Use Cases: News Monitoring, Market Intelligence

Get Your Product to Market Faster with Diffbot

The entrepreneurs at Topic saw many of their customers struggle with creating trustworthy SEO content that ranks high in search engine results.

They realized that while many writers may be experts at crafting a compelling narrative, most are not experts at optimizing content for search. Drawing on their years of SEO expertise, this two-person team came up with an idea that would fill that gap.

They came up with Topic, an app that helps users create better SEO content and drive more organic search traffic.They had a great idea. They had a fitting name. The next step was figuring out the best way to get their product to market.

Key Takeaways

- Rule-based web extraction is hard to scale when dealing with pages from across the web
- Knowledge-as-a-service circumvents the cost of messy data and maintaining freshness in extracted data
- Extracted data is only as good as its structure and metadata

The Problem

Topic is an application that analyzes the top results in Google to surface valuable insights about your own content. The app delivers an evaluation that shows you exactly what your audience wants to learn and how to improve your content to rank higher.

The first step of that process requires reliably extracting the content from web pages displayed in the top search results. The Topic team needed a scalable and precise method to extract text and other page elements from these news articles and blog posts. They knew that building and maintaining a data extraction system in-house would be time consuming and beyond their current resources. And it would take time away from developing core features and solving problems for their customers. They also wanted to test Topic with an initial set of pilot customers as soon as possible.

The Solution

The team needed a third-party solution that would allow Topic to extract content from the top results in Google. The app also needed to identify which search results are articles and which are not, and then get the HTML and text from each result. They found the exact web extraction functionality they were looking for in Diffbot's Article Extraction API.

The Article Extraction API automatically extracts clean text from news articles and blog posts. It returns all related information from the page, including author, date, images, and videos. Diffbot also provides more than thirty sponsored client libraries making integration easy and seamless.

Products Used

- Diffbot's Article Extraction API uses computer vision to extract normalized HTML, content, and a range of metadata fields from content in any language. No rules or training are required for this API to extract unstructured web data and return it in structured JSON. Natural language processing provides subject tags from content in any language. While sentiment analysis gauges the tone of comments on the articles extracted.

Extraction Challenges

Many data teams spend up to 80% of their time and resources aggregating and cleansing data that may be messy, unstructured, outdated, or downright inaccurate.

While the team did look at other web extraction solutions, most of the solutions rely on the user to create a script or train the system to extract the desired data from a web page- tasks that can be very time consuming. However, Diffbot APIs are automatic; **no rules or training required.**

The Results

The team considered building a web scraping system from scratch, but it would have taken at least four weeks of development up front. Also, an in-house system would require one or two days of maintenance each month - a recurring cost. With Diffbot, they didn't have to spend weeks building a web extraction system from the ground up. All they had to do was create an account, install the gem, and add a line of code to their codebase. The integration took a single day, and the API simply runs in the background; there is no maintenance.

Faster Time To Market

Thanks to Diffbot, the two entrepreneurs were able to bring Topic to market within a month. And they don't have to spend any time scaling or maintaining the article extraction system. They instead spend that extra time developing exciting features that grow their customer base."The Article Extraction API has been reliable, and the data that Diffbot is retrieving provides value to our customers with every content brief that is created," said Ryo Chiba, Topic Co-Founder. "Diffbot also helped us bring our product to market in under a month. Any company looking for a reliable content extraction solution should look no further."

A Structured Web

Diffbot's Automatic Extraction APIs, Knowledge Graph, and Enhance products provide access to the largest collection of web data commercially available.

Our AI-enabled extraction and processing systems parse the web into contextually linked entities including organizations, products, people, articles, and more.

While clients can point automatic extraction APIs to sites of their choosing, Diffbot also crawls the entirity of the web and parses it into semantic data every 3-5 days.

Currently this data is available in the Knowledge Graph in the form of over 10 billion entities and 3 trillion facts.

Want to find out how Diffbot can help you get your app to market faster?
Start a Free Trial Today or contact us at sales@diffbot.com