Scrapy Docker Setup Documentation

# 1. Overview

This document provides the setup instructions for building and running Docker containers to run the three Scrapy spiders:  
- Amazon Spider (`amazon`)  
- Asda Groceries Spider (`asda`)  
- Curry Products Spider (`curry`)

# 2. Building the Docker Image

## Prerequisites:

- Docker should be installed and running on your machine.  
- Ensure that the project contains the `Dockerfile`, `requirements.txt`, and `scrapy.cfg` files at the root of the directory.  
- The spiders (`amazon`, `asda`, `curry`) should be placed in the `/spiders` folder.

## Steps to Build Docker Image:

1. Navigate to the project directory where the `Dockerfile` and `requirements.txt` are located.  
2. Run the following command to build the Docker image:  
 ```bash  
 docker build -t shared\_scraper .  
 ```  
 This command will:  
 - Build the Docker image using the `Dockerfile` in the current directory.  
 - Tag the image as `shared\_scraper`.  
3. Once the build is complete, the Docker image will be available for use.

# 3. Running the Scrapers

You can run any of the three Scrapy spiders (`amazon`, `asda`, or `curry`) by passing the spider name as an environment variable. Additionally, you can run all spiders concurrently.

## Running a Specific Spider by Name

To run a specific spider, use the `SPIDER\_NAME` environment variable when executing the `docker run` command.  
  
- For Amazon Spider:  
 ```bash  
 docker run -e SPIDER\_NAME=amazon shared\_scraper  
 ```  
- For Asda Spider:  
 ```bash  
 docker run -e SPIDER\_NAME=asda shared\_scraper  
 ```  
- For Curry Spider:  
 ```bash  
 docker run -e SPIDER\_NAME=curry shared\_scraper  
 ```

## Running All Spiders Concurrently

To run all the spiders (`amazon`, `asda`, and `curry`) concurrently, set the `SPIDER\_NAME` environment variable to `all`:  
```bash  
docker run -e SPIDER\_NAME=all shared\_scraper  
```

# 4. Required Runtime Arguments and Environment Variables

## Environment Variables:

- `SPIDER\_NAME`: Specifies which spider to run. If set to `all`, all spiders will be executed concurrently. Possible values are:  
 - `amazon`  
 - `asda`  
 - `curry`  
 - `all` (for running all spiders concurrently)  
  
Additional Runtime Configurations (if any):  
You may need to provide specific runtime configurations such as proxies, user agents, or authentication credentials depending on the spider and target website (e.g., Zyte proxy API keys). These configurations should be added to the `settings.py` file within each spider's project or set as environment variables.

# 5. Troubleshooting

If you encounter any issues during the build or run process, here are a few things to check:  
  
- \*\*Ensure the correct spider name is provided\*\*: Use `scrapy list` inside the Docker container to list all available spiders and verify that the correct name is passed in the `SPIDER\_NAME` environment variable.  
  
 ```bash  
 docker run -it shared\_scraper /bin/bash  
 scrapy list  
 ```  
  
- \*\*Check logs for errors\*\*: If the spider fails to start or encounters errors, view the Docker logs for more details:  
  
 ```bash  
 docker logs <container\_id>  
 ```

**Specific Requirements:  
The input folder contains the following two files:**

1. **amazon\_search\_keywords.txt: Contains the keywords to be used for scraping the Amazon website.**
2. **curry\_search\_keywords.txt: Includes the keywords for scraping the Currys website.**
3. **Zyte Api is used In Amazon and Curry’s Spiders**