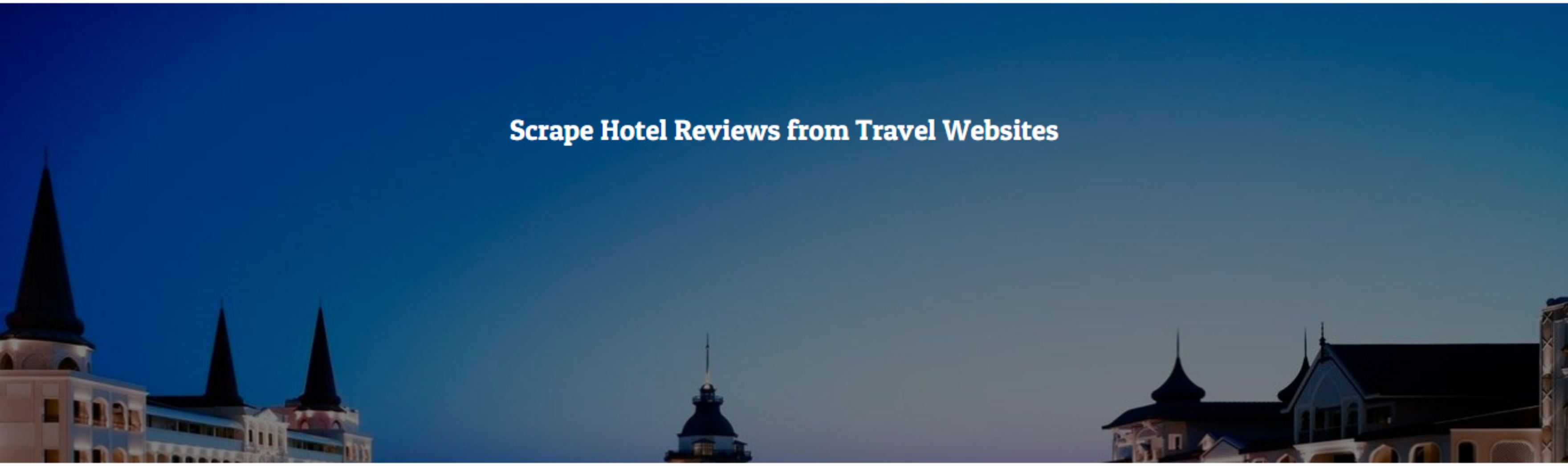

Scrape Hotel Reviews from Travel Websites

Domain: Travel

Clients: Social Travel Engine

Challenge:

The client was looking to build one of the web's largest review database by aggregating scattered reviews on hotels and destinations across multiple sources. They had tried few solutions around crawling but issues started creeping in as data scaled, given they needed new data regularly. Also the number of sources were increasingly exponentially on the web and so was the data. Additionally, they wanted reviews from all countries in all languages and the author profiles, images, etc. from the web pages.

Solution:

All historical data from each source (~100) was extracted in parallel with incremental data as reviews were published. Data was de-duped before delivery so only new data got uploaded. Machine learning techniques were employed for adaptive crawling thereby crawling the more active pages more often than others. Site list was dynamically modified based on client requirements. Over 20 million structured records were delivered in a period of 2 months.

Benefits to the client:

- Scalable platform took care of high data volumes without affecting data quality
- Development and maintenance costs dropped to zero
- Abstracted clients from technical specifics
- Having only relevant data helped the client gain credibility in the market and rocketed growth figures

One of our Slideshare decks discusses generic travel use cases.