Task: Extract Information from Websites (use any 100 websites)

Approach and Methodology

1. Web Scraping

We used the requests library to fetch web pages and BeautifulSoup to parse the HTML content. For each website, the following data is extracted:

- **Meta Information**: Title and description tags are extracted using BeautifulSoup.
- **Social Media Links**: Links to social media profiles are identified by checking href attributes.
- **Technology Stack**: The presence of specific keywords and script sources helps identify technologies.
- **Payment Gateways**: Keywords are checked within the text content.
- **Website Language**: The lang attribute of the html tag is extracted.
- **Category**: Based on keywords in the meta title and description, the website is categorized.

2. Error Handling

The script incorporates robust error handling to manage various potential issues:

- **Network Errors**: Implemented retries with exponential backoff to handle network-related errors. This is exemplified by the errors encountered and the subsequent retries for websites like www.uol.com.br and www.dailymail.co.uk.
- **HTTP Errors**: Managed HTTP errors by checking status codes and implementing appropriate retries for errors such as 403 Forbidden and 503 Service Unavailable.
- **General Exceptions**: Captured to log any other unexpected issues, ensuring the scraper does not crash unexpectedly.

Example error handling output:

```
Error scraping https://www.uol.com.br:
HTTPSConnectionPool(host='www.uol.com.br', port=443): Max retries exceeded
with url: / (Caused by ConnectTimeoutError)
Retrying in 5 seconds...
Error scraping https://www.espn.com: 403 Client Error: Forbidden for url:
https://www.espn.com/
Retrying in 10 seconds...
Failed to scrape data from https://www.dailymail.co.uk
```

3. Database Connection

We used mysql.connector to interact with a MySQL database. The connection is established using provided credentials, and a stored procedure (AddWebsiteInfo) is called to insert data into the database.

4. Storing Data

Data is stored in a MySQL database with a stored procedure to handle the insertion. This ensures a clean separation of database logic and application logic.

5. Efficiency and Performance

Efforts were made to ensure the script runs efficiently:

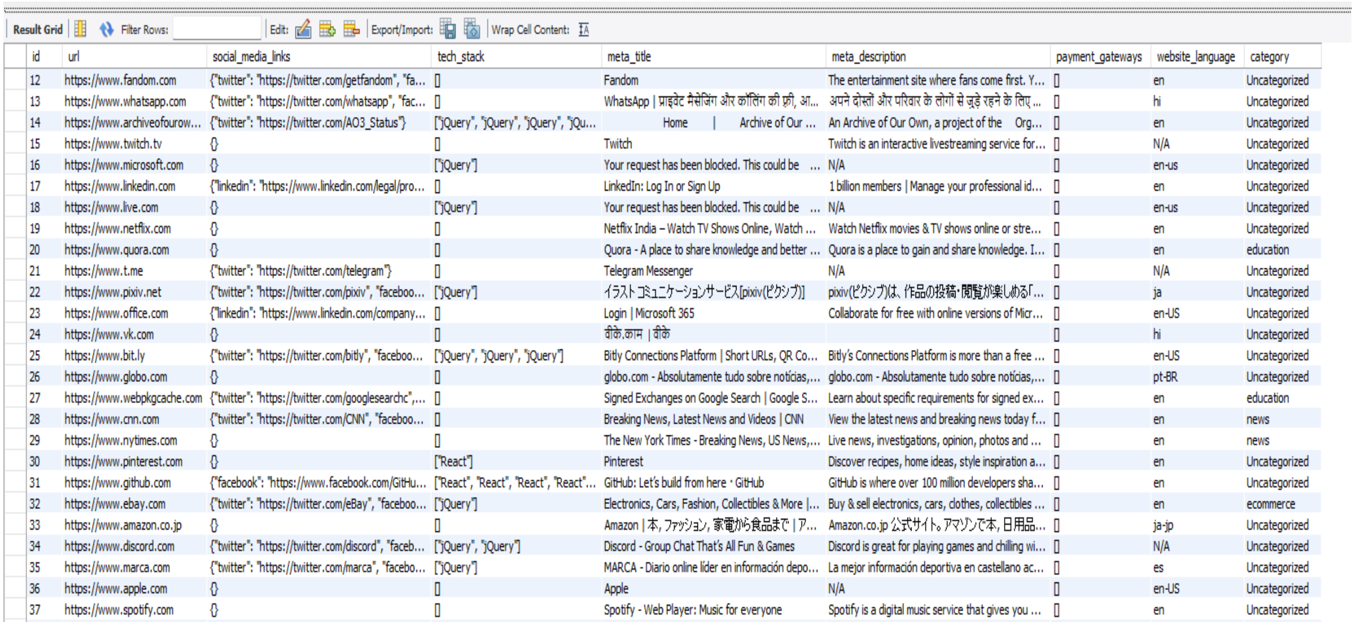
- **Batch Processing**: URLs are processed in sequence with minimal delay between retries.
- **Selective Parsing**: Only necessary parts of the HTML are parsed to reduce memory usage and processing time.

Challenges and Solutions

- **Handling Different HTML Structures**: Websites vary significantly in structure. The solution involved writing flexible and adaptive code to handle different layouts.
- **Rate Limiting and Captchas**: Some websites implemented rate limiting or captchas. We addressed this by incorporating delays and retries.
- **Data Consistency**: Ensuring that data extracted from different websites is consistently formatted and stored.

Conclusion

This web scraper effectively extracts relevant information from a diverse set of websites and stores it in a structured manner in a MySQL database. The code is organized, readable, and follows best practices, with comprehensive error handling to ensure robustness. Despite encountering various errors, such as connection timeouts and HTTP errors, the scraper is resilient and continues processing other websites.



id	url	social_media_links	tech_stack	meta_title	meta_description	payment_gateways	website_language	category
12	https://www.fandom.com	{"twitter": "https://twitter.com/getfandom", "fa...	[]	Fandom	The entertainment site where fans come first. Y...	[]	en	Uncategorized
13	https://www.whatsapp.com	{"twitter": "https://twitter.com/whatsapp", "fac...	[]	WhatsApp | प्राइवेट मैसेजिंग और कॉलिंग की फ़्री, आ...	अपने दोस्तों और परिवार के लोगों से जुड़े रहने के लिए ...	[]	hi	Uncategorized
14	https://www.archiveofourow...	{"twitter": "https://twitter.com/AO3_Status"}	["jQuery", "jQuery", "jQuery", "jQu...	Home | Archive of Our ...	An Archive of Our Own, a project of the Org...	[]	en	Uncategorized
15	https://www.twitch.tv	{}	[]	Twitch	Twitch is an interactive livestreaming service for...	[]	N/A	Uncategorized
16	https://www.microsoft.com	{}	["jQuery"]	Your request has been blocked. This could be ...	N/A	[]	en-us	Uncategorized
17	https://www.linkedin.com	{"linkedin": "https://www.linkedin.com/legal/pro...	[]	LinkedIn: Log In or Sign Up	1 billion members | Manage your professional id...	[]	en	Uncategorized
18	https://www.live.com	{}	["jQuery"]	Your request has been blocked. This could be ...	N/A	[]	en-us	Uncategorized
19	https://www.netflix.com	{}	[]	Netflix India – Watch TV Shows Online, Watch ...	Watch Netflix movies & TV shows online or stre...	[]	en	Uncategorized
20	https://www.quora.com	{}	[]	Quora - A place to share knowledge and better ...	Quora is a place to gain and share knowledge. I...	[]	en	education
21	https://www.t.me	{"twitter": "https://twitter.com/telegram"}	[]	Telegram Messenger	N/A	[]	N/A	Uncategorized
22	https://www.pixiv.net	{"twitter": "https://twitter.com/pixiv", "faceboo...	["jQuery"]	ｲﾗｽﾄ ｺﾐｭﾆｹｰｼｮﾝｻｰﾋﾞｽ[pixiv(ﾋﾟｸｼﾌﾞ)]	pixiv(ﾋﾟｸｼﾌﾞ)は、作品の投稿・閲覧が楽しめる「...	[]	ja	Uncategorized
23	https://www.office.com	{"linkedin": "https://www.linkedin.com/company...	[]	Login | Microsoft 365	Collaborate for free with online versions of Micr...	[]	en-US	Uncategorized
24	https://www.vk.com	{}	[]	वीके.काम | वीके		[]	hi	Uncategorized
25	https://www.bit.ly	{"twitter": "https://twitter.com/bitly", "faceboo...	["jQuery", "jQuery", "jQuery"]	Bitly Connections Platform | Short URLs, QR Co...	Bitly's Connections Platform is more than a free ...	[]	en-US	Uncategorized
26	https://www.globo.com	{}	[]	globo.com - Absolutamente tudo sobre notícias,...	globo.com - Absolutamente tudo sobre notícias,...	[]	pt-BR	Uncategorized
27	https://www.webpkgcache.com	{"twitter": "https://twitter.com/googlesearchc",...	[]	Signed Exchanges on Google Search | Google S...	Learn about specific requirements for signed ex...	[]	en	education
28	https://www.cnn.com	{"twitter": "https://twitter.com/CNN", "faceboo...	[]	Breaking News, Latest News and Videos | CNN	View the latest news and breaking news today f...	[]	en	news
29	https://www.nytimes.com	{}	[]	The New York Times - Breaking News, US News,...	Live news, investigations, opinion, photos and ...	[]	en	news
30	https://www.pinterest.com	{}	["React"]	Pinterest	Discover recipes, home ideas, style inspiration a...	[]	en	Uncategorized
31	https://www.github.com	{"facebook": "https://www.facebook.com/GitHu...	["React", "React", "React", "React"...	GitHub: Let's build from here · GitHub	GitHub is where over 100 million developers sha...	[]	en	Uncategorized
32	https://www.ebay.com	{"twitter": "https://twitter.com/eBay", "faceboo...	["jQuery"]	Electronics, Cars, Fashion, Collectibles & More |...	Buy & sell electronics, cars, clothes, collectibles ...	[]	en	ecommerce
33	https://www.amazon.co.jp	{}	[]	Amazon | 本, ﾌｧｯｼｮﾝ, 家電から食品まで | ア...	Amazon.co.jp 公式サイト。アマゾンで本, 日用品...	[]	ja-jp	Uncategorized
34	https://www.discord.com	{"twitter": "https://twitter.com/discord", "faceb...	["jQuery", "jQuery"]	Discord - Group Chat That's All Fun & Games	Discord is great for playing games and chilling wi...	[]	N/A	Uncategorized
35	https://www.marca.com	{"twitter": "https://twitter.com/marca", "facebo...	["jQuery"]	MARCA - Diario online líder en información depo...	La mejor información deportiva en castellano ac...	[]	es	Uncategorized
36	https://www.apple.com	{}	[]	Apple	N/A	[]	en-US	Uncategorized
37	https://www.spotify.com	{}	[]	Spotify - Web Player: Music for everyone	Spotify is a digital music service that gives you ...	[]	en	Uncategorized