IR Project

Members:

Mahmoud Ibrahim al Molla 224417

Youssef Ahmed Hany 220427

Ziad Mohamed 224875

Crawlability Summary

Amazon.com

1-Robots.txt File Analysis

General Info:

**File Location**: <https://www.amazon.com/robots.txt>

Disallowed Paths:

 **Account & Login** pages  
e.g., /exec/obidos/account-access-login, /gp/sign-in, /ap/signin

 **Cart & Checkout**  
e.g., /gp/cart, /dp/shipping/, /gp/flex

 **Review & Rating Systems**  
e.g., /gp/customer-reviews/write-a-review.html, /review/dynamic/sims-box

 **Wishlist**  
General disallow: /wishlist/, /gp/wishlist/, /registry/wishlist/

 **Digital Media / Video / Music**  
e.g., /gp/video/, /gp/dmusic/, /gp/video/library, etc.

 **Other UX/Non-essential content**  
Includes sitemaps, redirectors, voting tools, legacy APIs, etc.

2-CrawlabilityChecker Code Review & testing

 Fetches and parses Amazon’s robots.txt.

 Uses Python’s built-in urllib.robotparser to check crawl permissions.

 Prints whether each tested URL is ALLOWED ✅ or BLOCKED ❌.

**Limitation (False Negatives):**

The urllib.robotparser **does not support advanced robots.txt features**, such as:

| **Feature** | **Supported by urllib?** | **Example from Amazon** | **Result** |
| --- | --- | --- | --- |
| Allow: directives | ❌ **Not reliably** | Allow: /wishlist/universal\* | ❌ False negative |
| Wildcards (\*) | ❌ | Disallow: \*/gcrnsts | ❌ Ignored |
| Rule specificity | ❌ | Allow within disallow | ❌ Misinterpreted |

Example Cases:

| **URL** | **Expected** | **Reported by Code** | **Notes** |
| --- | --- | --- | --- |
| https://www.amazon.com/dp/B09G3HRMVP | ✅ | Allowed | Product pages are allowed |
| https://www.amazon.com/wishlist/universal | ✅ | ❌ BLOCKED | False negative (Allowed by robots.txt) |
| https://www.amazon.com/gp/customer-reviews/write-a-review.html | ❌ | ❌ BLOCKED | Correctly blocked |

Content Extraction :

JS/API Handling :

**How to Know if a Website is JS-Heavy**

**Code output:**

Titles found in raw HTML: 19

- 1-16 of over 100,000 results for

-

- "smartphones"

**Interpretation:**

* The "titles" you're seeing are **not actual product titles** — they are layout labels or page metadata.
* You're missing product data like:
  + Product **names**
  + **Prices**
  + **Ratings**
  + **Images**

What This Means

| **Signal** | **Explanation** |
| --- | --- |
| ✅ Found < 20 <h2> or <span> tags | Not a lot of data for a page claiming 100,000+ results |
| ❌ No product titles in raw HTML | JS is used to inject product listings |
| ❌ No pricing data visible | Often fetched via JS from backend API |
| ✅ Amazon uses anti-bot + obfuscation | HTML is intentionally cleaned to prevent scraping |

**Amazon is JavaScript-heavy.**  
You need to use **Playwright or Selenium** to get fully rendered content with products.