DATA SCIENCE - EXTRACTING INSIGHTS
Open Journal

Review

Tracking Reviews and Scraping the Web. Gaining Understanding from Business Records

Julio C. Fernández-Travieso; José Illnait-Ferrer; Lilia Fernández-Dorta; Sarahi Mendoza-Castaño*

Laboratory for Atmospheric and Space Physics, Boulder, CO, USA

Corresponding author
Sarahi Mendoza-Castaño
Laboratory for Atmospheric and Space Physics, Boulder, CO, USA

Article information
Received: April 7th, 2024; **Revised:** October 19th, 2024; **Accepted:** November 9th, 2024; **Published:** November 26th, 2024

Cite this article
Fernández-Travieso JC, Illnait-Ferrer J, Fernández-Dorta L, Mendoza-Castaño S. Tracking reviews and scraping the Web. gaining understanding from business records. 2024; 2(2). doi: https://doi.org/10.70705/ppp.dsei.2024.v02.i02.pp153-156

ABSTRACT

There are a lot of uses for web scraping. To get more out of web pages, you may utilize it in conjunction with APIs. As an example, there is a lot of commercial data out there, but it may not be relevant in its current form on websites. Our proposal in this paper is to analyze the reviews and derive valuable insights from the data using web scraping techniques (specifically, BeautifulSoup and Selenium, two popular libraries) and other Python libraries and techniques (vaderSentiment, SentimentIntensityAnalyzer, nltk, n consecutive words). We constructed a web scraper to extract prices and track their fluctuations. In addition, the evaluations are culled and examined to find pertinent comments, including consumer concerns.

Keywords
Web scraping; Price tracking; Sentiment analysis; Alerts.

INTRODUCTION

Human resources (HR) firms, banks looking to get an edge in the market, marketing, public opinion on Bitcoin, psychology, trends in depression and suicide ideation, price comparison, and many more uses for web scraping have been found (Boegershausen et al., 2022; Landers et al., 2016; https://www.sas.com/en_ca/insights/articles/analytics/using-big-data-to-predictsuicide-risk-canada.html). A possible inclusion of online scraping data in the big data paradigm. According to a case study in (Landers et al., 2016), huge data in the field of psychology may be extracted utilizing Python and web scraping. Additionally, they discuss the ethical and legal considerations that come up in online scraping efforts. The veracity of the data, as well as legal and ethical considerations, were the primary foci of a few scholars that examined the topic of online data gathering. More than 300 articles concerning online marketing were analyzed. Questions like "What information to extract?" were among the many intriguing ones raised by this study. What is the best method for sampling? How often should data be extracted? "How can we best analyze this data?" The stated source is Boegershausen et al. (2022).

Khder (2021) notes that a plethora of programming languages, methodologies, and tools are available to make online scraping easier (Saurkar and Gode, 2018). Automated data extraction, parsing, and organization from the web is known as web scraping. In order to access the data in an organized way, many websites nowadays provide an API. When looking at the reliability of Twitter content, both application programming interfaces and web scraping are utilized. They provide several perspectives for data extraction. One finding is that site scraping alone is unable to collect all Twitter properties, according to the research. They both provide comparable credibility levels, but both involve a lot of work before the tweets are processed (such normalization). Scraping the web was more efficient and versatile than using the Twitter API. However, as a result of website updates, web scraping was more prone to failure (Dongo et al., 2021).

Still, the intended functionality might not be exposed by APIs (Glez-Peña et al., 2013). A rate limit might be in place for the API, limiting the amount of times data can be retrieved per second or day. If the website exposes all the essential data but the API does not, then scaping will be needed. Two writers asked, "Why is it more efficient to combine BeautifulSoup and Selenium in scraping for data under energy crisis?" (https://stec.univ- ovidius.ro/html/anale/RO/2022-issue2/Section%201%20and%202/19.pdf). When data extraction and dynamic actions are needed to acquire data, the combination of BeuatifulSoup and Selenium is preferable, according to their comparison of the two Python libraries' scraping performance.

2. Theoretical background

In their 2018 publication, vanden Broucke and Baesens detailed many applications. Cases when data that isn't readily available on re-

search sites (like Kaggle or ResearchGate) are given by the authors. For proof-of-concept purposes, it is common practice to scrape once and then extract data. In order to compile up-to-date data for new projects, it is common practice to use several data sources that allow for the merging of different time series. As a further example, data from inverters may be scraped for daily forecasts, for example, during application execution. The latter necessitates a more comprehensive strategy, the accessibility of the website, and the consideration and treatment of additional circumstances.

The US Copyright or Trademark Infringement defines what is supposedly a reasonable use of data collected by web scraping. Including the data in academic study does not need specific authorization. Nevertheless, commercial use of copyrighted content often need prior authorization. In addition, the CFAA states that someone who "intentionally accesses a computer without authorization... and as a result of such conduct recklessly causes damage" is regarded to be abusing their position, particularly if the owner of the webpage can provide evidence of any harm or loss. Not to mention the Computer Misuse Act and the Trespass to Chattels statutes in Europe. However, without proper authorization, online scraping—also called crawling—on a big scale for commercial reasons might run into legal trouble (Krotov, Johnson and Silva, 2020).

The BS and Selenium libraries aren't the only ones out there. They include R (the rvest package), PHP (the curl package), and so on. Also, there is an alternative that may be found in Python libraries (such as Scrapy). One major issue with Scrapy is that it doesn't simulate a full browser. Because of this, using this library to handle JavaScript could be challenging. Python provides the CatchControl utility to prevent servers from being overwhelmed with requests. This is particularly useful when creating scraping programs, as scripts are restarted often to verify if bugs have been resolved, expected results have been received, and so on. Graphical scraping programs like Portia, Parsehub, Kapow, Fminer, and Dexi are also available. One major drawback is that scraping becomes problematic when JavaScript is heavily loaded. When the page is constructed in a less-than-simple way, these pre-made tools occasionally fail. Since some website designers make a concerted effort to evade or prohibit scraping, online scraping might feel like a game of cat and mouse. For example, according to Olufemi et al. (2021), AI and deep learning versions of the Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) pose a threat to the designers of these tests.

Determining the positive, neutral, or negative tone of a text is known as sentiment analysis or opinion mining. A customer's attitude may be clearly shown by extracting an opinion from letters, even if such words are somewhat extensive. Marketers, politicians, and public servants frequently employ this method. Finding out clients' ideas and sentiments is important for business enterprises since it helps them design plans. Their goal is to learn how consumers react to advertisements and new products, as well as the reasons behind the lack of interest in certain items. Because it shows how consistent or inconsistent political views are, text mining is also useful in politics. Estimates of the result of elections can also be developed from analysis of public opinion.

Political statements are investigated to measure their impact on the economy and on particular assets, such as Bitcoin, electric cars, RES technologies, etc. Public actions can be monitored to understand social movements and identify moods and events that can emerge from public initiatives.

In this paper, the goal is to investigate how do the prices change for a popular category of tech products in a short period of time, as well as retrieving and studying the attitude of the customers towards that specific product. Only the products with a significant number of reviews will be used for analysis, reducing the large sample of products to a much smaller one, with more significant elements. A similar thinking will be applied for the price tracking over the selected period.

3. Research methodology

In this paper, a combination of Python libraries is proposed to extract and investigate data from the web. Python has a variety of libraries that interact with HTTP. For instance, httplib2 is a small, fast HTTP client library. Originally it was developed by Googler Joe Gregorio, and now it is supported by the community. Another library is urllib3 that is a powerful HTTP client for Python. Requests is also a popular library. Grequests extends requests to deal with asynchronous HTTP. Aiohttp is another library focusing on asynchronous HTTP. Requests and get method with a webpage string as parameter (url) create a page object. For static page elements, in order to obtain page's information, a BeautifulSoup (BS) object is created. This library is especially useful for finding html elements in the webpage source. For this purpose, the BS library relies on an HTML parser. In Python, multiple parsers do exist, such as: html.parser that is a built-in Python parser that is useful especially when using recent versions of Python 3 and requires no extra installations. Lxml is very fast parser, but it requires extra installations. Html5lib parses the webpage as a web browser, but it is slower. BS's main purpose is to transform the HTML into a tree-based representation. Extracting the content is facile using a BS object, and the two methods (find and findAll) fetching the data from the webpage. However, requests and BS are not enough to deal with script tags. Selenium is a powerful web scraping tool that was originally developed for the purpose of automated website testing. Selenium operates by automating various browsers to load a webpage, retrieve its contents, and control it like a user would when using the browser, clicking on buttons, on links, etc. Thus, it is a powerful tool for web scraping focusing on the dynamic actions that a user might do. Selenium can be used by several programming languages (Java, C#, PHP, and of course, Python). However, Selenium does not come with its own web browser and requires a WebDriver to interact. WebDrivers are available for browsers, including Internet Explorer, Chrome, Firefox, Edge, Safari, etc. With these WebDrivers, a browser window will open and simulate actions included in the Python code. The WebDriver has to be downloaded (https://sites.google.com/a/chromium.org/chromedriver/downloads) and its path inserted into Advanced System Setting Environment variables PATH or locate the WebDriver in the same directory as the Python scripts. Selenium finds elements and perform actions using several methods (like send_keys). Both libraries run on online platforms such as Google Collaboratory. Sentiment analysis using vaderSentiment, SentimentIntensityAnalyzer and nltk libraries are applied to identify whether the reviews are positive, neutral or negative. Moreover, the content of reviews is

investigated measuring the combination of n consecutive words in order to identify the most frequent complaints.

4. Findings

The purpose of this section is to provide the price tracking results and review analysis. The data analysed was extracted from an e-commerce website of a Romanian retailer which focuses on technology related products and the category chosen for the analysis is phones. Using BeautifulSoup, the following data was extracted and processed to match the purpose of the analysis: the product's code, the URL to the product's information page, details about the price (the initial price, discount and the final price), as well as data related to reviews which will be used later, such as the total rating from the reviews and the number of them.

The movement of the prices has been followed during two separate weeks: 15-April to 21-April and 01-May to 07-May. The number of products varied from one day to another, but after merging the tables, 386 products remained. Since there are over 385 products daily in the chosen category, the purpose of the analysis was to seek the products for which prices fluctuated the most in each of those two weeks. A sample was taken from the total number of products for each week, consisting in 11 products for April and 7 products for May, so as to clearly see the evolution for the selected products in the diagrams from Figure 1 and Figure 2.

- SMTIP14PM1BK → APPLE iPhone 14 Pro Max 5G, 128GB, Space Black
- SMTIP14PM1GD → APPLE iPhone 14 Pro Max 5G, 128GB, Gold
- SMTS23U1TBBK → SAMSUNG Galaxy S23 Ultra 5G, 1TB, 12GB RAM, Dual SIM, Phantom Black
- SMTIP141YL → APPLE iPhone 14 5G, 128GB, Yellow
- SMTIP145BL → APPLE iPhone 14 5G, 512GB, Blue
- SMTIP14P2BL → APPLE iPhone 14 Plus 5G, 256GB, Blue
- SMTIP14P2PP → APPLE iPhone 14 Plus 5G, 256GB, Purple

As we can see, the products which suffered multiple price changes within a single week were the latest phones from the brands Apple and Samsung. SMTS23U1TBBK, the Galaxy S23 Ultra 5G, had the most sudden change, a raise of 17%, followed by a fall in the next two days. For this week the prices had a predominantly upward trajectory, with visible changes from one day to another. For the week 15 April to 21 April, there were not many sudden price fluctuations, but more subtle differences suggesting that mostly the prices went down. We see that the most visible change was for the product with the code SMTZFOLD42BE, which had a dramatic decrease on 20 April, then went back up a day later. The name of the product impacted by this change was "SAMSUNG Galaxy Z Fold4 5G, 256GB, 12GB RAM, Dual SIM, Beige", another one of most recent Samsung releases.

It can be concluded that the products which are most prone to price changes from one day to another during the mentioned two-week period are, more often than not, the latest released phones.

A sample of 392 tech products was extracted from the website, only from the phones category. From each device, all the reviews were extracted, alongside the review score and the number of reviews, which will be later used to compare the response received from the sentiment analysis to the overall opinion of the product. The products were filtered by the number of reviews, excluding those with no reviews (6,12%) and those which received less than 100 reviews (75,77%). All the reviews were translated to English using googletrans Translator, a free Python library which uses the Google Translate API to be able to translate at once pieces of text with a maximum of 15000 characters (https://py-googletrans.readthedocs.io/en/latest/). The reviews were selected and analysed with the purpose of establishing the customer's feeling towards the acquired product.

Sentiment analysis was used to determine the opinion of the buyer by placing it in the pertaining category: positive, negative or neutral. Each review returned a scoring system which consisted of three values: negative, neutral and positive, all of which should add up to 1 to create a whole. The last component of the scoring system is called 'compound', which was used as the overall score for each review (https://vadersentiment.readthedocs.io/en/latest/pages/about_the_scoring.html). The review was considered positive if it had a compound of at least 0.5 and negative if it was less than or equal to -0.5. Anything in between was considered neutral. Only the devices with more than 100 written reviews were selected, the percentage of the positive reviews was calculated for each one of them and the results are presented in Figure 3.

In Figure 3, it is clearly seen that the percentage of positive reviews is well over 65%. The lowest one (69.15%) belongs to code SMTIP132BL, corresponding to the product "APPLE iPhone 13 5G, 256GB, Blue" and the biggest score (89.57%) belongs to product SMTIP121BK which is the product "APPLE iPhone 12 5G, 128GB, Black ". Product SMTIP121BK had an overall score of 4.87 out of 115 reviews, while product SMTIP132BL had an overall score of 4.90 out of 295 reviews. By looking at the difference between the grade-only reviews and the written ones, for SMTIP121BK there are 113 written reviews and 2 grade-only, whereas for SMTIP132BL there are 110 written reviews and 185 grade-only. From this, it can be deducted that the higher grade for the device with the lowest percentage of positive scores comes from the grade-only reviews which could not be included in the sentiment analysis. The top 10 devices with the most written reviews are represented in Figure 4. Upon searching the product codes, it was observed that all the products correspond to the latest iPhones models from Apple: iPhone 13, iPhone 14 Pro and iPhone 14 Pro Max, with varying specifications.

Product code SMTIP131MN, which is one of the devices with the most written reviews, corresponds to "APPLE iPhone 13 5G, 128GB, Midnight" with a review score of 4.91 from a total of 290 reviews. Out of these 290 reviews, 284 consisted of written reviews with 252 having a positive sentiment (86.9%) and 33 having a negative sentiment (11.38%). On the other side of this top, product SMTIP14PR1BK corresponds to "APPLE iPhone 14 Pro 5G, 128GB, Space Black" and has 160 written reviews out of 160 reviews in total. The written reviews were analyzed, which led to the conclusion that 127 reviews were found with a positive sentiment (79.38%) and 25 (15.62%) with a negative sentiment, having a 4.89 review score. It can be concluded that the latest iPhones were the most request-

ed, leading, thus, to a great number of written reviews, whereas for many other devices the review was only resumed to the grade itself.

Figure 5 aims to highlight the devices with a perfect review score, 5 out of 5, taking into account only the phones which have at least 15 reviews, sorted in a descending order by the number of reviews. As we can see, the first 6 devices have the same number of reviews, 41, while the remaining 4 have a significantly lower number, 21 and 20. All the devices present in the above chart are different versions of the same product, iPhone 14 Plus 5G, which have a few distinctions such as color and storage capacity. Just as in Figure 4, it seems that the latest versions of iPhones have the highest ratings, being among one of the most popular devices.

A n-gram analysis is performed for three devices in order to identify common segments of words based on their frequency. The first device is Phone: SAMSUNG Galaxy A04s, 32GB, 3GB RAM, Dual SIM, Black; Review score 4.73; Number of reviews: 116; Code: SMTA04BK. The total number of 4-grams is 1243. Analyzing the frequencies that are higher than 2, we obtained the following text elements that all reveal positive opinions of customers. A selection of opinions is displayed in Figure 6.

5. Conclusions

In this article, we analyzed data from a prominent category of technology devices and monitored how their pricing changed over a short period of time. Using sentiment analysis tools, we looked at how customers felt about that product in particular. In order to narrow the sample down to a more manageable size, we only included goods with a big number of reviews in our research. This allowed us to focus on the most important factors.

Over the course of two weeks in April and May, we looked at price data and discovered that newly introduced phones from Samsung and Apple were the most volatile, with numerous price adjustments in a single week.

Review positivity, neutrality, and negativity were determined by applying sentiment analysis with the help of the vaderSentiment, SentimentIntensityAnalyzer, and nltk libraries. In order to find the most common complaints, we also looked at the review content by evaluating the combination of n consecutive terms.

REFERENCES

A. Borah, J. Boegershausen, H. Datta, and A.T. Stephen published this in 2022. Harvesting Marketing Intelligence from Web Data: A Field Guide. Journal of Marketing, volume 86, issue 5, pages 1–20, doi:10.1177/00222429221100750.

It was published in 2018 by vanden Broucke and Baesens. Web Scraping made easy for data scientists. How to Scrape the Web for Data Science: A Practical Guide, https://doi.org/10.1007/978-1-4842-3582-9.

This information is sourced from a publication by Dongo et al. (2021) under the authors' names. An examination of Twitter credibility using web scraping and API technologies, comparing their qualitative and quantitative aspects. Published in volume 17, issue 6, pages 580-606, this article may be accessed online at this URL: https://doi.org/10.1108/IJWIS-03-2021-0037.

This information is from a 2013 publication by Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato, and Fdez-Riverola. Web scraping tools in the API era. Briefings in Bioinformatics, volume 15, issue 5, pages 788–797, doi:10.1093/bib/bbt026.

Khander (M.A.), year 2021. The latest developments, methods, strategies, and applications in web scraping, also known as web crawling. The article may be found in volume 13, issue 1, pages 145–168, and the DOI code is 10.15849/IJASCA.211128.11.

(2020) Krotov, V., Johnson, L., & Silva, L. Explainer: The morality and lawfulness of web scraping. Publication date: https://doi.org/10.17705/1CAIS.04724, volume 47, issue 1, pages 555–581. Published by the Association for Information Systems.

Brusso, R.C., Cavanaugh, K.J., and Collmus, A.B. (2016) were the authors of the study. An introduction to theory-driven web scraping: the procedure of automatically extracting large amounts of data from the internet with the purpose of doing psychological study. "Psychological Methods" (Vol. 21, Issue 4, Pages 475–492), with the DOI 10.1037/met0000081.

This information is sourced from a publication by Olufemi et al. (2021). Research trends on CAPTCHA: A comprehensive literature. Article citation: International Journal of Electrical and Computer Engineering, Volume 11, Issue 5, Pages 4300–4312, Online DOI: 10.1590/ijece.v11i5.pp4300-4312.

Saurkar and Gode (2018) found this information. Getting Started with Web Scraping Methods and Software. The article may be found at this URL: https://api.semanticscholar.org/CorpusID:198993824, published in the International Journal on Future Revolution in Computer Science & Communication Engineering.