

The logo for Zyte, featuring the word "zyte" in a lowercase, bold, sans-serif font. The letters are white with a slight shadow effect, making them stand out against the dark blue background. The background of the entire page is a dark blue gradient, with a prominent vertical bar on the right side that transitions from a lighter blue at the top to a darker blue at the bottom. This bar is overlaid with a series of vertical white lines of varying heights, creating a stylized bar chart or data visualization effect. The lines are spaced out and vary in length, with some reaching near the top of the page and others being shorter. The overall aesthetic is clean, modern, and data-oriented.

ZYTE PAPER

Generating alpha from web data

An inside look at how web
scraped alternative financial
data is used to inform
investment decision-making.

Introduction

The demand for alternative data is surging, due to its availability, scale, and striking ability to generate alpha.

Hedge funds and investors are increasingly incorporating alternative datasets into their workflows to inform the investment decision-making process. Among them, web data has consistently proven the most powerful and popular.

Web data is the #1 source of alternative data, helping to inform the world's best asset managers of market opportunities and arming them with the insight needed to act quickly and develop positions carefully.

When it comes to extracting high-quality data from the web, there's virtually no limit to the type and quantity of data available. The once-esoteric realm of alternative data is now increasingly essential and central to modern investment models.

Greenwich Associates research suggests 50% of institutional investors aim to increase their consumption of alternative data in 2019, and while available datasets range widely in both type and size, web data remains the most popular.

Jefferies stated in 2017 that 20% of hedge funds with over \$1 billion in AUM have at least one analyst dedicated to alternative data, and the Financial Times reports that investment groups have quadrupled their number of alternative data analysts in the past five years.

The uncovering of trading signals and opportunities for alpha buried in today's data-rich environment has triggered a "coming-of-age" for web data, as well as kicked off an alternative data arms race between investors. In this guide, we'll explain several types of web data available for scraping and how they can be used to inform the investment decision-making process.



Type 1: Product data

Scraping product data from complex online marketplaces is a difficult task, but one providing huge insights into a number of factors critical for evaluating company fundamentals and stock performance.

Product data can be scraped from the web from a number of sites, from the familiar to the obscure; Amazon, Best Buy, eBay, Walmart, and Target are great examples of sites from which thousands of data points can be scraped to reveal company and category performance.

Use Cases

Using product data scraped from online retailers yields a great deal of actionable information.

A product's rank on an online retailers bestseller list can indicate strong performance, while poor ranking and a discount history revealing early price reductions can expose inventory troubles and margin reductions.

This data unlocks lucrative opportunities for investors when determining market orders and positions, while also providing insights into long term trends.

Actionable Insights



YoY share trends from online retailers are powerful predictive indicators of revenue momentum and stock performance.



Aggregated pricing histories can allude to problems with inventory and signal trouble or demonstrate improved sell-through trends.



Data related to frequency and timing of online search data for a product or company provide strong signals for product sales, especially when observed against competitor averages.



Advertising data can provide unique insight into the corporate spend behind a product launch and anticipate revenue changes or slow-moving inventory, along with measurements for product traction and adoption.

Type 2: SEC filing data

The demand for web scraped SEC filing data is extraordinary.

Taking a dive into the lengthy and oft nebulous pages of a company's SEC filings can lead to amazing investment insights. It arms investors with high-quality, reliable information — data already held to strict standards by the US government.

Now, by scraping SEC filing information, the discovery process familiar to investors can be replicated tens of thousands of times across an unlimited amount of filings, unearthing precious alpha in otherwise unseen places.

It's crucial to interpret company information in a framework of its utility. To echo David Blitz at Robeco Institutional Asset Management, "the data that we find is most impactful for our quantitative investment team includes thousands of companies, compared against each other, on a plethora of points." These critical points are made available by scraping SEC filings, a process yielding data nearly impossible to acquire otherwise.

Use Cases

Most investors wouldn't want to read a 256-page, 110,000 word annual report, like that of the infamous 10-K from General Electric.

Scraping SEC filings, 10-Ks and 10-Qs produce enormous datasets, revealing non-trivial patterns in huge aggregations of data. Investors can use this information to identify rare opportunities for alpha as well as high-performing outliers.

The challenge with this type of data is maintaining quality and consistency in its delivery, as the algorithms involved in producing datasets from SEC filings need careful tuning and validation.

Actionable Insights



Identification of trends with regards to debts or material impairments can be identified and understood historically - alongside scraped Schedule 13D and 13Gs, this data can measure risk and assist in predicting company fundamentals.



SEC filings have an immediate causal impact on stock performance. Markets are highly responsive to filings, whose submission has a direct relationship to increased trade volume. trading periods is increased for investors with access to more meaningful insights.

Type 3: Product reviews

To accurately predict how a stock is going to perform for any given company, it helps to understand how key products are tracking in the marketplace.

Unfortunately, the delay between company performance and quarterly earnings can dampen the usefulness of these reports for real-time analysis. Scraping product reviews can allow investors to proactively gather information on a product life cycle and make more up to date assumptions about company earnings.

Use Cases

Analyzing consumer reviews from sites like Amazon can yield valuable predictive information on stock performance.

Jiekun Huang at UI-Urbana Champaign analyzed 6 million Amazon reviews and found a positive correlation between the Amazon reviews and the company stock performance. The study found that when consumers have more favorable views of a company and/or its products, one can assume greater revenue, higher cash flow and therefore a rise in stock prices. It's intuitive enough, should we take the time to examine our own buying behavior.

What's one of the first metrics many of us look to when deciding on a product? The reviews.

Actionable Insights



Utilizing scraped product reviews and average customer rating data from large retailers grants foresight into category performance - in virtually any retail industry.



Evaluating how a company or product is trending against its competition and against expectations helps hone the perception of risk in the investment decision-making process.



Negative consumer reviews can increase stock volatility and pose a serious risk to shareholder value. By scraping product reviews historically and in real-time, smarter and more timely analyses (or corrective actions) can be performed.

Type 4: Company news

Where can we get information straight from the horse's mouth? A company's website and its communications are a good place to start.

Scraping data from the web may include analyzing company web traffic, announcements, and hiring activity in the aim of accessing otherwise exclusive insights.

More than just the company's communications themselves, one can scrape the web for the frequency of company mentions across social media platforms and business content providers, generating extremely useful data on a company's trajectory.

Use Cases

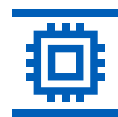
The reality is that companies often report vital information outside of traditional reporting windows, and ignoring this can lead to serious investment blunders. To get the full picture of a company's longevity and performance, keeping track of developments on a more frequent basis is essential. The following are key pieces of data to monitor:

- Press releases
- Company web traffic
- Hiring activity (via company website and employment websites)
- Product announcements

Actionable Insights



Textual analyses across high-dimensional datasets of press releases, company communications and blog posts can yield a more nuanced view of a company's confidence and (true) projected outlook.



Artificial intelligence and natural language processing can aid investors in staying on top of market news. Technology can find, track, and aggregate keywords to determine when a company is rising up the news food chain - this relevant, timely attention also protects investors from scandals.



Corporate news and mentions of a company may precede new product releases and signal investment - predicting the trends towards new releases via historical data means valuable foresight into coming market signals.

Type 5: Sentiment analysis

To shed light on the sheer amount of data embedded in Twitter, reflect on the fact that Twitter users tweet 347,222 times per minute.

Public sentiment on social media gives insight into both the public at large and also any given subset we might want to glean more information about.

As the largest share of the world's consumers becomes those who have grown up immersed in social media, the value in this data set is expected to grow.

Use Cases

Tweets have incredible predictive value. Companies announce developments on social media channels like Twitter, Instagram, and Facebook, and with an opportunistic mindset and consistent dataset, investors can analyze the timing of company announcements to make financial decisions.

Bloomberg recently declared that access to the Twitter stream offers one of the largest and most nutritious alternative data sets for alpha-seeking investors, and new developments in behavioural economic research suggest that “collective mood states derived from large-scale Twitter feeds” could predict Dow Jones movements with an astonishing 87.6% accuracy.

Beyond the qualitative benefits, the speed and volume of social media content generation also gives the medium an edge. Breaking news with huge stock market impact can sometimes hit social media before mainstream news.

Actionable Insights



Company popularity and reputation can be monitored carefully and tracked against competitors.



Investors can track changes in a company's internal framework and gain financial insight on a company outside of quarterly reports.



Influential public figures or celebrities can be monitored for company mentions, along with the tracking of tweets which gain momentum, enabling investors to minimize risk or seize immense opportunities. High-value social media posts should be treated as any other piece of news content, with serious potential consequences for company earnings and stock performance.

Conclusion

Harnessing the power of quality alternative data will be a decisive advantage in the search for alpha.

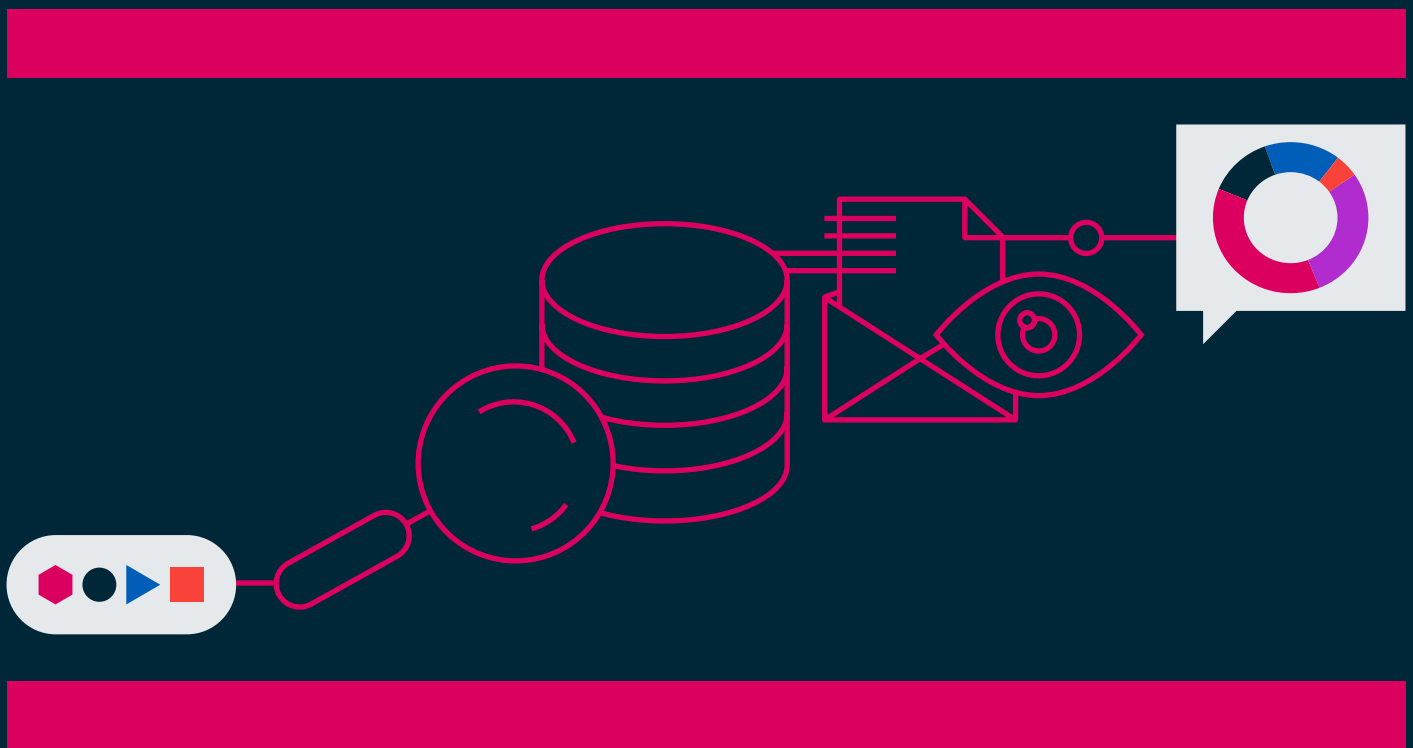
Modern data sources span such a large variety, from those yet discussed to geolocation data, emailed receipts, and even satellite imagery to name a few. The possibilities seem endless, but sophisticated investors will achieve the best results by combining machine learning, human analytics, and large, high-quality sets of alternative data.

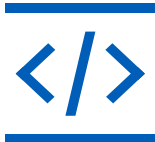
More and better data means your investment decision-making process produces more value, more consistently. Furthermore, adjusting to the new practices of alternative data ensures your models are in sync with the pending data-based transformation of virtually every sector of business.

As a tectonic shift continues to disrupt the investment research landscape, getting on top of these alternative datasets as early as possible is key to capturing the immense alpha left in this data.

But as alternative data slowly enters the mainstream, the insights which stand to be captured today will be gone tomorrow.

If you have a need to start or scale web data acquisition for your alternative data needs then our Solution Architecture team are available for a free consultation, where we will evaluate and architect a data extraction solution to meet your data and compliance requirements.





At Zyte we turn websites into data with industry leading technology and services.

Our solutions include:

- **Data Extraction Service**

Let our web scraping experts build and manage the bespoke data extraction solution for your business needs.

- **Smart Proxy Manager (formerly Crawlera)**

Forget about proxy lists. We manage hundreds of thousands of proxies, so you don't have to.

- **Automatic Extraction powered by AI**

Instantly access accurate web data through our user-friendly interface or various Extraction APIs and save time getting the data you need.

- **Data extraction platform**

Access developer tools, data extraction APIs and documentation, built and maintained by our world-leading team of over 100 extraction experts.

zyte

It's yours. The web data you need.

Access clean, valuable data with web scraping services that drive your business forward.

[Talk to us](#)

www.zyte.com
Copyright 2021 © Zyte

Cuil Greine House, Ballincollig Commercial
Park Link Road, Ballincollig / Co. Cork, Ireland