

Geographically-sparse Feeds Aggregation

Client: Social media intelligence house

Challenge:

The Client was looking to get news feeds and [social media data scattered](#) in various geographic locations coming from more than 5000 sources. They wanted this data to be delivered in a structured collated format which they could simply import every week. Earlier they tried it in-house, but results were unsatisfactory as data was lacking in both quality and quantity. Client reported that geographic location associated with feed was incorrect in numerous cases. Moreover, they also wanted this data to be searchable for more than 1000 set of keywords and specific queries.

The Solution:

We addressed this requirement by setting up a [mass scale crawl](#), that enabled crawling numerous sources in parallel at regular periodic intervals in a day, still adhering to the politeness policies by not excessively hitting the servers of these sources. [Feeds from various social media](#) were aggregated intelligently by developing a Geo-Intelligence API, that assured feeds were captured only from desired locations. List of locations, sources, keywords and queries was dynamically modified based upon the client requirements and feedback. Over 2,00,000 feeds were collected from various continents within 2 months of time. Every week fresh data is collated location-wise and delivered.

Benefits to the client:

- Parallel collection of data from numerous sources without any infrastructural concerns
- Uniform Data schema irrespective of number of sources and heterogeneity of content
- Periodic delivery of fresh data, reducing further data processing efforts
- Geo-Intelligence API assuring data to be belonging to the described geography
- Scalable solution with increasing number of sources, locations and keywords