

# ■ Lab Exercise Sheet – Scrapy Middlewares

## ■ Learning Objectives

- Understand what Downloader and Spider Middlewares are.
- Write custom middlewares in Scrapy.
- Enable/disable middlewares in settings.py.
- Use middlewares for request/response customization and filtering.

## Part 1 – Starter Code Setup

```
1. Create a Scrapy project: scrapy startproject quotes_middleware_demo cd quotes_middleware_demo 2. Add a spider (spiders/quotes_spider.py): import scrapy
class QuotesSpider(scrapy.Spider):
    name = "quotes"
    start_urls = [
        "http://quotes.toscrape.com/"
    ]
    def parse(self, response):
        for quote in response.xpath('//div[@class="quote"]'):
            yield {
                'text': quote.xpath('.//span[@class="text"]/text()').get(),
                'author': quote.xpath('.//small[@class="author"]/text()').get(),
            }
        next_page = response.xpath('//li[@class="next"]/a/@href').get()
        if next_page:
            yield response.follow(next_page, self.parse)
```

## Part 2 – Hands-On Tasks

### ***Task 1: Downloader Middleware – Rotate User-Agent***

Add RandomUserAgentMiddleware in middlewares.py and enable it in settings.py. ■ Question 1: Run the spider twice. Do you see different User-Agent values? Why is this useful?

### ***Task 2: Spider Middleware – Filter Short Quotes***

Add FilterShortQuotesMiddleware in middlewares.py and enable it. ■ Question 2: Run the spider. Do very short quotes appear?

### ***Task 3: Modify Middleware – Drop Einstein's Quotes***

Update FilterShortQuotesMiddleware to drop quotes by Albert Einstein. ■ Question 3: Are Einstein's quotes still present in output?

### ***Task 4: Custom Proxy Middleware (Optional Advanced)***

Write a ProxyMiddleware to apply a proxy for all requests. ■ Question 4: What real-world situations require using proxies?

### ***Task 5: Disable Built-in Middleware***

Set ROBOTSTXT\_OBEY = False in settings.py. ■ Question 5: Why does quotes.toscrape.com stop scraping if ROBOTSTXT\_OBEY is True?

## Part 3 – Reflection Questions

- Difference between Downloader and Spider Middleware?

- Can multiple middlewares work together? What determines the order?
- How would you use middleware to retry failed requests or handle captchas?

## ■ Expected Lab Outcomes

- Students run the spider and observe middleware effects on requests/responses. - Logs show filtering and User-Agent rotation. - Students understand the practical need for middlewares in real scraping (avoid blocking, clean data, handle errors).