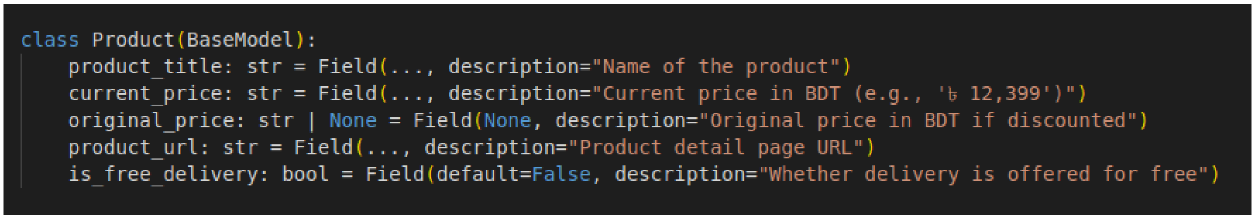
# Web Scraping Report: Daraz Smartphone Data Collection using Firecrawl

## Overview

This script implements web scraping functionality for Daraz's smartphone section using the Firecrawl service, with MongoDB for data storage and Pydantic for data validation.

## Components

1. Data Model Definition



```
class Product(BaseModel):
    product_title: str = Field(..., description="Name of the product")
    current_price: str = Field(..., description="Current price in BDT (e.g., '৳ 12,399')")
    original_price: str | None = Field(None, description="Original price in BDT if discounted")
    product_url: str = Field(..., description="Product detail page URL")
    is_free_delivery: bool = Field(default=False, description="Whether delivery is offered for free")
```

- Uses Pydantic for strict type checking
- Defines required and optional fields
- Includes field descriptions for documentation

2. Helper Functions

- to_abs(): Converts relative URLs to absolute
- text_or_none(): Safely extracts text from HTML elements
- normalize_price(): Standardizes price format (৳ symbol)
- parse_float_or_none(): Converts string ratings to floats
- unique_by_url(): Removes duplicate products

3. Firecrawl Implementation

app = FirecrawlApp(api_key=os.getenv("FIRECRAWL_API_KEY"))

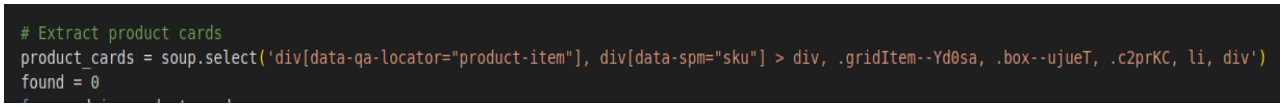
4. Data Extraction Pipeline

1. Page Fetching
    - Handles pagination (up to 10 pages)
    - Random delays (2-4 seconds) between requests

- Custom user agent and headers

2. Content Parsing

   Multiple selector fallbacks



```
# Extract product cards
product_cards = soup.select('div[data-qa-locator="product-item"], div[data-spm="sku"] > div, .gridItem--Yd0sa, .box--ujueT, .c2prKC, li, div')
found = 0
```

Extracts:
Product titles
Current prices
Product URLs
Delivery information

1. Data Processing

   - HTML parsing with BeautifulSoup
   - Price normalization
   - URL absolutization
   - Validation against Pydantic model
   - Deduplication by URL

2. MongoDB Integration

   - Database: "daraz_scraping"
   - Collection: "products3"
   - Bulk insert operations
   - Connection management

## Technical Requirements

- firecrawl
- beautifulsoup4
- pydantic
- pymongo
- python-dotenv

## Mongodb Link : Products