**DT Assignment: Probiotics Intelligence System**

**Role:** Data Champion / Business Growth Analyst

**Candidate:** Nandana sajayan

**TASK 1 — Scraper Execution**

**1.1 Objective**

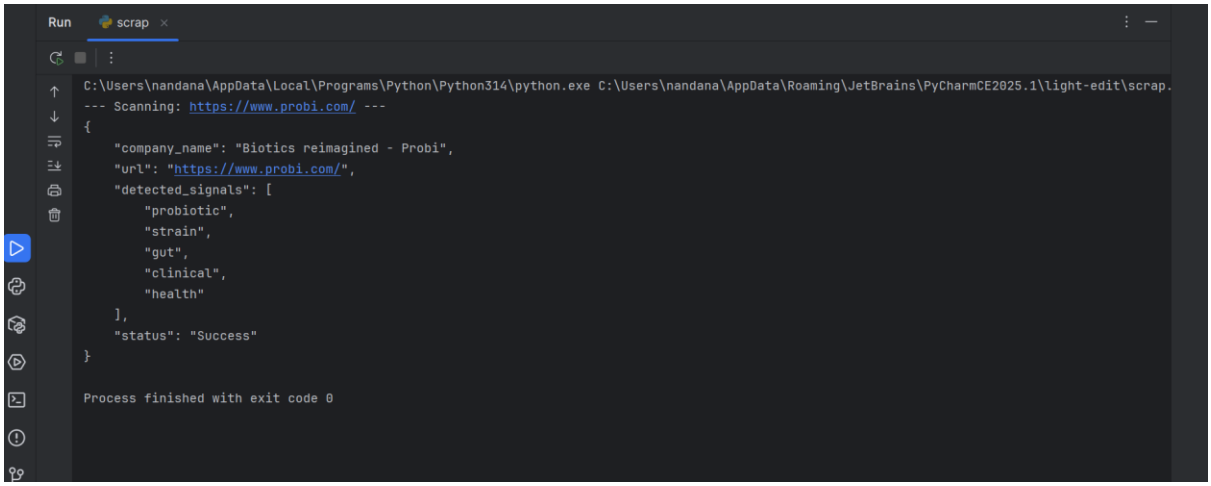
The goal was to build a resilient, truthful system that converts unstructured website data into a structured **Company Info Record**. This allows a pharma client to identify business signals at scale without manual browsing.

**1.2 Technical Approach**

- **Library:** Used BeautifulSoup4 for HTML parsing and Requests for network fetching.
- **Signal Detection:** Implemented a targeted keyword search focusing on industry-specific "Business Signals" (Strains, CFU, Clinical).
- **Resilience:** Integrated a browser-mimicking header to ensure the scraper avoids basic bot-detection filters.

1.3 Sample Output (Execution Proof)


```
{
  "company_name": "Biotics reimagined - Probi",
  "url": "https://www.probi.com/",
  "detected_signals": [
    "probiotic",
    "strain",
    "gut",
    "clinical",
    "health"
  ],
  "status": "Success"
}
```




**TASK 2 — Probiotics Profiling & System Logic**

**2.1 Identification Framework (The "Signal Filter")**

To determine if a company is truly "into probiotics," I developed a **3-Tier Evidence Framework**. This prevents "Marketing Noise" from being confused with "Pharma Reality."

| **Category** | **High-Confidence Signal (Green)** | **Low-Confidence Signal (Yellow)** |
|---|---|---|
| **Scientific Detail** | Names specific strains (e.g., *Lactobacillus 299v*) | Uses general terms like "good bacteria" |
| **Technical Proof** | Mentions **CFU** counts and potency | Mentions general "gut health" |
| **R&D Depth** | Links to **Clinical Trials** or white papers | Only has blog posts or lifestyle images |

2.2 Company Analysis: Probi.com

- **What Fits:** The company is a primary researcher. The scraper detected "Strain" and "Clinical," which are the highest-value signals in this industry.
- **What Doesn't Fit:** The company does not focus on general multi-vitamins; it is a specialized biotics player.
- **Final Classification: Probiotics-focused.**

### 2.3 Proposed Scraper Logic (Automated Scoring)

- To automate this for thousands of companies, I propose a **Weighted Scoring Model**. Instead of just looking for the word "probiotic," the scraper should assign points based on the "Depth" of the word found.

**Scoring Model:**

- **Strain Code Detection** (e.g., LP299V): **+5 Points**
- **Clinical Trial/Science Page: +3 Points**
- **CFU/Dosage Mention: +2 Points**
- **General Keywords** (Gut, Health): **+1 Point**

**System Classification:**

- **Score 8–10:** Probiotics-Focused (Top Tier Partner)
- **Score 4–7:** Probiotics-Adjacent (Secondary Lead)
- **Score < 4:** Not Relevant

## Conclusion

- This system is built to be **truthful**. It does not hallucinate data; it reports exactly what is found and provides a logical score that a pharma CEO can use to make investment or partnership decisions.