IS 3005 Statistics In Practice I

Building Alternative data
Created by
Darshana Rupasinghe

Contents

Introduction

ABC is a small team of financial analysts who are conducting an analysis of the iPhones market. To collect data for their analysis, the team has decided to explore the use of alternative data sources. This report will discuss the possible ways of collecting alternative data, advantages of using web-scraping to gather alternative data, limitations and challenges of web-scraping, and the major python libraries used for web scraping.

Possible ways of collecting alternative data

There are several possible ways to collect alternative data.
It depends on the specific requirements and objectives of the analysis.

Social media data

This includes data from social media platforms such as Facebook,
Instagram, Twitter, LinkedIn, and others. Social media data can provide
insights into product reviews, consumer sentiment, and customer feedback,
among other things.

Web scraping

Web scraping involves extracting data from websites using specialized
tools or scripts. This can be useful for collecting data on competitors, market
trends, or product features.

Satellite imagery

Satellite imagery can be used to collect data on things like crop yields,
weather patterns or changes in land use. This data can be useful for predicting
market trends or identifying potential risks.

Sensor data

Sensors can be used to collect data on a wide range of variables, including
air quality, traffic patterns, temperature, humidity and more. This data can be
useful for predicting consumer behavior, identifying market trends, or assessing risk.

Government data

Government organizations frequently collect and publish data on a wide range
of topics, economic indicators, including demographics and industry-specific
information. This data can be useful for understanding market trends,
identifying regulatory risks, or assessing market opportunities.

Surveys

Surveys can be used to collect data on product usage, consumer preferences and other important variables. Surveys can be conducted through phone calls, online platforms or in-person interviews.

These are just a few examples of the many ways in which alternative data can be collected. The specific methods used will depend on the goals of the analysis, the available resources, and other factors.

Advantages of web-scraping

Web scraping can be a powerful way to gather alternative data, but it also has some unique advantages and disadvantages compared to other methods. Here are some advantages of web scraping for gathering alternative data.

Cost-effective

Web scraping can be a cost-effective way to collect large amounts of data from a wide range of sources. Compared to other methods like satellite imagery or surveys, web scraping can be much less expensive.

Timeliness

Web scraping can provide real time data on market trends, consumer behavior, and other variables. This can be valuable for businesses that need to make quick decisions based on current market conditions.

Customizability

Web scraping can be customized to collect specific types of data from specific sources. This can be useful for businesses that need to collect data on a particular market or industry.

Scalability

Web scraping can be scaled up or down depending on the needs of the business. This means that businesses can collect as much or as little data as they need to make informed decisions.

Granularity

Web scraping can provide very granular data on various companies, products or market segments. This can be useful for identifying niche opportunities or assessing the competition in a particular market.

It's worth noting that web scraping also has some potential disadvantages, including legal and ethical concerns as some websites prohibit web scraping the need for specialized technical expertise, and the potential for inaccuracies in the data collected. Ultimately, the choice of method for collecting alternative data will depend on the specific needs and goals of the analysis.

Limitations of web-scraping

Web scraping can be a powerful way to gather alternative data, but there are also several limitations and challenges associated with this method. Here are few potential limitations and challenges to consider.

Legality and ethics

Web scraping can be illegal if it violates website terms of service or copyright laws. It's important to ensure that any data collected through web scraping is obtained both legally and ethically.

Technical complexity

Web scraping requires specialized technical expertise, including knowledge of programming languages, web architecture, and data parsing. It can be challenging to set up and maintain a web scraping system, particularly for businesses without dedicated technical staff.

Data accuracy and reliability

Web scraping can be subject to errors and inaccuracies, particularly if the website being scraped changes its formatting or content. It's important to carefully validate and clean any data collected through web scraping.

Limited scope

Web scraping is limited to data that is publicly available on websites. It may not be possible to collect data on certain variables or from certain sources, depending on the availability of that data online.

Website restrictions

Many websites prohibit web scraping, and some use technical measures to prevent it. This can make it difficult or impossible to collect data from certain sources.

Data overload

Web scraping can generate vast amounts of data, which can be difficult to manage and analyze. It's important to have a clear plan for how to process and analyze the data collected through web scraping.

Data privacy

Web scraping may collect personal information about individuals, which can raise privacy concerns. It's important to ensure that any personal information collected through web scraping is handled in compliance with relevant privacy laws and regulations.

Overall, web scraping can be a powerful way to gather alternative data, but it's important to be aware of these limitations and challenges and to carefully consider whether this method is appropriate for our specific needs and goals.

Major python libraries used in web-scraping and their uses

There are many Python libraries that can be used for web scraping, but here are three popular ones and their uses.

BeautifulSoup

BeautifulSoup is a Python library used for parsing HTML and XML documents. It is used to extract data from HTML and XML files, including web pages. It allows the developer to parse and navigate the HTML tree structure in a simple and intuitive way. This library can be used to find specific tags, attributes, and text within HTML and XML documents.

Scrapy

Scrapy is a powerful and flexible web scraping framework for Python that is used to extract structured data from websites. It provides a complete framework for building web spiders that can crawl websites, extract data, and store it in various formats. It can handle complicated web scraping tasks and design for large-scale data extraction projects.

Selenium

Selenium is a web testing library that can also be used for web scraping. It allows the developer to interact with web pages programmatically and simulate user interactions. This can be useful for scraping dynamic web pages or websites that require user authentication. Selenium can also be used to automate web-based tests and to control web browsers programmatically.

These libraries can be used together or separately depending on the needs of the web scraping project. For example, BeautifulSoup can be used to parse HTML and XML documents, while Scrapy can be used to crawl websites and extract structured data. Selenium can be used to interact with web pages programmatically and simulate user interactions.

Conclusion

In conclusion, alternative data sources such as web-scraping can provide valuable insights for financial analysts. While web-scraping has several advantages, it is important to consider the limitations and challenges of this method, especially legal considerations, technical challenges and data quality.

By using the right tools and techniques, however, analysts can overcome these challenges and leverage alternative data sources to gain a competitive edge in their analysis.