🧠 WEB SCRAPING — STAGE 1 (PART 1)

HONEST CHEATSHEET (WHAT YOU ACTUALLY KNOW)

0️⃣ SCRAPING KA BIG PICTURE (CLEAR HAI)

Web scraping ka matlab:

> Browser jo cheez dikha raha hai,
> Python se wahi cheez server se uthana.

Flow hamesha ye hota hai:

```
Website
 → HTML
 → DOM Tree
 → Records (parent)
 → Data (children)
 → Structure (list/dict)
 → Output
```

1️⃣ STATIC vs DYNAMIC (CLEAR)

- **Static website**
  → Data HTML me hi hota hai
  → `requests + BeautifulSoup` kaam karta hai
- **Dynamic website**
  → Data JS / XHR / API se aata hai
  → BS alone kaam nahi karta

Examples:

- `quotes.toscrape.com` → static
- `books.toscrape.com` → static

👉 Tum ab browser dekh ke ye judge kar sakte ho.

2️⃣ DEVTOOLS THINKING (CLEAR)

Tumne ye seekh liya hai:

- Inspect element ka use
- Hover karke **1 complete record** pakadna
- Parent identify karna

Golden rule:

> **1 hover = 1 record = 1 parent**

Examples:

- Quotes: `div.quote`
- Books: `article.product_pod`

3️⃣ HTML TREE MENTAL MODEL (CLEAR)

HTML = tree jaisa structure:

- Parent → repeat hota hai (record)
- Children → data fields hote hain

Tum ye soch sakte ho:

> “Pehle parent pakdo,
> phir uske andar children nikaalo.”

4️⃣ BASIC REQUEST + HEADERS (CLEAR)

```
requests.get(url, headers=headers)
```

- `url` → kaunsa page
- `headers` → tum kaun ho (identity)

Headers ka role:

- 403 se bachna
- browser jaisa behave karna

Tum ye bhi samajh chuke ho:

> Har site ko same headers nahi chahiye.

5️⃣ BEAUTIFULSOUP BASICS (CLEAR)

```
soup = BeautifulSoup(html, "lxml")
```

Meaning:

- Raw HTML → searchable DOM tree

Important methods:

- `find()` → ek cheez
- `find_all()` → list of cheezein

Tum ye bhi jaante ho:

> Parsing ka kaam fetch se alag hona chahiye.

6️⃣ EXTRACTOR THINKING (CLEAR)

Extractor ka role:

- Parent select karna
- Us parent ke andar:
  - title
  - price
  - rating
  - availability

Pattern tumhe aata hai:

```
parents = soup.find_all(parent_tag)

for parent in parents:
    child = parent.find(...)
```

Tum ye samajh chuke ho:

> Pehle record decide hota hai,
> phir loop ke andar data.

7️⃣ ATTRIBUTE vs TEXT (CLEAR)

Tumne real example dekha:

- `.text` → visible text (kabhi truncated)
- `["title"]` → full data (attribute)

Example:

```
title = tag["title"]
```

👉 Tum ab ye farq samajhte ho.

8️⃣ STAR RATING LOGIC (CLEAR)

HTML:

```
<p class="star-rating Three"></p>
```

BeautifulSoup:

```
["star-rating", "Three"]
```

Extraction:

```
rating = tag["class"][-1]
```

Tum ye samajh chuke ho:

> Kabhi data text me nahi hota,
> kabhi class ke naam me hota hai.

9️⃣ PAGINATION THINKING (CLEAR)

Pagination ka matlab:

> “Next page exist karta hai ya nahi?”

Pattern:

- `li.next`
- `a[href]`
- relative link → full URL banana

Tum ye jaante ho:

- Pagination = **flow control**
- Data extraction ka kaam nahi

🔟 FIVE ROLES MODEL (VERY CLEAR)

Tumne **system ko roles me todna seekh liya** —
ye sabse badi win hai.

1. Network

- Kaam: HTML laana
- Nahi karna: parse / extract / paginate

2. Parser

- Kaam: HTML → DOM tree
- Nahi karna: fetch / extract / paginate

3. Extractor

- Kaam: DOM se data nikaalna
- Nahi karna: fetch / parse / pagination

4. Flow / Pagination

- Kaam: decide next page / stop
- Nahi karna: fetch / parse / extract

5. Output

- Kaam: final data save / return / print
- Nahi karna: process control

👉 **Ye mental model tumhe aa chuka hai.**

1️⃣1️⃣ WHAT YOU DO NOT NEED TO MEMORISE (CLEAR)

Tumhe ye **ratna nahi** hai:

- Exact syntax
- Poora class structure
- Boilerplate code

Tumne ye seekh liya hai:

> **Human = WHAT + WHY**
> **AI = HOW**

1️⃣2️⃣ QUOTES SCRAPER CLASS — STATUS

Jo tumhe AA GAYA HAI ✅

- Kyun class bani
- Kyun methods alag-alag hain
- Kaunsa method kis role ka hai
- Fetch → Parse → Extract → Flow sequence

Jo ABHI HALF-CLEAR / UNCLEAR HAI ❌ (KAL)

- Orchestration vs Output ka fine difference
- Error handling ka decision kaun leta hai
- “Crash vs continue” logic

👉 **Ye kal ka kaam hai.**

1️⃣3️⃣ IMPORTANT TRUTH (PLEASE REMEMBER)

Tumhara problem **learning ka nahi tha**.
Tumhara problem tha:

> “Mujhe lag raha tha mujhe sab khud likhna chahiye.”

Tumne ye realise kar liya:

> **Thinking > typing**

Ye 2026 ka correct mindset hai.