

Media Monitoring using Web Crawling

Client: A popular marketing agency from the USA

Offering: Site-specific crawl and extraction

Challenge:

The client wanted to extract blog details from a blog feed reader platform. The data was required from blogs that belonged to following categories – Street Style, Fashion, Beauty, Food & Drink, Home Decor, Lifestyle and Fitness. The data points needed were Site name, URL, RSS feed URL and follower count.



Solution:

After taking the requirements from the client, our team started building the crawler to extract the required data. Since the client wanted fresh datasets every week, we set the frequency of crawl to 7 days. The data started getting accumulated once our team finished setting up the crawlers. We ran cleaning and formatting processes on the data with the dedicated systems for the same and made the data available via our API in CSV format as opted by the client. We delivered about 33k records in the first crawl.

Benefits to the client:

- The client didn't have to deal with any of the technical aspects in the process
- The setup was completed in just 3 days and the data flow was consistent since then
- We also setup monitoring for the target site to ensure consistent crawling and to avoid data loss
- Our tech stack could efficiently handle the dynamic coding practices used by the target sites
- The client could get ready to use data to power their marketing activities
- The cost of extraction was 58% less than the cost of an in-house crawling setup projected by the client