Leo Richard Hermann Giesen

Extraction of Consumer Insights with Scalable Web Scraping of Forums

Seminar Thesis
in the context of the seminar "Principles of Entrepreneurship"

at the Chair for Information Systems and Information Management
(Westfälische Wilhelms-Universität, Münster)

Supervisor:	Prof. Dr. David Bendig
Tutor:	Lucas Mantke, M.A.

Presented by: Leo Richard Hermann Giesen
Bismarckallee 51
48151 Münster
+49 176 83386614
l_gies10@uni-muenster.de

Date of Submission: 2020-07-11

Content

Figures

Tables

Abbreviations

API	Application Programming Interface
BFS	Breadth-First Search
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CSS	Cascading Style Sheet
DFS	Depth-First Search
HTML	Hypertext Markup Language
IP	Internet Protocol
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
NLP	Natural Language Processing

1 Contextualization of Project Work

The Information Systems specialization module 'Principles of Entrepreneurship' held by Prof. Dr. David Bendig, Lucas Mantke and Kathrin Teupe, was supplemented by the project group work with the startup pivoty. This project strengthens pivoty's product development as well as gives the project team the opportunity to gain practical experience. Thus, both parties benefitted from the external input and work together.

pivoty is located in Münster and develops "an AI-based analytics software that tries to discover innovation potentials" (Schäper et al. 2021a) for companies "by analyzing unbiased customer insights" (Schäper et al. 2021a). To clarify, the extracted ideas are subjective, but the positivity bias may be overcome (see 30). These insights "draw on a variety of online sources like social networks, forums, blogs, and product reviews to systematically derive inspirations for new products and services" (Schäper et al. 2021b). The corresponding two-step process portrays that the innovation potential is derived with an AI analysis from extracted information from discussion forums using web scraping (see Derivation of Innovation Potential).

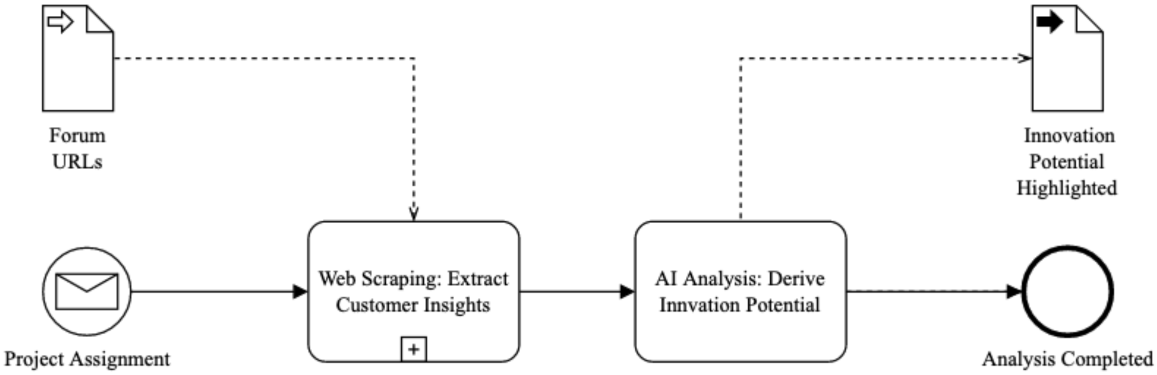


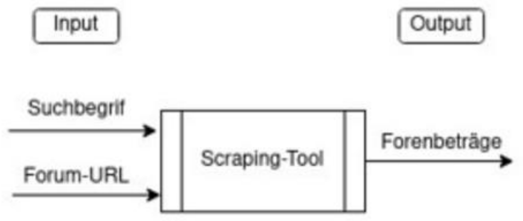
Figure 1 Derivation of Innovation Potential

As outlined in Derivation of Innovation Potential, a company first intends to innovate a product and contacts pivoty to exploit consumer insights to fulfill their innovation potential. Second, relevant information from discussion forums is extracted by a web scraper. Third, the large amount of extracted data is evaluated and analyzed by "AI-powered text-analysis algorithms […] to derive innovation potentials" (Schäper et al. 2021a). Fourth, pivoty highlights those potentials and therefore invokes innovation.

The special prominent value to the client lies in the process' efficiency, as unlike "time-consuming customer interviews and costly market studies, pivoty quickly collects unbiased customer experiences from various internet sources" (Schäper et al. 2021a). Additionally, pivoty may stand out from competitors by positioning itself at the end of the product development process and focusing on a specific product (see 28). This enables them to a high level of customer centricity as a central competitive advantage (Schäper et al. 2021a).

The AI-based algorithm is currently performed by the software WordStat (see 29). Though, the long-term goal is to build an AI-based algorithm for pivoty itself (see 30). Due to its complexity and the duration of the project, working on the AI in this project was discarded. Instead, the team focused on data collection, which is required by the AI.

The task was defined by pivoty to implement a scalable process, which enables the scraping of various primarily German-speaking discussion forums in a consistent format. Furthermore, discussion forum posts are primarily derived from search, because it enables company or product-related research of consumer insights in contrast to scraping randomly. In detail, the web scraper receives a search term and various discussion forum URLs and outputs all the information of the resulting discussion forum posts (see Figure 2).



Source: pivoty (2021)

Figure 2 Web Scraping Task

2 Data Extraction Approaches

2.1 Evaluation of Data Extraction Approaches

The goal of the web scraping process is to extract the evaluation and ideas regarding a product by consumers. To determine the best approach, four possible options considered by pivoty are weighed up against each other:

The first option is to buy data, e. g. Twitter offers different levels of data access. However, this only works for pivoty in the short-term with no better alternatives, because it provides insufficient data quality according to pivoty's experience (see 30). Moreover, the required premium plan comes with a monthly payment (Twitter 2021a, 2021b, 2021c). Consequently, buying data is not the desired approach.

The second option is to directly contact the discussion forum database administrator to export the desired data. This has the advantage of high data quality, though this approach is inefficient, because each discussion forum must be targeted separately and there is a long communication delay, which disqualifies this approach.

The third option is to check if a discussion forum provides an API. This would provide high data quality, though this case rarely occurs. Furthermore, this approach would only be scalable in the unlikely situation that the sources provide uniformly structured APIs.

Thus, the only other considered option is web scraping, which requires post-processing to provide high data quality and maintenance, if a website changes its structure. Depending on the given reliability and scalability, web scraping might not be an approach worth pursuing. For instance, if each discussion forum requires a separate implementation, it is not economically feasible. Therefore, scalability is of high importance. Further, it is worth keeping in mind that web scraping may come with limitations and ethical questions as explained in 5.3.

2.2 Feasibility of Selected Approach

Because the web scraping process is directly linked to the discussion forums and the AI-based algorithm as an input and output, its feasibility is dependent on them. Thence, the possibility of realization of the web scraper in the context of pivoty's business model is determined by

1. the value the discussion forums offer through consumer insights,
2. the difficulty of the implementation of the scalable web scraper and

3. the ability of the AI-based algorithm to derive innovation potential.

First, the AI-based algorithm needs to be supplied with high quality consumer insights by the web scraper to derive consumer needs. This requires the existence of consumer insights in discussion forums in the first place. Although one could claim that only few posts offer insights. It is important to note that once the web scraper is completed, it is usually able to retrieve thousands of posts, which almost inevitably contains insights. The literature confirms that the user-generated content in online discussions forums offer significant knowledge (Bradley and James 2021, p. 59; Pretzsch et al. 2012, p. 821; Zhao et al. 2019, p. 151). Additionally, the AI is powerful enough to identify these consumer ideas and understand if a topic is highly controversial using topic modeling (see 28).

Second, the difficulty of the web scraper's implementation can be estimated by comparable approaches. It can be concluded that "[w]eb scraping is a feasible strategy with which to explore […] forums" (Mintz et al. 2020, p. 1915) and extract valuable consumer insights (Mintz et al. 2020, p. 1915; Pretzsch et al. 2012, p. 821; Zhou and Palma 2017, p. 1). This statement is confirmed by pivoty and the student team's personal experience before and during the project (see 31, 32).

Third, to maximize the overall effectiveness of the selected approach, it is helpful to understand how the AI-based algorithm works and whether or not it is feasible. It combines

1. Natural Language Processing (NLP),
2. text mining and
3. topic modeling

to understand the consumer insights and derive innovation potential. Since these three AI fields or concepts differ, they are dissociated and defined in the next section to highlight in how far they contribute to pivoty's AI-based analysis (see 29).

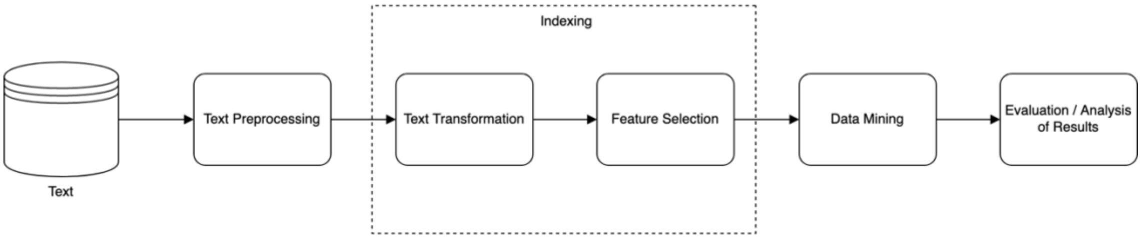
3 Techniques of AI-based Algorithm

The first AI-field applied in the algorithm is NLP, which is difficult to define as “there is not a single agreed-upon definition” (Liddy 2001, p. 1) because it is a broad field in AI (see 29). Nevertheless, one could define NLP as “a collective term referring to automatic computational processing of human languages. This includes both algorithms that take human-produced text as input, and algorithms that produce natural looking text as outputs” (Goldberg 2017, p. xvii). In comparison, Liddy (2001, p. 1) defines NLP as “a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications”. Further definitions similarly state that NLP is able to comprehend, analyze, process and manipulate verbally and textually based human language (Brownlee 2019; Lutkevich 2021; SAS Institute Inc. 2021).

The discussion forum posts of pivoty’s AI-based algorithm are generated by consumers and the input and innovation potential is expressed in natural language. Therefore, NLP is a fitting and helpful technology to understand and process the consumer insights from discussion forums.

The second AI-field applied in the algorithm is text mining, which “is the art of data mining from text data collections. The goal is to discover knowledge (or information, patterns) from text data, which are unstructured or semi-structured. It is a subfield of Data Mining (DM), which is also known as Knowledge Discovery in Databases (KDD)” (Cai and Sun 2009, p. 3061). Additionally, it creates value from large amounts of unstructured text data (Ananiadou and Mcnaught 2006, p. 1; IBM 2021; Kremer 2019).

The relevance of text mining for the AI-based process is clarified by the fact that based on the unstructured discussion forum posts, knowledge about potential innovation is acquired. Hence, KDD is applied, which creates valuable information from discussion forum posts, which represent a large text data collection. Additionally, the AI-based process may follow the text mining process as it utilizes discussion forum posts as text input and applies text preprocessing to clean up the text for impurities and inconsistencies. An example for impurities is the usage of the Unicode encoding, which could be transformed into the standard character encoding UTF8 because the former displays the letter ‘ü’ as ‘\u00fc’, which may hinder feature selection (see Figure 3; App. 6D.d; App. 6D.e). Next, indexing is performed consisting of text transformation and feature selection. The former creates a document matrix to structure the discussion forum posts according to their focus and the latter clusters discussion forum posts to create document groups and extract basic concepts. After that, data mining identifies patterns within categories, which enable to derive e. g. wishes or ideas from consumers in the evaluation and analysis of the results (see Figure 3 Text Mining Process).

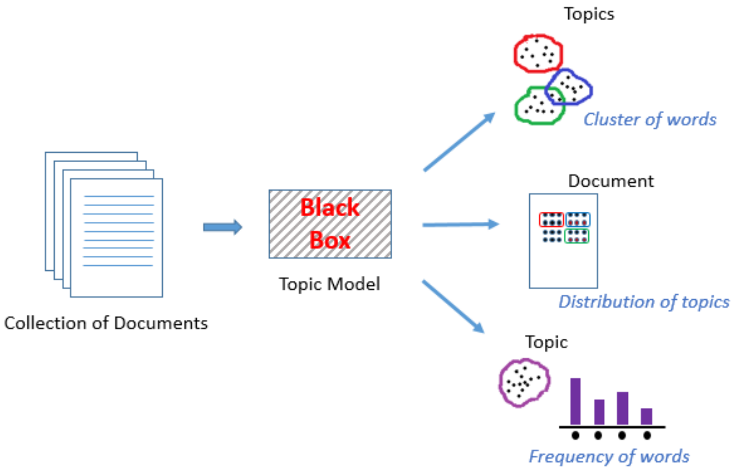


Source: Openminted Communications (2018); Paudyal (2014, p. 9); Rashid et al. (2016, p. 5145)

Figure 3 Text Mining Process

The third AI-field applied in the algorithm is topic modeling, which, enables the understanding of a large number of documents with a statistic process. This is achieved by determining the relevance of various words on the basis of their frequency of occurrence. Additionally, the words are clustered to derive central topics and the distribution of topics within the documents is analyzed to further enable insights (see Figure 4 Topic Modeling Process Akwei 2019; Khare 2017; Li 2018; Naskar 2021). Moreover, it is an "emerging text mining technique[,] which helps in mitigating limitations of traditional text mining concepts" (Khare 2017). Thus, topic modeling classifies a number of documents, for instance blog posts (Silge and Robinson 2021).

The classification of core topics and words using clusters of words and their frequency is essential in estimating the relevance of specific innovation potentials. Further, the distribution of topics across posts and discussion forums enable to identify controversial topics. This analysis may detect that many consumers focus on one specific topic, which may be addressed and acted upon by the company early on. For instance, if one product is highly appreciated by the consumers online, it may help to forecast future demand. In other cases, the brand image and awareness might be derived from conversations and debates online, so that the respective company may react in accordance with that to improve their image.



Source: Akwei (2019)

Figure 4 Topic Modeling Process

Concluding, NLP, text mining and topic modeling are relevant to the AI-based algorithm as they play different roles in the derivation of innovation potential. Though, in practice those techniques overlap and work together seamlessly (see 29).

Since these AI-based techniques need to be able to derive innovation potential, their feasibility was tested and confirmed in previous tests from pivoty (see 28). Furthermore, there are numerous successful approaches, which resemble pivoty's (see Table 1). Therefore, the data quality from web scraping and the ability to derive innovation potential has been demonstrated. The prerequisites for the comparable approaches to qualify as such consist of

1. the application of the AI to analyze text-based language and
2. the identical type of input for the AI.

I. e. the text input needs to be scraped from online discussion forums to realistically represent the Artificial Intelligence's capabilities to derive insights from them. The AI-based process also utilizes at least one of the three presented text analysis techniques. Consequently, the web scraping and AI-based processes are similar to the ones applied in pivoty's context.

NLP	• Khymytsia et al. (2019) • Naveen et al. (2020) • Nicolas et al. (2019)
Text mining	• Abdellaoui et al. (2018)
Topic modeling	• Bradley and James (2021) • Koloveas et al. (2019) • Pee et al. (2020) • Porter (2018) • Törnberg and Törnberg (2016) • Zhao et al. (2019)

Table 1 Literature Approaches Using Web Scraping and AI in Combination

4 Functionality of Web Scraping

Apart from the AI in the process of deriving innovation potential, web scraping is applied to extract information. In order to understand the implementation of the web scraper, it is crucial to comprehend its functionality first. Generally, web scrapers and crawlers are two terms, which are widely (mis-)used. Therefore, it is important to dissociate them.

A web crawler or spider automatically fetches links from web pages. Subsequently, the web crawler is applied to the newly fetched web pages resulting in an iterative process of crawling hyperlinks. In contrast to that a web scraper processes a web page and extracts specific data out of it (Google 2021; J. and Ben 2012; Parsers VC 2019). Hence it is more adjusted to the analyzed website. In contrast to web crawlers, the web scraper is usually not affected by the robots.txt, which "applies any automated process that accesses a web page" (J. and Ben 2012). E. g. it specifies what pages should not be crawled (see 6B). Since a web scraper does not put any significant burden on the website traffic, it may not consider the robots.txt. This can be achieved by specifying `ROBOTSTXT_OBEY = False` in the settings.py file. This is controversial because it might be morally reprehensible depending on the context since this file explicitly directs crawling bots what they should not do.

The focus of this project lies on web scraping rather than crawling because web scraping is often applied in data analysis, data mining and information processing to provide or extract data when no API is available (Traversy Media 2020). It can be performed with the library BeautifulSoup, the frameworks Selenium or Scrapy. In this project, the last one is used because the scalability requires high speed performance and is relatively complex in comparison to other web scraping projects (Glez-Peña et al. 2013, p. 791; Palakollu 2019).

The functionality of a web scraper is comprised in the processing and retrieval of the website data as a response from a server request. The targeted URLs are defined in the `start_urls`, which allow to scrape multiple discussion forums at once. The server response is handled in the parse function, where the processed response may be stored in a JSON file using `yield { 'attribute': ... }`. The function applies a combination or chain of CSS selectors or XPath to navigate to specific tags of the HTML response. In the example in Table 2 Exemplary Application of CSS Selectors and XPath, the first division `div` with the class `className` and ID `id` is selected. Within that division the selection of the first link's text `a::text` is chained. In comparison to that the XPath only uses the HTML structure for navigation (see Table 2

Exemplary Application of CSS Selectors and XPath; Traversy Media 2020).

CSS selectors	`response.css(‘div.className#id a::text’).get()`
XPath	`response.xpath(‘/html/body/div[1]/div/div/div[1]/div/div[2]/div[4]/div/a[1]::text’).extract()`

Table 2 Exemplary Application of CSS Selectors and XPath

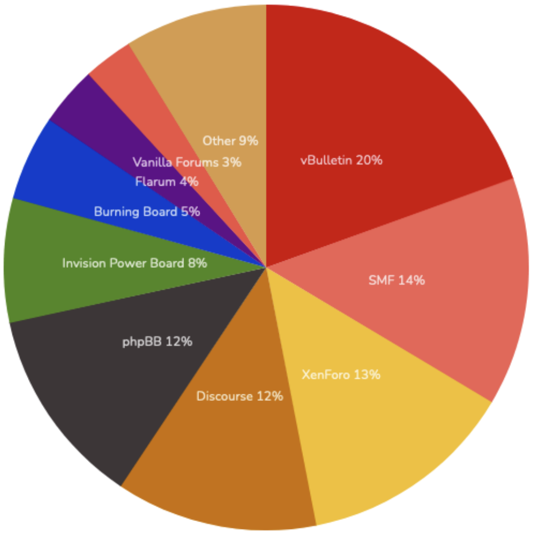
5 Implementation of Web Scraping

5.1 Strategy for Scalability

The functionality of web scraping lies a foundation for the overall scalability of scraping discussion forums, which was tested at the hand of a potential project with BabyOne, where the company's online awareness was investigated in various discussion forums. The insights from that investigation should aid BabyOne to form a decision about expanding their product features and their product range. The scalability of the web scraper allows it to be applied to almost any discussion forum. So the web scraper may also aid companies from other industries, is independent and is not constrained by a particular application background.

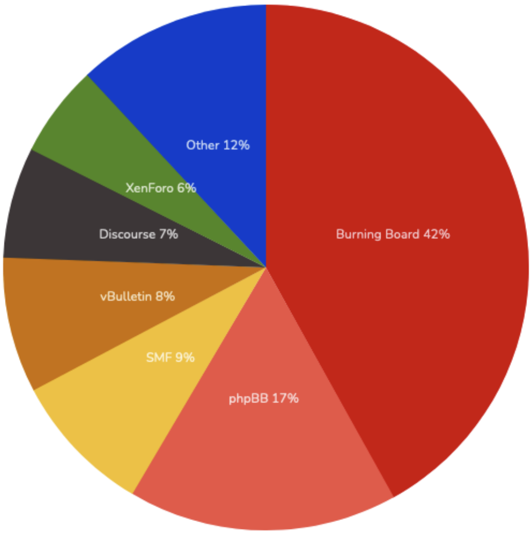
To develop a scalable web scraper, a research about possible scaling techniques is conducted and results in the finding that the CSS selectors and HTML structure differ across discussion forums. However, discussion forums, which utilize the same framework exhibit identical CSS selectors apart from few exceptions. The XPath is usually a viable alternative to CSS selectors. Nevertheless, in this particular case it is not an option as many frameworks offer numerous User Interface themes, which alter the HTML structure resulting in low scalability with XPaths.

The most crucial frameworks in Germany and worldwide were determined (see Figure 5 Usage of Discussion Forum Frameworks WorldwideFigure 6 Usage of Discussion Forum Frameworks in Germany). And by implementing web scrapers for the six most popular ones, namely phpBB, Burning Board, vBulletin, SMF, XenForo and Discourse, the web scraper is able to cover about 88,14% of German-speaking discussion forums and 76,9% worldwide. Consequently, discussion forum frameworks enable scalability in scraping discussion forums.



Source: BuiltWith Pty Ltd (2021a)

Figure 5 Usage of Discussion Forum Frameworks Worldwide



Source: BuiltWith Pty Ltd (2021b)

Figure 6 Usage of Discussion Forum Frameworks in Germany

Existing web scrapers could be studied to accelerate the forum-specific implementation, which is tested at the example of the web scraper from Ascienzo (2020). This idea is dismissed because the test emphasizes that a new implementation by the student project team is faster and more reliable. The reason for that is that other web scraping projects utilize redundant features. For instance web scraping after log in, which is not permitted because it extracts nonpublic data (IONOS 2020; Toth 2017). Moreover, the time for testing, integration and adjustment of other scrapers exceeds the time for a new implementation. Thus, an implementation by the student project team was conducted from the start.

Concluding, the six most popular frameworks were analyzed regarding their CSS selectors to identify the web scraping parameters, which enable scalability for each framework. Moreover, the selected scraping parameters were tested and corrected on the basis of the 60 most popular German-speaking discussion forums (Stenzel 2021). The HTML structure from discussion forums using vBulletin and Discourse varied widely. In these cases the CSS classes and IDs cannot be applied as effectively as with other discussion forum frameworks and the scalability with these spiders is not given.

5.2 Procedure of Web Scraping

The final problem-solution approach for web scraping is inspired by the execution presented in the literature in similar domains and use cases (see 6C.b; Guo et al. 2006,

p. 745). First, the utilized discussion forum framework is determined. Since the implemented frameworks cover almost nine in ten German-speaking frameworks, it is very likely that it is scrapable with the implemented scalable web scraper.

Because the scraper knows the applied framework, the central scraping can be executed using the breadth-first search approach, which means that the scraper begins to scrape on the search result page to identify all relevant threads and proceeds to extract all posts from those threads. This is repeated for every discussion forum provided in the input (see Figure 7 BFS and DFS Applied in Forum Web Scraping; 6C.b).

In contrast to a breadth-first approach as similarly realized by Guo et al. (2006, p. 745), a depth-first search could have been implemented. I. e. first selecting one thread and scraping every post of it and repeating that for every search result thread (see Figure 7 BFS and DFS Applied in Forum Web Scraping). This would require scraping the search result page as many times as there are threads. This puts redundant and artificial pressure on the website traffic, which should generally be avoided. If implemented carelessly and it may amount to a brute force attack in some cases and cause the server to crash (Roberts 2018). In the case of overwhelming server requests, the server can be unburdened by inserting short breaks (e. g., 10ms) between each server query, causing the web scraping to take significantly longer. The depth-first search enables the storage of posts within the thread they belong to, whereas the breadth-first search results in separate documents for threads and posts (see 6D.d; App. 6D.e).

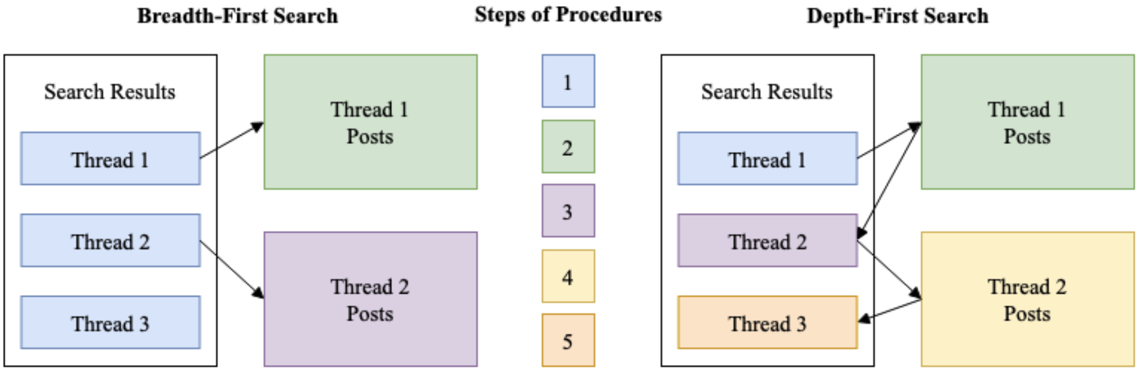


Figure 7 BFS and DFS Applied in Forum Web Scraping

In most cases, the web scraper can access the search results using the URL, which is effortless to implement and reduces server utilization (Pavkovic and Protic 2013, p. 817). Fortunately, Pavkovic and Protic (2013, p. 817) outline that although “forums have different designs, and are built on different technologies, they always have identical logic navigation that connects homepage and particular posts through discussion forum lists and threads by specific URLs". This concept was confirmed in the project research and can also be partly extended to the search URL, because it exhibits similar logic across frameworks. Nevertheless, the slight syntactic differences of the URL depending on the framework should be noted (see Table 3 Exemplary Variations of Framework-Specific Search URLs.

Since some search result pages have multiple pages, pagination must be considered and implemented in the web scraper as well. This is the case, because the HTML from the other pages is not loaded directly and hence cannot be scraped from the original page. Although most discussion forums of the same framework utilize the same pagination technique, there are variations across frameworks. For instance, vBulletin and phpBB allow pagination by manipulating the URL, e. g. in the case of vBulletin, a specific page can be targeted by appending "`&pp=&page=<PageNumber>`" to the URL, while most phpBB use the post results for pagination: "`&start=<StartSearchResultsAtPost>`" (see Tab. 3).

Framework	Discussion Forum URL of Search Result Threads

phpBB	`<BaseURL>/search.php?st=0&sk=t&sd=d&sr=posts&keywords=<SearchTerm>`
	`&start=<StartSearchResultsAtPost>`

vBulletin	`<BaseURL>/tags.php?tag==<SearchTerm>&pp=&page=<ResultPageNumber>`

Table 3 Exemplary Variations of Framework-Specific Search URLs

Other discussion forums do not utilize pagination via the URL; instead a button to call the next page must be pressed:

```
next_button = response.css(‘a.Next’).attrib[‘href’]
if next_button: #if button with link tag a and hyperlink href exists
    #follow the link and reapply the scraper
    yield response.follow(next_button, callback=self.parse_post)
```

In some cases the URL contains a search ID instead of the search term, e. g., `<BaseURL>/search.html?searchid=<SearchID>`. This prevents the scraper from using the URL to directly receive the search results. Consequently, in these cases one needs to utilize different techniques, for example through an intermediate step: A form is submitted, which signals the website to display the search results (see 6D.a). Hence, the desired response is received (see 6D.a). In few edge cases the discussion forums may not offer any search function, e. g. IGN Boards (2021). Then, the discussion forum cannot be considered.

The applied discussion forum framework may be determined in various ways, e. g. to get a manual overview, BuiltWith Technology Lookup (BuiltWith Pty Ltd 2021a) can be used. The scraper however searches the web scraping response to determine the framework. The latter can be performed because the framework's name is almost always incorporated in the HTML response in various ways. For instance, the website may link to the framework's stylesheets, use the framework's name or abbreviation in classes or IDs. The framework's name might also be present in the meta tag or is visible in the footer (see Table 4 Usage of the Framework Name in the Server Response). Because each discussion forum may be based on a different framework, this process is applied for every discussion forum in the input of discussion forum URLs (see 6C.b).

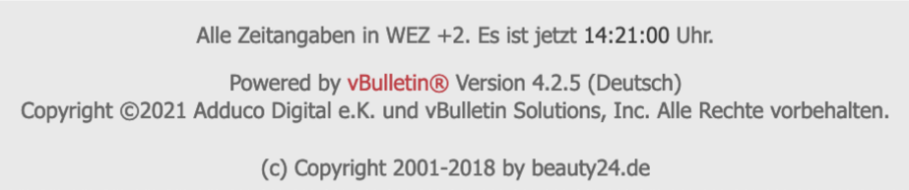
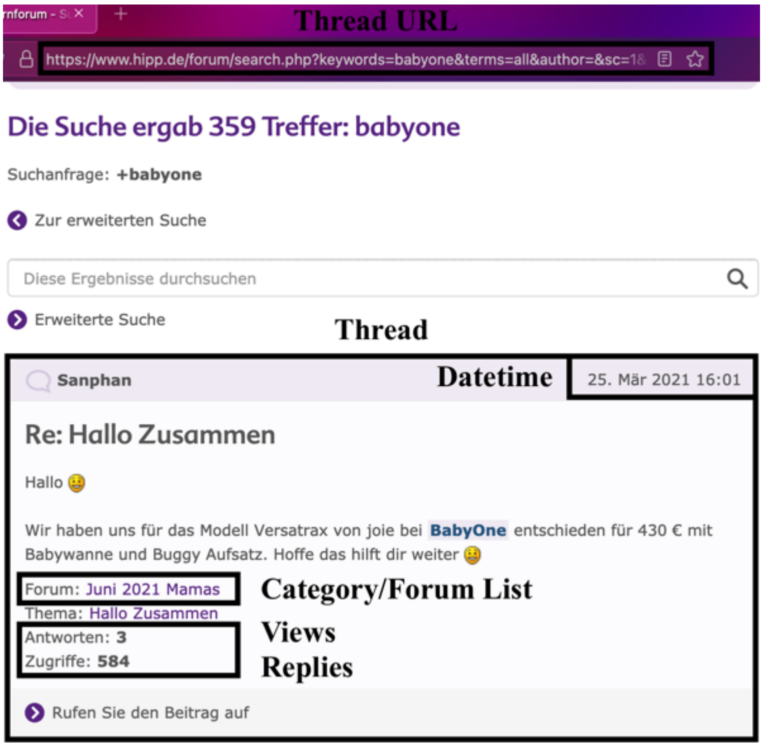
Type	Examples from Beauty Board (2021)
Stylesheets	`<link rel=”stylesheet” type=”text/css” href=”clientscript/vbulletin_css/style000071/search-rollup.css?d=1614339022”>`
Class or ID	`<a id=”vbulletinlink” href=”http://www.vbulletin=germany.com”> vBulletin® </a>`
Meta tag	`<meta name=”generator” content=”Bulletin 4.2.5”/>`
Footer	 Alle Zeitangaben in WEZ +2. Es ist jetzt 14:21:00 Uhr. Powered by vBulletin® Version 4.2.5 (Deutsch) Copyright ©2021 Adduco Digital e.K. und vBulletin Solutions, Inc. Alle Rechte vorbehalten. (c) Copyright 2001-2018 by beauty24.de

Table 4 Usage of the Framework Name in the Server Response

As an overview, the web scraping process begins by determining the discussion forum framework to enable the web scraper to scrape accordingly. Subsequently, the relevant threads are extracted on the search result page (see Figure 8). These single thread pages are scraped to extract their posts (see Figure 7; 6C.b). An exemplary spider code for the thread and post extraction is presented in the appendix (see 6D.b; 6D.c). As a site note, various web scraping techniques are utilized, which exceed the scope of this thesis but can be inspected at w3schools (2021).



https://www.hipp.de/forum/search.php?keywords=babyone&terms=all&author=&sc=1&

Die Suche ergab 359 Treffer: babyone

Suchanfrage: **+babyone**

Zur erweiterten Suche

Diese Ergebnisse durchsuchen

Erweiterte Suche

Sanphan 25. Mär 2021 16:01

Re: Hallo Zusammen

Hallo

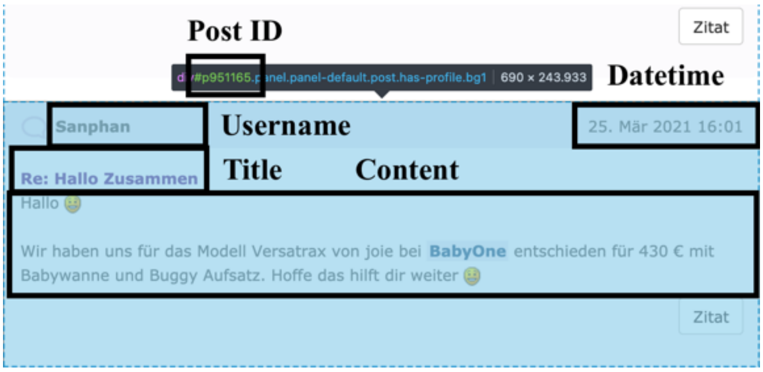
Wir haben uns für das Modell Versatrax von joie bei **BabyOne** entschieden für 430 € mit Babywanne und Buggy Aufsatz. Hoffe das hilft dir weiter

Forum: Juni 2021 Mamas
Thema: Hallo Zusammen
Antworten: **3**
Zugriffe: **584**

Rufen Sie den Beitrag auf

Source: HiPP (2021b)

Figure 8 Scraping of Search Result Threads


Post ID
Zitat
d #p951165.panel.panel-default.post.has-profile.bg1 | 690 × 243.933
Datetime
Username
Title
Content
Zitat

Sanphan 25. Mär 2021 16:01

Re: Hallo Zusammen

Hallo

Wir haben uns für das Modell Versatrax von joie bei **BabyOne** entschieden für 430 € mit Babywanne und Buggy Aufsatz. Hoffe das hilft dir weiter

Source: HiPP (2021)

Figure 9 Scraping of Posts

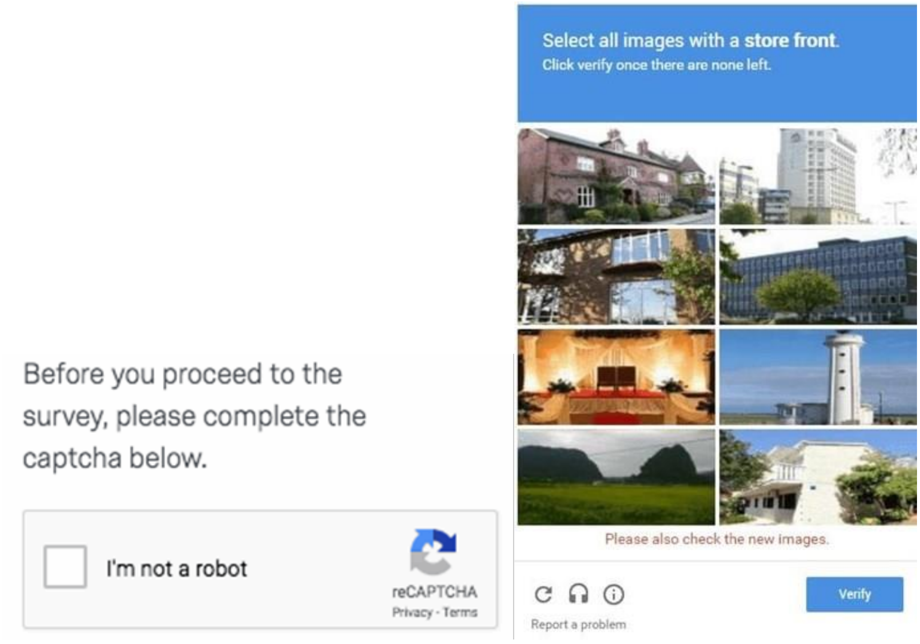
Based on the dissociation of web scrapers and crawlers (see 4), the process ‘Scrape Threads’ only applies web scraping as it collects thread data and processes the posts’ username, date, time, title, content, ID, URL and index in the applied search ranking (see 6D.c; 6D.e). By contrast, the preceding process ‘Scrape Search Results’ might be considered a mixture of web scraping and crawling. On the one hand, it utilizes scraping since it collects thread information, for instance the thread URL and ID, the number of replies and views and the discussion forum category (see 6D). On the other hand, one could argue that it displays characteristics of a web crawler because it fetches links from web pages, which are scraped thereafter. However, the implemented data extraction from this project does not act iteratively in contrast to crawlers. As a result, it does portray any risk to the website and cannot be fully classified as a crawler. Despite the request for each search result thread, the website is not overwhelmed because short breaks between each server query may be inserted. As a result, the website is not flooded with requests in a short period of time, which may allow careful scraping of discussion forums.

The significance of this categorization and strict dissociation lies in the fact that it is recommended for web crawlers to act in accordance with the rules from the robots.txt file (see 4). As mentioned previously, the point of that file is to allow e. g. search engines or (price) comparisons portals crawlers to index the pages without putting significant burden on the website traffic.

If the process is based on crawling, it should obey the robots.txt, whose legal background is worth to be examined. Most importantly, the district court Hamburg (2008) denoted that the Robots Exclusion Standard Protocols are non-binding for legal assessment. This is affirmed by Schellekens (2013) because "the mere ignoring of the robots.txt file cannot give rise to illegal access in a criminal sense." And since it is a text file, it does not represent a technical barrier (dpa 2010; Schellekens 2013; Sen 2001). However, it is worth noting that it is a strong guideline and directive, which almost every crawler obeys (de Campos et al. n.d.; Sen 2001; Ulbricht 2012; van Vessum and Rangel 2021). Accordingly, the robots.txt is not disregarded, but it may also not be completely enforced.

5.3 Limitations and Obstacles

On top of robots.txt, there are other techniques that hinder web crawling and may protect the website from Distributed Denial of Service (DDoS) attacks, where a system is flooded with requests making the server unavailable (Turk et al. 2020, p. 428). For example, websites may detect and block a single IP address, if it sends requests "periodically or within a short period of time" (Wu 2019), which can be overcome by rotating the IP address or randomizing the timing and frequency requests (Chauhan 2020). Another example is CAPTCHA, which is designed to differentiate between humans and bots and is extremely difficult to bypass (see Figure 10 Exemplary CAPTCHA). Moreover, authentication using login makes scraping more difficult and illegal (Wu 2019). However, it is worth noting that the project research only encountered the last anti-crawling technique in discussion forums few times.



Before you proceed to the survey, please complete the captcha below.

Select all images with a store front.
Click verify once there are none left.

Please also check the new images.

Source: Wu (2019)

Figure 10 Exemplary CAPTCHA

Moreover, it is illegal to disregard a website's terms of service (Bernand 2017). Thus, the terms and conditions of each discussion forum framework have to be checked, which is not economically feasible in a scalable web scraping approach. Moreover, the copying of whole websites or whole data bases to create shadow databases is not permitted as it violates the copyright (Cologne Higher Regional Court 2020; Onlinerecht Blog 2020).

6 Lessons Learned and Outlook

Concluding, the feasibility of a scalable discussion forum web scraper was validated in the project with pivoty, which extracts consumer insights. The resulting data set is the foundation of an AI-based analysis using NLP, text mining and topic modeling to derive innovation potential.

The scalability of the data extraction is based on discussion forum frameworks and their application of forum-consistent CSS selectors. On the basis of the configuration file the implementation may be extended, for instance by adding further frameworks freely. In contrast to other frameworks, vBulletin and Discourse did not have a uniform CSS structure. However, both of them offer an API (Discourse 2021; vBulletin 2021). As an outlook, one could integrate them into the data extraction, though it was not within the scope of this project.

The web scraping process begins with extracting the search result threads and proceeds with the corresponding thread posts. This approach still requires some further testing since only individual use cases were implemented. Hence, a central and all-encompassing test comprising different and unidentifiable frameworks is advisable.

References

Abdellaoui, R., Foulquie, P., Texier, N., Faviez, C., Burgun, A., and Schück, S. 2018. "Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach," *Journal of Medical Internet Research* (20:3), JMIR Publications Inc.

Akwei, J. 2019. "ContextBase - Topic Modeling," *Think Infi*. (https://rstudio-pubs-static.s3.amazonaws.com/509287_4e38623a817e4473a053d0a34f441c7c.html, accessed July 1, 2021).

Ananiadou, S., and Mcnaught, J. 2006. "Text Mining for Biology and Biomedicine," London, United Kingdom.

Ascienzo, D. 2020. "PhpBB Forum Scraper: Python-Based Web Crawlers for Scraping PhpBB Forum Posts," *GitHub*. (https://github.com/Dascienz/phpBB-forum-scraper, accessed July 2, 2021).

Beauty Board. 2021. "Suchergebnisse Für Parfüm." (https://www.beautyboard.de/tags.php?tag=parfüm, accessed July 8, 2021).

Bernand, B. 2017. "Web Scraping and Crawling Are Perfectly Legal, Right?" (https://benbernardblog.com/web-scraping-and-crawling-are-perfectly-legal-right/, accessed July 3, 2021).

Bradley, A., and James, R. J. E. 2021. "Defining the Key Issues Discussed by Problematic Gamblers on Web-Based Forums: A Data-Driven Approach," *International Gambling Studies* (21:1), Routledge, pp. 59–73.

Brownlee, J. 2019. "What Is Natural Language Processing?" (https://machinelearningmastery.com/natural-language-processing/, accessed July 1, 2021).

BuiltWith Pty Ltd. 2021a. "Forum Software Usage Distribution on the Entire Internet." (https://trends.builtwith.com/cms/forum-software/traffic/Entire-Internet, accessed June 29, 2021).

BuiltWith Pty Ltd. 2021b. "Forum Software Usage Distribution in Germany." (https://trends.builtwith.com/cms/forum-software/country/Germany, accessed June 29, 2021).

BuiltWith Pty Ltd. 2021c. "BuiltWith Technology Lookup." (https://builtwith.com/, accessed July 9, 2021).

Cai, Y., and Sun, J.-T. 2009. "Text Mining," in *Encyclopedia of Database Systems*, Boston, MA: Springer US, pp. 3061–3065. (http://link.springer.com/10.1007/978-0-387-39940-9_418).

de Campos, J., Dbr, B., and Tallent, R. (n.d.). "Ethics of Robots.Txt," *Stack Overflow*. (https://stackoverflow.com/a/999088, accessed July 3, 2021).

Chauhan, B. 2020. "How to Bypass Anti-Scraping Tools on Websites," *Datahut*.

(https://www.blog.datahut.co/post/web-scraping-how-to-bypass-anti-scraping-tools-on-websites, accessed July 4, 2021).

Cologne Higher Regional Court. 2020. *Rechtswidriges Auslesen Einer Datenbank Mittels Web-Scraper (20 241 E)*, (Vol. 6), U 128/19.

Discourse. 2021. “Discourse API Docs.” (https://docs.discourse.org/, accessed July 10, 2021).

District Court Hamburg. 2008. “Urteil: Bildersuche Suchmaschine Haftung,” *Kanzlei Dr. Bahr*. (https://www.suchmaschinen-und-recht.de/urteile/Landgericht_1-Hamburg-20080926/, accessed July 3, 2021).

dpa. 2010. “BGH-Urteil - Google Darf Weiter Minifotos Bei Bildersuche Anzeigen,” *Hannoversche Allgemeine Zeitung*. (https://www.haz.de/Nachrichten/Medien-TV/Uebersicht/Google-darf-weiter-Minifotos-bei-Bildersuche-anzeigen, accessed July 3, 2021).

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., and Fdez-Riverola, F. 2013. “Web Scraping Technologies in an API World,” *Briefings in Bioinformatics* (15:5), Oxford University Press, pp. 788–797.

Goldberg, Y. 2017. “Neural Network Methods for Natural Language Processing,” *Synthesis Lectures on Human Language Technologies* (10:1), (G. Hirst, ed.), pp. 1–309.

Google. 2021. “Robots.Txt-Spezifikationen.” (https://developers.google.com/search/docs/advanced/robots/robots_txt?hl=de, accessed July 3, 2021).

Guo, Y., Li, K., Zhang, K., and Zhang, G. 2006. “Board Forum Crawling: A Web Crawling Method for Web Forum,” in *Proceedings of the 2006 International Conference on Web Intelligence*, Hong Kong, China: Institute of Electrical and Electronics Engineers Inc., pp. 745–748.

HiPP. 2021a. “Suche Babyone,” *HiPP Baby- Und Elternforum*. (https://www.hipp.de/forum/search.php?keywords=babyone&terms=all&author=&sc=1&sf=all&sr=posts&sk=t&sd=d&st=0&ch=1000&t=0&submit=Suche, accessed July 5, 2021).

HiPP. 2021b. “Hallo Zusammen,” *HiPP Baby- Und Elternforum*. (https://www.hipp.de/forum/viewtopic.php?f=75&t=88172&p=951165&hilit=babyone#p951165, accessed July 5, 2021).

IBM. 2021. “What Is Text Mining?” (https://www.ibm.com/cloud/learn/text-mining, accessed July 1, 2021).

IGN Boards. 2021. “IGN Boards Forum.” (https://www.ignboards.com/, accessed July 9, 2021).

IONOS. 2020. “Web Scraping: What Is It and How Does It Work?” (https://www.ionos.com/digitalguide/websites/web-development/what-is-web-scraping/, accessed July 2, 2021).

J., D., and Ben. 2012. "What Is the Difference Between Web-Crawling and Web-Scraping?," *Stack Overflow*. (https://stackoverflow.com/a/4327523, accessed June 30, 2021).

Khare, S. 2017. "Topic Modeling Using Latent Dirichlet Allocation," *LinkedIn*. (https://www.linkedin.com/pulse/topic-modeling-using-latent-dirichlet-allocation-saket-khare-frm/, accessed July 1, 2021).

Khymytsia, N., Ustyianovych, T., and Dronyuk, I. 2019. "Identification and Modeling of Historiographic Data in the Content of Web Forums," in *Proceedings of the 2019 CEUR Workshop* (Vol. 2392), Lviv, Ukraine: CEUR-WS, pp. 297–308.

Koloveas, P., Chantzios, T., Tryfonopoulos, C., and Skiadopoulos, S. 2019. "A Crawler Architecture for Harvesting the Clear, Social, and Dark Web for IoT-Related Cyber-Threat Intelligence," in *Proceedings of the 2019 World Congress on Services*, Milan, Italy: Institute of Electrical and Electronics Engineers Inc., July 1, pp. 3–8.

Kremer, P. 2019. "Text Mining - Definition Und Anwendungsbeispiele," *Taod Consulting GmbH*. (https://www.taod.de/insights/text-mining-definition-und-anwendungsbeispiele/, accessed July 1, 2021).

Li, S. 2018. "Topic Modeling and Latent Dirichlet Allocation (LDA) in Python," *Towards Data Science*. (https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24, accessed July 1, 2021).

Liddy, E. D. 2001. *Natural Language Processing*, (2nd ed.), New York: Marcel Decker, Inc.

Lutkevich, B. 2021. "What Is Natural Language Processing? An Introduction to NLP." (https://searchenterpriseai.techtarget.com/definition/natural-language-processing-NLP, accessed July 1, 2021).

Mintz, I., Weisman, A., Springer, S., and Gottlieb, U. 2020. "Individuals with Back and Neck Pain on Medical Forums: What Do They Mention? What Do They Fear?," *European Journal of Pain* (24:10), Blackwell Publishing Ltd, pp. 1915–1922.

Naskar, A. 2021. "Latent Dirichlet Allocation for Beginners: A High Level Overview." (https://thinkinfi.com/latent-dirichlet-allocation-for-beginners-a-high-level-overview/, accessed July 1, 2021).

Naveen, P., Nair, P. C., and Gupta, D. 2020. "Predicting the Degree of Emotional Support in an Online Health Forum for HIV Using Data Mining Techniques," in *Lecture Notes in Electrical Engineering* (Vol. 569), Springer Verlag, pp. 81–94.

Nicolas, G., Bai, X., and Fiske, S. T. 2019. "Exploring Research-Methods Blogs in Psychology: Who Posts What About Whom, and with What Effect?," *Perspectives on Psychological Science* (14:4), SAGE Publications Inc., pp. 691–704.

Onlinerecht Blog. 2020. "Webscraping, Screenscraping Und Das

Datenbankurheberrecht," *Dury Legal Rechtsanwälte*. (https://www.dury.de/onlinerecht-blog/webscraping-screenscraping-und-das-datenbankurheberrecht, accessed July 10, 2021).

Openminted Communications. 2018. "Text Mining 101," *OpenMinTeD*. (http://openminted.eu/text-mining-101/, accessed July 2, 2021).

Palakollu, S. M. 2019. "Scrapy vs Selenium vs Beautiful Soup for Web Scraping," *Medium*. (https://medium.com/analytics-vidhya/scrapy-vs-selenium-vs-beautiful-soup-for-web-scraping-24008b6c87b8, accessed July 10, 2021).

Parsers VC. 2019. "What Is the Difference Between Web Crawling and Web Scraping?" (https://parsers.me/what-is-the-differences-between-web-crawling-and-web-scraping/, accessed June 30, 2021).

Paudyal, S. 2014. "Text Mining from Online CMS Forums - Steps of Text Mining Process."

Pavkovic, M., and Protic, J. 2013. "Intelligent Crawler for Web Forums Based on Improved Regular Expressions," in *Proceedings of the 21st Telecommunications Forum Telfor*, Belgrade, Serbia: Institute of Electrical and Electronics Engineers Inc., pp. 817–820.

Pee, L. G., Pan, S. L., Li, M., and Jia, S. 2020. "Social Informatics of Information Value Cocreation: A Case Study of Xiaomi's Online User Community," *Journal of the Association for Information Science and Technology* (71:4), John Wiley and Sons Inc., pp. 409–422.

Porter, K. 2018. "Analyzing the Dark Net Markets Subreddit for Evolutions of Tools and Trends Using LDA Topic Modeling," in *Proceedings of the 10th Digital Forensic Research Conference*, New Orleans, USA: Digital Forensic Research Workshop, pp. S87–S97.

Pretzsch, S., Muthmann, K., and Schill, A. 2012. "FODEX - Towards Generic Data Extraction from Web Forums," in *Proceedings of the 26th International Conference on Advanced Information Networking and Applications Workshops*, Fukuoka-shi, Japan: Institute of Electrical and Electronics Engineers Inc., pp. 821–826.

Rashid, A., Shoaib, U., and Shahzadsarfraz, M. 2016. "Knowledge Discovery in Database Using Intention Mining," *Science International (Lahore)* (28:6), pp. 5145–5151.

Roberts, E. 2018. "Is Web Scraping Illegal? Depends on What the Meaning of the Word Is," *Imperva*. (https://www.imperva.com/blog/is-web-scraping-illegal/, accessed July 2, 2021).

SAS Institute Inc. 2021. "What Is Natural Language Processing?," *SAS Institute Inc.* (https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html, accessed July 1, 2021).

Schäper, D. T., Lauterjung, J., and Everding, J. S. 2021a. "Pivoty - Simplifying

Innovation." (https://www.pivoty.de/, accessed June 29, 2021).

Schäper, D. T., Lauterjung, J., and Everding, J. S. 2021b. "Pivoty," *LinkedIn*. (https://www.linkedin.com/company/pivoty/, accessed June 30, 2021).

Schellekens, M. 2013. "Are Internet Robots Adequately Regulated?," *Computer Law and Security Review* (29:6), Elsevier Advanced Technology, pp. 666–675.

Sen, D. W. 2001. "Was Dürfen Suchmaschinen Indexieren?," *Digitalwelt Magazin*. (https://www.digitalwelt.org/ratgeber/seo/was-duerfen-suchmaschinen, accessed July 3, 2021).

Silge, J., and Robinson, D. 2021. "Topic Modeling - Text Mining with R." (https://www.tidytextmining.com/topicmodeling.html, accessed July 1, 2021).

Stenzel, S. 2021. "Die Größten Deutschsprachigen Internetforen," pp. 1–3. (https://www.beliebte-foren.de/top500_nach_beitraege, accessed June 29, 2021).

Törnberg, A., and Törnberg, P. 2016. "Combining CDA and Topic Modeling: Analyzing Discursive Connections between Islamophobia and Anti-Feminism on an Online Forum," *Discourse and Society* (27:4), SAGE Publications Ltd, pp. 401–422.

Toth, A. 2017. "Is Web Scraping Legal? Six Misunderstandings about Web Scraping," *Import.Io*. (https://www.import.io/post/6-misunderstandings-about-web-scraping/, accessed July 2, 2021).

Traversy Media. 2020. "Intro To Web Crawlers & Scraping with Scrapy," *YouTube*. (https://youtu.be/ALizgnSFTwQ?t=34, accessed July 4, 2021).

Turk, K., Pastrana, S., and Collier, B. 2020. "A Tight Scrape: Methodological Approaches to Cybercrime Research Data Collection in Adversarial Environments," in *Proceedings of the 5th European Symposium on Security and Privacy Workshops*, Genoa, Italy: Institute of Electrical and Electronics Engineers Inc., September 1, pp. 428–437.

Twitter. 2021a. "Twitter Premium APIs." (https://developer.twitter.com/en/products/twitter-api/premium-apis, accessed June 30, 2021).

Twitter. 2021b. "Twitter Standard v1.1 API." (https://developer.twitter.com/en/docs/twitter-api/v1, accessed June 30, 2021).

Twitter. 2021c. "Twitter Search API: Enterprise." (https://developer.twitter.com/en/docs/twitter-api/enterprise/search-api/overview, accessed June 30, 2021).

Ulbricht, D. C. 2012. "Big Data Und Recht - Wem „Gehören" Eigentlich Daten Oder Wie Google Nun Gegen SEO-Tools Vorgeht," *Recht 2.0 Internet, Social Media Und Recht*. (http://www.rechtzweinull.de/archives/623-big-data-recht-wem-gehoeren-eigentlich-daten-oder-wie-google-nun-gegen-seo-tools-vorgeht.html, accessed July 3, 2021).

vBulletin. 2021. “VBulletin 5.6.5 API.” (http://vb5support.com/resources/api/packages/vBApi.html, accessed July 10, 2021).

Vegan-Forum. 2021. “Robots.Txt,” *Vegan-Forum*. (https://vegan-forum.de/robots.txt, accessed June 30, 2021).

van Vessum, S., and Rangel, J. 2021. “Robots.Txt for SEO: The Ultimate Guide,” *ContentKing*. (https://www.contentkingapp.com/academy/robotstxt/#what-are-the-limitations-of-robots-txt, accessed July 3, 2021).

w3schools. 2021. “CSS Selectors Reference.” (https://www.w3schools.com/cssref/css_selectors.asp, accessed July 4, 2021).

Wu, J. 2019. “Five Anti-Scraping Techniques You May Encounter,” *Octoparse*. (https://www.octoparse.com/blog/5-anti-scraping-techniques-you-may-encounter#, accessed July 4, 2021).

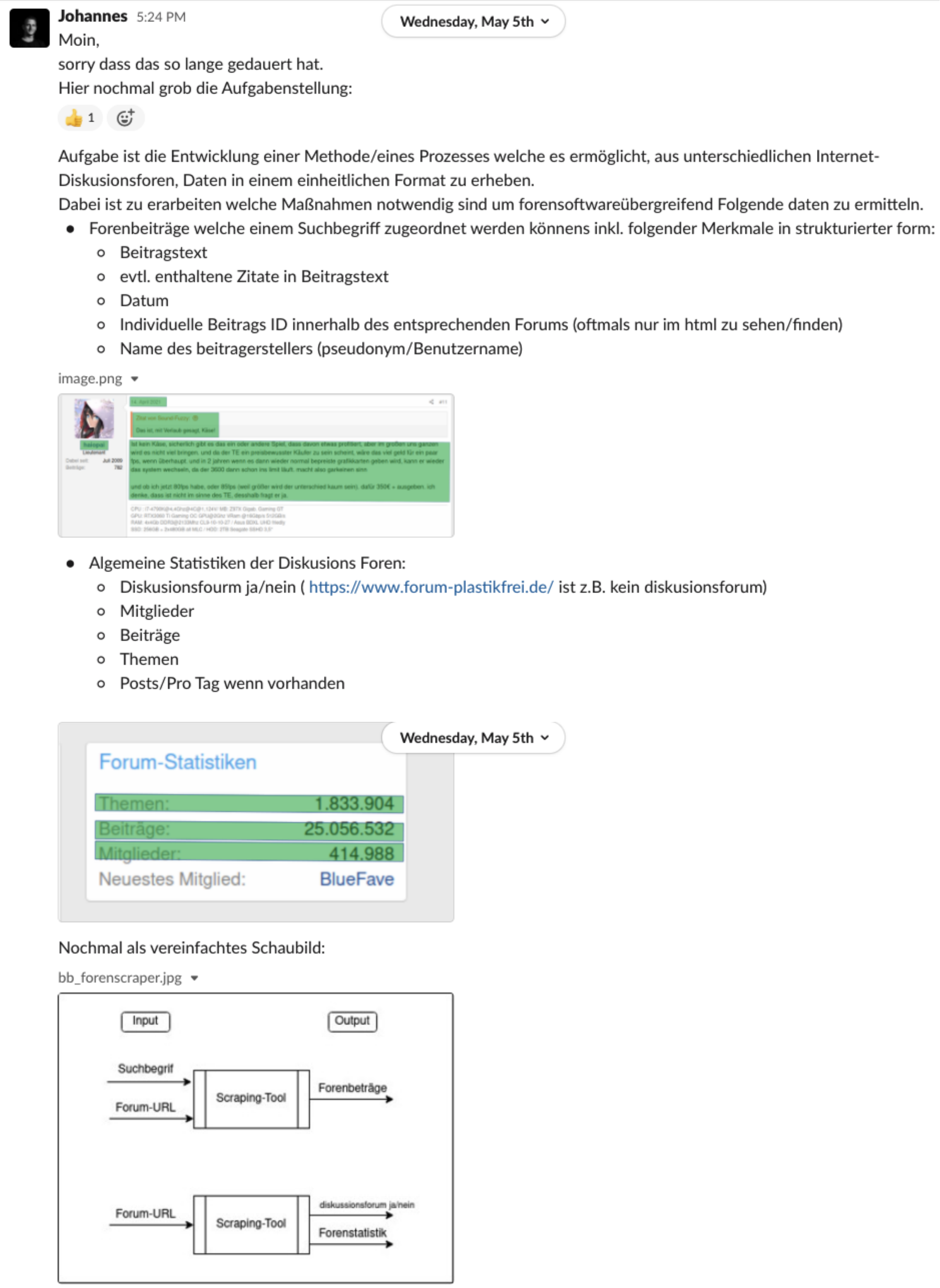
Zhao, X., Jiang, Z., and Gray, J. 2019. “Text Classification and Topic Modeling for Online Discussion Forums: An Empirical Study from the Systems Modeling Community,” in *Trends and Applications of Text Summarization Techniques*, IGI Global, pp. 151–186.

Zhou, S., and Palma, M. 2017. “A Web Scraper for Forums - Navigation and Text Extraction Methods,” KTH Royal Institute of Technology.

Appendix

A Input by pivoty

A.a Project Task Definition by pivoty



Johannes 5:24 PM

Wednesday, May 5th

Moin,
sorry dass das so lange gedauert hat.
Hier nochmal grob die Aufgabenstellung:

👍 1

Aufgabe ist die Entwicklung einer Methode/eines Prozesses welche es ermöglicht, aus unterschiedlichen Internet-Diskusionsforen, Daten in einem einheitlichen Format zu erheben.
Dabei ist zu erarbeiten welche Maßnahmen notwendig sind um forensoftwareübergreifend Folgende daten zu ermitteln.

- Forenbeiträge welche einem Suchbegriff zugeordnet werden könnens inkl. folgender Merkmale in strukturierter form:
  - Beitragstext
  - evtl. enthaltene Zitate in Beitragstext
  - Datum
  - Individuelle Beitrags ID innerhalb des entsprechenden Forums (oftmals nur im html zu sehen/finden)
  - Name des beitragerstellers (pseudonym/Benutzername)

image.png

14. April 2021 #11

haiopai
Lieutenant
Dabei seit: Juli 2009
Beiträge: 782

Zitat von Sound-Fuzzy:
Das ist, mit Verlaub gesagt, Käse!

Ist kein Käse, sicherlich gibt es das ein oder andere Spiel, dass davon etwas profitiert, aber im großen uns ganzen wird es nicht viel bringen. und da der TE ein preisbewusster Käufer zu sein scheint, wäre das viel geld für ein paar fps, wenn überhaupt. und in 2 jahren wenn es dann wieder normal bepreiste grafikkarten geben wird, kann er wieder das system wechseln, da der 3600 dann schon ins limit läuft. macht also garkeinen sinn

und ob ich jetzt 80fps habe, oder 85fps (weil größer wird der unterschied kaum sein). dafür 350€ + ausgeben. ich denke, dass ist nicht im sinne des TE, desshalb fragt er ja.

CPU : i7-4790K@4,4Ghz@4C@1,124V/ MB: Z97X Gigab. Gaming GT
GPU: RTX3060 Ti Gaming OC GPU@2Ghz VRam @16Gbps 512GB/s
RAM: 4x4Gb DDR3@2133Mhz CL9-10-10-27 / Asus BDXL UHD Hedy
SSD: 256GB + 2x480GB all MLC / HDD: 2TB Seagate SSHD 3,5"

- Algemeine Statistiken der Diskusions Foren:
  - Diskusionsfourm ja/nein (https://www.forum-plastikfrei.de/ ist z.B. kein diskusionsforum)
  - Mitglieder
  - Beiträge
  - Themen
  - Posts/Pro Tag wenn vorhanden

Wednesday, May 5th

Forum-Statistiken

Themen:	1.833.904
Beiträge:	25.056.532
Mitglieder:	414.988
Neuestes Mitglied:	BlueFave

Nochmal als vereinfachtes Schaubild:

bb_forenscraper.jpg

A.b Interview with Johannes Everding

The audio and video of the interview can be retrieved on sciebo from the link https://uni-muenster.sciebo.de/s/aHS47sClmz2Q2Z3 using ‘pivoty_interview_2021’ as a password. The link expires on the 1st of October 2021. The transcript of the interview is provided here:

Interviewer (I): Leo Giesen

Respondent (R): Johannes Everding, co-founder of pivoty

00:00 I: Just to have it on tape. I'm recording and that's okay with you.

00:08 R: Yeah.

00:09 I: Great. Okay. So first I've got some questions about your business model then about the resources you utilize or access and about the web scraping process and the AI. Okay. So first, what does pivoty do? What's the basic idea of the business model?

00:32 R: Yeah, so the basic idea… Let’s start by the first idea we had was to help companies to improve or just to build an innovation process. If they have an innovation process to improve it or just to implement an innovation process. Yeah, but in some interviews we came to the decision that most companies… The main problem is not to have no process or to do an ideation. It's more to get some data to improve the decision-making and so we came to the point to say: “Okay, let's look on the internet which data we could use”. And then say to test yeah, like, like discussion forums to search for terms companies are interested in. And to make a topic modeling to show which topics are there. And to classify problems or solutions, because there are a lot of people on the internet, who tackle their own problems with all solutions and often the solutions are much better than things you can buy. And so that could be something which companies are interested in. And so we decided to build a platform, where companies can put in a search term and find results in terms of innovation.

02:48 I: Oh, I knew about just like general insights, but it's really helpful for me that you clarified: okay, there are different ideas and adjustments, maybe to certain products that the company can do based on the ideas of the customer. And do you solely... I imagine you focus on the consumer, not the customer only.

03:25 R: So we don't focus actively, but the data on the internet [is] best for consumer products, where the consumer is researching and thinking about it before he buys something. So for example, a laptop is a thing people researched, which is the best laptop.

So, and another example, which would be not so good is let's say Zahnpasta…. Toothbrush.

04:11 I: Toothpaste

04:14 R: Yeah, you just buy it, so you don't do any research. So there is no discussion on the internet about it. So we need consumer goods, where people discuss on the internet.

04:30 I: So others it's probably based on the product probably. Yeah. And I imagine there are various other customers… businesses, who also utilize web scraping what data extraction in some form and implement an AI based algorithm to gain some insights or get results. Do you have a specific unique selling point or how could pivot team stand out?

05:08 R: Yes, so the other companies, who are doing quite similar things are social listening companies, for example. They look for example your brand on social media. So they watch Adidas on social media and then screen what are the people talking about? And do something like sentiment analysis and so on. So relatively brand specific, not directly on product base. And it's on another point in the product development process. It’s more at the end, so the product is already finished and we want to watch: What is the market talking about? Okay, so we are more at the front and before the very beginning before the ideation. So there's a new place where you want to, to get in as a company. And you want to understand it and so you do some research and based on this research, you go into ideation and that's more or focused. And yes, the unique selling point is… Yeah, the classification we do, so problem-solutions and best practice and so on that is something nobody else is doing actively.

06:48 I: Okay, so you are just positioning yourself in a different part of the process and you have a slightly different focus. Okay, great. Have you done a proof of concept or is there something in that or are you still working on one?

07:10 R: Yes. So our proof of concept is actually mostly based on interviews or let's say that it was our first proof of concept. And we also did some pilot projects with [pause] three projects with two companies. Or four projects and yeah, we do a lot by hand. So, classification: we don't have a big data set to learn classifiers based to 90% or so. So we do it by hand, so our data set is growing so we can build a lengthy classification algorithm and so on. And to do research by hand but supported by a lot of paths and scripts. Topic modeling is mostly based on clustering algorithms and so on. And do the research, put it in the Excel file and give it to the companies and talking with them. What is good? What is bad? And so on. For example, sentiment analysis, we also did sentiment analysis. But the feedback from the company says: “Yes, it's nice. We give it to our marketing guys. But for innovation process, it's not so important”. And so proof of concept is a lot of interviews and talking with other people.

08:54 I. Okay, great. You mentioned sentiment analysis and I was wondering if that is part of your process. Though, you've already answered that. Apart from sentiment analysis, Natural Language Processing, text mining and topic modeling play a role in your AI based process?

09:21 R: So we have three tasks to do. So first we do clustering, so topic modeling to see which topics are in the data. And then we do a summarizing part to give a quick look into the topics. So this topic is about XY and so on. So the user can decide: “Is it interesting for me”, because maybe a lot of topics are there and I'm just interested in one. So a quick overview and then inside this topic, we classify the data points by best practice, problem solution and context. So you can filter and say: “I just want to see the problems. Maybe I have a solution to solve those problems. So I have a new product.” And that's yeah, that's the main thing.

10:30 I: Okay, so a mixture of various …

10:38 R: Two points that are unsupervised learning and one is supervised learning.

10:52 I: Okay, NLP or Natural Language Processing, text mining and topic modeling are unsupervised, I believe.

11:04 R: Hmm, let's say NLP is a whole field, so Natural Language Processing is everything you do with text.

11:10 I: Yeah, and there's some text classification in supervised learning. That's true.

11:18 R: And mainly clustering is unsupervised and classification often is a supervisor.

11:25 I: Yeah, that's true. Okay, so those different AI fields overlap in some way. Okay, just a general question: So do you use software for the AI-based process at the moment? And if yes, what is it called?

11:57 R: Yes, so the very first topic modeling was with WordStat.

12:07 I: How do you spell that?

12:12 R: Word like word and S T A T. But actually we use a Python script and use pre-trained models for embeddings and clustering. A lot of experiments; HBD scan is what actually works best.

12:46 I: Okay, so this is still in process and will be optimized later on. Okay, now a bit more about the resources and web scraping. On LinkedIn, you say that you draw on a variety of online sources, like social networks, forums, blogs, and product reviews to systematically derive inspiration for new products and services. Have you considered other possible information sources to gain customer insights? For example, surveys, social benchmarking or company internal resources?

13:28 R: Yeah, so our very first idea we talked about was the whole process: To build a platform with the feature, where you can do online surveys. So when you can get direct customer interactions. But the feedback in interviews was that on one side companies already have this kind of infrastructure to do their own service and on the other side, a lot of competitions there is, like SurveyMonkey and so on. And you have the problem that in a survey the data is biased. So the opposite is very nice and don't want to say that you are building crap and then just say: “Yeah, I would buy it really”. Really at the end that's just a lie. And so, we did some research. So what we are doing also has the name netnography. It's from … Das Beobachten von Gruppen, also das Gruppenverhalten. Und Netnographie ist dann das Beobachten von Gruppen im Internet in ihrer natürlichen Umgebung. So you get unbiased data.

15:07 I: Yeah, okay. So you are able to extract data of higher quality from forums because the users do not express their opinion in a supervised context.

15:29 R: Yeah, so quality in this special case. Yeah, I wouldn't say that interviews are less quality… of a bad quality, if they are well done. But it's, we use other kinds of data.

15:47 I: Okay, so it's usually more biased, if you do a survey. Okay. Because on LinkedIn, you say, also say unbiased customer insights though of course every voice of a customer is slightly biased, but you are able to extract relatively unbiased insights.

16:17 R: Yes, so every customer insight is biased. You're right, in this case. So, if you have bought a product. So you already did the decision to say this product is good because I have bought it. So there is already a bias, but not bias to you. So if you ask me: “Would you buy my product?” So I don't want to … So I want to be a nice guy and say: “Yeah, I like your product very well.”

17:05 I: Okay. So it's more subjective, not unbiased. Is that what you're trying to say?

17:14 R: Yes, the current word is just ‘positivity bias’. You can Google it, maybe. So in surveys mostly, the guys are more positive than they really are.

17:45 I: Okay, great. Now over to web scraping. The goal of web scraping is to extract [pause]. Give me one sec. The goal of the web scraping is to extract customer insights or consumer insights. What qualifies as an insight you mentioned it's about ideas from the customer or how they view a certain product. Does an insight have to be new, unexpectedly relevant or does it need to include any specific, clear course of action. What exactly are you looking for as insights? I'm not sure. I understand. So we are searching for data based on the search team our customer is giving to us. So we are scraping every post, who contains the search term. And the decision, if this is relevant or not is by our user. So we just do a nice and quick-to-understand analysis dashboard.

19:02 I: Okay, so since it's relevant to the customer or user on the online forum, it might also impact the … or it might be important to the company as well to adjust maybe their course of action.

19:24 R: Yes, so when a company wants to develop a new [pause] let's say bicycle and you want to improve a special thing on this bicycle. And then you search for this special thing in a bike forum. So every post is relevant, which contains the search term. So for our scraping algorithm, but maybe every [inaudible] users saying, yeah, I like this power from company

20:05 I: XY

20:06 R: XY, but it's too heavy. So then a conclusion for the company could be: “Yeah, we have to get lighter”. Okay. But maybe that's not the target for the company. They just want to add features and don't want to get lighter or so on. And then this post may be not relevant, but then the topic ‘lightweight’ is not relevant. So they don't go deeper in this topic. So the decision if it's relevant or not is by our user. We just provide the data.

20:47 I: Yes, okay. But still to the company the customer opinion might still be worth something. Maybe the bike company could start like… build a new bicycle, which is even lighter or may be able to create add-ons for certain bicycles, if some requests new features. Okay, could you just confirm, is it feasible to build a scalable web scraper, to extract valuable information out of discussion forums as we've built in?

21:31 R: Yeah, so it's a big topic for us. Data acquisition is quite challenging, especially a big amount of data. And so when you think that maybe 100 of our customers want to go inside the same forum - that's a lot of calls in this forum. So we have to look to don't do any [pause]. I don't know the word. Too lot request. Too much request. So web scraping and scalability is I would say; I'll call it. Okay, it's not perfect. But we are also thinking about maybe other solutions. Like companies like Bosch or Telecom have own forums with their own community to provide our service. So the analysis to their data. Yes, it's a possibility.

22:48 I: So you could just like export their forum posts from their databases.

22:53: Yep, directly. Continuous web scraping and safe the data in our own database to provide them. But there are some legal questions about that. So we are just in development of these questions. So web scraping is in my opinion, a good way to... for prototyping and for productive products, but to scale up we have to consider other solutions as well, which might be better.

23:43 I: Okay, so it's good for the start where you can scale slightly, but there might be some problems as we can encounter in the project. For example, two frameworks did not allow for great scalability.

24:03 R: So, yep. Technically, I think that's all not that big challenge, but you also have to consider some moral [pause].

24:16 I: Ethical questions.

24:20 R: Yeah, because if you have a small form and some other company is thinking to do millions of requests in 10 seconds. Yeah, I can totally understand that. You not so happy to about that. So we have to be a little bit careful there.

24:44 I: Okay, so are there any other problems you or we encountered in the implementation of the scalable web scraping of forums?

24:57 R: Yeah, I mean it's not especially for web scraping. I mean, the whole big data or a lot of data management is a thing. So extract, transform, learn.

25:24 I: ETL. Extract, transform, load.

25:27 R: Load nicht learn. Is a thing when you are working with data, but it's not especially on scraping. So yeah, you have to take care of your data, that they are clean and yeah, scheduling tasks and asynchronous tasks in your back end and so on. But it's, I would say it’s like every other data mining process.

26:02 I: So you mentioned that web scraping is good for starting out. When you try to scale, will you or you could consider using a service for web scraping. So for example ‘zyte’ offers data extractions of forums. Have you considered that or have you tried using that?

26:28 R: Yes, we tested some services in case of data acquisition of social media data. We also used some direct APIs from Twitter. We have to go into deeper research for that, but the first book was not so satisfying on external services in case of social media.

27:12 I: Low data quality, too pricy? What...

27:18 R: Quality. So this is the data were not consistent and not complete. So if I go to the website and we're looking for the data. So it was not the same feedback, but maybe it was a bad service or too big technically challenge in the case of Facebook data. Okay, so we go back to forums because with our own scrapers it was ok so. But we are [pause]. Wir sind gewillt dafür zu bezahlen. If there is a good API, we can pay for data. We also would pay for data. So if that answers the question.

28:20 I: I think that answers it very well. Okay, so now some concluding questions: Where do we stand in the project? What did we accomplish so far and what is there left to implement?

28:33 R: Yeah, I would say we are at the very beginning. So we started a half year ago with this idea of the innovation platform and did a little bit of pivoty with this data acquisition idea. And yeah, I think we have a proof of concept to say that there are companies, who want to buy and we also have know-how how to solve the single steps, but we are not quite ready to go into production with our process and our software. So we now have to build the software itself on a production-ready quality. But also have to clarify some tasks. Let's say [pause] Darstellungsart. So what our customers want. So, we are still doing some pilot projects by hand to just to find out what kind of data and what kind of information are most of interests.

30:08 I: Okay, so we started out at the beginning of the web scraping process and then worked our way through various forum frameworks and the possibility of scaling.

30:24 R: Yeah, so at the end, your product is the first step to scale up a little bit. So it's not the big scale-up which will work in let's say three years. But today, it's the first step from last single use scripts to a more generalized solution and after will come the next step. Ok, so in small taps we iterate to production.

30:58 I: Okay, so that's great; thank you for the answer. Are you worried about GDPR, so General Data Protection Regulation, privacy laws, or any of the sort?

31:11 R: Yeah, so the honest Antwort is: I don't know. But I think nobody knows. It's very chaotic. But in general, we don't really process any personality data. So we focus on date and content of the posts and maybe the name, the username, but mostly it's already a pseudonym. And I think we will maybe clarify, on demand, is it a man or a woman. And then put the username completely in trash. So don't save it. So I think this will not be a big problem.

32:10 I. Okay. So you just don't store any personal data. Though, you extract the value, if there is [pause] if you can extract the value out of a customer username out of a forum. Okay, thank you very much for the interview. If you've got any final remarks or reviews about our work. Or you can also give them to us later on in a private session.

32:51 R: I think about you, we can talk again, if you send me the document and I have read it. And maybe after the presentation of the code and so on. But I think, you build a good basis to improve further.

33:10 I: Thank you, I'm very flattered. Okay, I think that concludes our interview. Thank you very much again for making this possible.

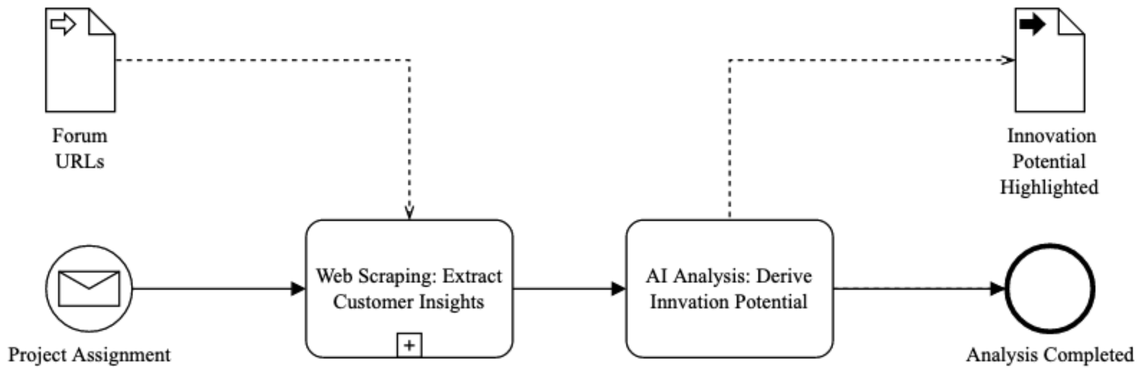
B Exemplary robots.txt from Vegan-Forum (2021)

```
“User-agent: *

Disallow: /memberlist.php

Disallow: /memberlist.php*

Disallow: /search.php

Disallow: /search.php*

Disallow: /ucp.php

Disallow: /ucp.php*

Disallow: /posting.php

Disallow: /posting.php*
```

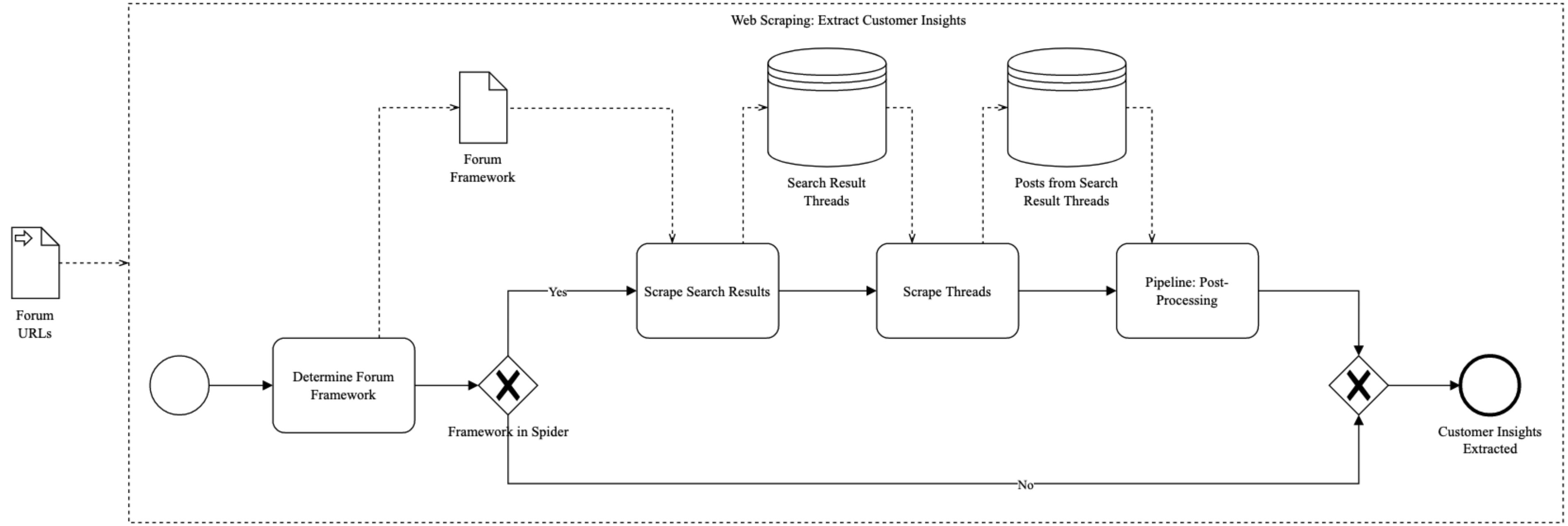
```
Disallow: /report.php

Disallow: /report.php*

Disallow: /viewonline.php

Disallow: /download.php

Disallow: /download.php*

Disallow: /*?sid=*”
```

C Processes

C.a Innovation Potential Derivation Process


Forum URLs
Project Assignment
Web Scraping: Extract Customer Insights
AI Analysis: Derive Innvation Potential
Innovation Potential Highlighted
Analysis Completed

C.b Web Scraping Process


Web Scraping: Extract Customer Insights
Forum URLs
Forum Framework
Search Result Threads
Posts from Search Result Threads
Determine Forum Framework
Framework in Spider
Yes
No
Scrape Search Results
Scrape Threads
Pipeline: Post-Processing
Customer Insights Extracted

D Implementation

D.a Form Submission

```
import scrapy

class Search_Spider(scrapy.Spider):
    name='searchagent'
    #Exemplary URLs to send the form to
    start_urls = ['https://www.dogforum.de', 'https://securitytreff.de/']

    search_pattern = ['hund','katze'] #search terms

    def parse(self, response):
        for pattern in self.search_pattern:
            data = {'q':pattern}
            yield scrapy.FormRequest.from_response(response, formdata=data,
               formcss='div.pageHeaderSearch > form',
                callback=self.parse_something)


    def parse_something(self,response):
        print("response_ps: "+str(response)) #print response
        yield {
            "title":response.css('div.section.sectionContainerList').get()
        }  #set response
```

D.b Exemplary Extraction of Threads

```
# This spider is applied to forums based on phpBB
from scrapy import Request, Spider
class Spider_threads_phpBB(Spider):
    # run the following command to store the output: scrapy crawl
    threads_phpBB -o ../../output/phpBB_threads.json
    name = 'threads_phpBB'
    start_urls = [
    'https://www.hipp.de/forum/search.php?keywords=babyone&terms=all&author=
    &sc=1&sf=all&sr=posts&sk=t&sd=d&st=0&ch=1000&t=0&submit=Suche',...
    ]

    def parse(self, response):
        try:
            for article in response.css('div.search.post'):
                yield {
                    'username': article.css('dl.postprofile dt.author
                    a::text').get(),
                    'datetime': article.css('dl.postprofile
                    dd::text').get().replace('on ', ''), #processing
                    'title': article.css('dl.postprofile dd a::text')
                     [-1].get(), # get last elem using [-1]
                    'content': article.css('div.postbody div.content').get(),
                    'category': article.css('dl.postprofile dd
                    a::text').get(),
                    'replies': article.css('dl.postprofile dd
                    strong::text').get(),
                    'views': article.css('dl.postprofile dd strong::text')
                     [-1].get()
                }
        except:
            yield {
                'yield': "error"
            }
```

D.c Exemplary Extraction of Posts

```
# This spider is applied to forums based on phpBB
from scrapy import Request, Spider
class Spider_posts_phpBB(Spider):
    # run the following command to store the output: scrapy crawl posts_phpBB
-o ../../output/phpBB_posts.json
    name = 'posts_phpBB'
    start_urls = [
    'https://www.hipp.de/forum/viewtopic.php?f=75&t=88172&p=951165&hilit=
    babyone#p951165', ...
    ]
    def parse(self, response):
        try:
            index = 0
            for article in response.css('div.panel.panel-default.post.has-
            profile'):
                yield {
                    'test': article.css('div.panel.panel-default.post.has-
                    profile strong::text').get()
                    .replace("\n","").replace("\t", ""), # post-processing
                    'username': article.css('strong::text').get()
                    .replace("\n", "").replace("\t", ""),
                    'datetime': article.css('div.pull-right::text').get(),
                    'title': article.css('strong a::text').get(),
                    'content': article.css('div.content').get(),
                    'post_id': article.css('div.panel.panel-default.post.has-
                    profile::attr(id)').get(),
                    'index': index
                }
                index += 1
        except:
            yield {
                'yield': "error"
            }
```

D.d Exemplary Output of Threads

```
[
  {
    "replies": "55",
    "views": "34099",
    "thread_id": "782905",
    "thread_url": "https://www.eltern.de/foren/schwangerschaftsguide/782905-
    kinderwagen-teutonia-fazit.html?highlight=babyone",
    "category": "Schwangerschaftsguide"
  }, ...
]
```

D.e Exemplary Output of Posts

```
[
  {
    "index": 1,
    "username": "Juju1982",
    "datetime": "06.10.2012 18:20",
    "title": " Re: Kinderwagenthema",
    "content": "Kenne mich mit dem Hartan gar nicht aus, kann Dir also nicht
    wirklich weiterhelfen, leider. \n \nWegen der Falt- oder Kombitasche
    w\u00fcrde ich mal in live und in Farbe testen gehen. Vielleicht bei
    Babyone oder in einem Babyfachgesch\u00e4ft? \n \nW\u00fcrde
    wahrscheinlich die nehmen, die sich stabiler tragen l\u00e4sst und
    gr\u00f6\u00dfer ist. Je nachdem wie zierlich Du bist, spielt ja auch das
    Gewicht eine Rolle. \n \nIn eine gr\u00f6\u00dfere, d\u00fcnnere Tasche
    kann man ja immer noch eine Extramatratze und einen warmen Fu\u00dfsack
    reinlegen, denke ich. Oder?",
    "post_id": "22401661"
  }, ...
]
```