**Day – 6 (task1)**

**1.Summary of robots.txt: Web Scraping Limits**

A **robots.txt** file is used to control how web crawlers and bots interact with a website. It specifies which parts of the site are accessible and which are restricted. The provided **robots.txt** files include rules from **Amazon, ESPNcricinfo, Instagram, and other websites**, setting various restrictions.

**Key Web Scraping Restrictions:**

1. **Amazon Website:**
   * Blocks access to login, cart, checkout, order management, and user profile pages.
   * Disallows pages related to customer reviews, voting, and media uploads.
   * Restricts certain pages related to Amazon Music and Amazon Video.
   * **Specifically disallows AI-based bots like GPTBot, PerplexityBot, ClaudeBot, and Google-Extended.**
2. **ESPNcricinfo:**
   * Blocks access to print versions of articles, error pages, fragments, and various site resources.
   * Prevents scraping of certain archived data, images, videos, and some series-specific content.
   * Restricts access to advertisements, support files, and interactive applications.
3. **Instagram:**
   * Blocks automated data collection unless explicitly authorized by Meta.
   * Prohibits access to login, AJAX requests, direct messages, and private API endpoints.
   * Restricts access for bots like Googlebot, Bingbot, Twitterbot, LinkedInBot, and several others.
   * Contains multiple **sitemap links** to guide search engines on permissible indexing.
4. **General Search Engine & Web Scraping Bot Restrictions:**
   * Many well-known bots, including Googlebot, Bingbot, Applebot, and Yandex, have restrictions on accessing login and user-related pages.
   * AI-based bots (e.g., GPTBot, ClaudeBot, PerplexityBot) are explicitly **blocked** from scraping content.
   * Some websites allow specific bots to access only certain parts of their content.

**Overall Impact on Web Scraping:**

* **Restricted Areas:** User accounts, login pages, payment processes, and private data sections are blocked to prevent unauthorized data collection.
* **Allowed Sections:** Some websites still allow access to publicly available data such as **product listings, general articles, or sitemaps** for search indexing.
* **AI & Data Collection:** Many sites now actively **block AI-powered bots** from scraping their content.

If you're considering web scraping, always check the **robots.txt** file of the target website and ensure compliance with legal and ethical guidelines.

import requests

# List of websites to check

websites = [

"https://www.amazon.com/robots.txt",

"https://www.espncricinfo.com/robots.txt",

"https://www.instagram.com/robots.txt"

]

# Loop through each website and fetch robots.txt

for url in websites:

print(f"\n Checking: {url}")

response = requests.get(url)

if response.status\_code == 200: # If request is successful

print(response.text[:500]) # Print first 500 characters for preview

else:

print(f" Could not fetch robots.txt for {url}")