Sadbait': Why we love depressi...	homepage	https://www.bbc.com/future/artic...
Melania Trump launches her ow...	homepage	https://www.bbc.com/news/articl...
Inside Iceland's futuristic farm gr...	homepage	https://www.bbc.com/news/articl...
Stock markets cautious as Trum...	homepage	https://www.bbc.com/news/articl...
Canada avoids Trump's tariffs - f...	homepage	https://www.bbc.com/news/articl...
Executive order delaying TikTok ...	homepage	https://www.bbc.com/news/articl...
'It's time to accept it - England a...	homepage	https://www.bbc.com/sport/crick...
Underdog Djokovic must 'raise ...	homepage	https://www.bbc.com/sport/tenni...
Chelsea beat Wolves to end five...	homepage	https://www.bbc.com/sport/footb...
Badosa stuns Gauff to reach firs...	homepage	https://www.bbc.com/sport/tenni...
The Finnish secret to happiness...	homepage	https://www.bbc.com/reel/video/...
Bulgaria country profile	homepage	https://www.bbc.com/news/worl...
Fine Gael agrees programme fo	homepage	https://www.bbc.com/news/articl



```python
import scrapy
from scrapegraphai.graphs import SmartScraperGraph
import json

class SmartScraperSpider(scrapy.Spider):
    name = "web"
    start_urls = [
        'https://www.sliit.lk/',  # Starting URL
    ]

    custom_settings = {
        'LOG_LEVEL': 'INFO',  # Adjust log level as needed
    }

    def __init__(self, *args, **kwargs):
        super(SmartScraperSpider, self).__init__(*args, **kwargs)
        # Define the configuration for the scraping pipeline
        self.graph_config = {
            "llm": {
                "model": "ollama/deepseek-r1:1.5b",
                "model_tokens": 8192,
                "temperature": 0.0  # Reduce creativity for more structured output
            },
            "verbose": True,
            "headless": False,
        }

    def parse(self, response):
        # Create the SmartScraperGraph instance with explicit JSON formatting instructions
        smart_scraper_graph = SmartScraperGraph(
            prompt="""Extract information from the webpage and format as valid JSON with:
1. Company description
2. List of founders (empty array if none found)
3. Social media links (empty array if none found)
4. Academic programs (if any)
5. Key features

Use this exact structure:
{
  "description": "text",
  "founders": [],
  "social_media_links": [],
  "academic_programs": {},
  "key_features": []
}

Return ONLY valid JSON, no additional text.""",
            source=response.url,
            config=self.graph_config
        )

        # Run the pipeline with error handling
        try:
            result = smart_scraper_graph.run()
```