

PROBE INTO COMPANY REVIEWS:
AN APPROACH TO INSPECT WORKSPACE ENVIRONMENT
Project report submitted in partial fulfilment of the requirement of degree of
BACHELOR OF TECHNOLOGY
IN
ELECTRONICS AND COMMUNICATION ENGINEERING

By

Vishal Srivastava (171062)

Akshat Gupta (171074)

UNDER THE GUIDANCE OF

Mr. Pardeep Garg



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

MAY 2021

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our teacher and mentor **Mr. Pardeep Garg** who gave us the golden opportunity to do this project on the topic **PROBE INTO COMPANY REVIEWS: AN APPROACH TO INSPECT WORKSPACE ENVIRONMENT** which also helped us in doing a lot of Research and we came to know about so many new things. We are thankful to him.

A handwritten signature in blue ink that reads "Vishal". The signature is written in a cursive style with a horizontal line underneath the name.

Vishal Srivastava (171062)

A handwritten signature in blue ink that reads "Akshat Gupta". The signature is written in a cursive style with a horizontal line underneath the name.

Akshat Gupta (171074)

DECLARATION

We hereby declare that the work presented in this report entitled “**PROBE INTO COMPANY REVIEWS:AN APPROACH TO INSPECT WORKSPACE ENVIRONMENT**” in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Electronics and Communication, submitted in the department of Electronics and Communication, Jaypee University of Information Technology, Wagnaghat, is an authentic record of our own work carried out over a period from January 2021 to May 2021 under the supervision of **Mr. Pardeep Garg**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

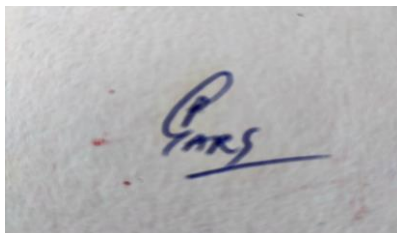


Vishal Srivastava, 171062



Akshat Gupta, 171074

This is to certify that the above statement made by the candidate is true to the best of my knowledge.



Mr. Pardeep Garg

Department of ECE

JUIT

TABLE OF CONTENT

Abstract	1
Aims and Objectives	2
Chapter-1 Introduction	3
1.1 Introduction to the project	3
1.1.1 Project Workflow	4
1.2 Literature review	4
1.2.1 General facts about Web Scraping	4
1.2.2 Purpose of Web Scraping	5
1.2.3 Analysis and research	5
1.2.4 Social Mining and Sentiment Analysis	6
1.2.5 Automatic data collection	6
Chapter-2 Legal Aspects of Data Extraction	7
2.1 Law areas for Web Scraper defence action	7
2.1.1 Website terms and conditions	8
2.1.2 Copyright, Intellectual Property Rights	9
2.1.3 Database rights	10
2.1.4 Trademarks	11
2.1.5 Data protection	12
2.1.6 Criminal Damage	13
2.2 General Legal Aspects	14
2.3 Conclusion	15
Chapter-3 Methods of Web Scraping & Available Software Tools	16
3.1 Methods of Web Scraping	16
3.1.1 Manual Scraping	16
3.2.2 HTML Parsing	16
3.3.3 DOM Parsing	17
3.4.4 XPath	18
3.4.5 API's	18
3.2 Available Software Tools	19
3.2.1 Cloud Software	19
3.2.2 Desktop Software	20
3.2.3 Programming libraries	20
3.2.4 Web Scraper	20

3.3	Conclusion	21
Chapter-4 Hands on Web Scraping & Implementation		22
4.1	Task	23
4.2	Prerequisites	24
4.3	Selected Software	25
4.4	Preparation	25
4.5	Execution	26
4.6	Results	27
4.7	Conclusion	28
Chapter-5 Data Pre-Processing		
5.1	Information into data	29
5.1.1	Data Cleaning	29
5.1.2	Data Transformation	30
5.1.3	Data Reduction	31
Chapter-6 Understanding Employees Reviews Using Sentiment Analysis		32
6.1	Topic Modelling	32
6.1.1	Identify Patterns Among Positive and Negative Employee Reviews	32
6.2	Observed Topics	33
6.2.1	Observed Topics Among Pros and Cons in Employee Reviews	33
6.3	Visualizing	33
6.3.1	Word Cloud	33
6.3.1.1	Word Cloud - Positive Reviews	34
6.3.1.2	Word Cloud - Negative Reviews	34
Chapter-7 Conclusion & Future Directions		34
7.1	Conclusion	34
7.2	Future Scope	35
References		35

LIST OF FIGURES

S.NO.	NAME OF FIGURE
1.1	Workflow
3.1	HTML Parsing
3.2	DOM Parsing
3.3	XPath
4.1	Target Website
4.2	Data Restructuring Format
4.3.1	CODE Installing scrapy libraries
4.3.2	CODE Importing scrapy
4.3.3	Code Starting Scrapy Project
4.3.4	Code creating scrapy spider
4.3.5	Code Setting Up URL
4.3.6	Code Spider indeed review
4.3.7	Code items.py
5.1	CODE Convert “?” to Nan
5.2	CODE Count missing values in each column
5.3	CODE to check the data type

List of Tables

S. No	Name of Tables
4.1	Company_profile.py output

ABSTRACT

In enhancing dynamics during the enrollment methodology, association analysis is important and can select the state of the workspace. In this project, using a technique called web scraping, we will create datasets. Basically, the web scraper assembles data from a local movement into sorted databases, for example, .csv documents. Web Scraping involves evaluations to be performed on data that is not provided in a sorted out setup from now on this analysis, we will scratch feedback for the IT associations that have onceover used Scrapy to save up to 5,000+ overviews of the data as contained in the data key to 'Indeed's The Top Rated Workplaces in 2020.' The true viewpoint, both positive and negative aspects, is taken into account here. A few incidents are equally thought of as to the real problems. The organizing norms and procedures of the Web Scraper are segregated, it tells how to organize a functioning Scraper. Additional evaluation of data is completed. Web Scrappers have a far reaching array of devices to browse. The initiative does not involve only the basic strategy and execution of the game. The legal piece of the profession evaluated and important advancements perceived by all Web Scraping adventures should be the legal piece.

Aims and Objectives

1. To ethically study and collect data through the web scraping process and to analyze it to give a better understanding of the organization and its working environment.
2. A web scraper will be created to automate the process of extracting data and information from Indeed.com and analyze the extracted text and visualize its content in order to do this.

The major objectives of this project are:

- To gain a better understanding of Internet data, web scraping methods, web scrapers, and data analysis.
- • Build a web scraper that crawls the internet and extracts feedback in form of data from indeed.com.
- • To keep track of the results in a database so that they can be worked upon later.
- To perform analyses on the obtained data and Visualize its outputs.

The deliverables of this project are:

- Web scraper
- Database
- Analysis and Visualized outputs
- Presentation
- Final report

CHAPTER-1

INTRODUCTION

The World Wide Web consists of an interlinked data system that is introduced to customers via sites. The way we provide, collect, and distribute information has completely altered the Internet. The measure of the data introduced evolves continuously.

With this disorderly growth, it is not conceivable to physically track and record every available source at this point. The second is when Web Scraping moved forward. In contrast to manual information extraction, robotized strategies allow the assortment of a gigantic measure of Web information. Another term, along with web scraping, turned out to be important meta data. A huge variety of information acquired by Web Scraping permits the investigation of Meta Data.

INTRODUCTION TO THE PROJECT

During the enrollment process, the organizational survey is fundamental to improving dynamics and can decide the nature of the workspace. We will make datasets using a technique called web Scratching in this investigation. Basically, the web scrubber collects information from site progression into organized databases, such as .csv documents. Web Scraping enables examinations to be performed on data that is not given in an organized configuration as of now. In this review, we will scratch surveys for the IT organizations that made it to Indeed's The Top-Rated Workplaces in 2020 using Scrapy to spare up to 5,000+ audits with the information as found in the data key. The legitimate viewpoint here both positive and negative sides are taken into account. In addition, a few instances are considered with regard to the legitimate issues. The planning norms and strategies of the Web Scraper are differentiated, it tells how a working Scraper is structured. Further research into information is conducted.

In this project, we collect datasets using a method called web Scratching. Web Scratching is essentially the way to collect data (such as writings) from site progression to organized databases, such as .csv records. Web Scratching allows examinations to be done on data that

has not been given in an organized arrangement as of now.

In this review, representative audits for the IT organizations that made it to Indeed's 50 Best Corporate Rundown will be scratched. Understanding the expectations and feelings associated with positive and negative audits (for example Conclusion Analysis). Understanding the subjects of positive and adverse audits (for example Subject Modelling) For each of the retailers on the Top-Rated Workplaces: The Top 50 List, Scrapy (a Python framework) was used to collect employee reviews.

This project is divided into four phases, as following

Phase 1 - Legal and ethical aspect to data scraping

Phase 2 - Structuring and optimizing the process of web scraping

Phase 3 - Designing and designing of web scraper

Phase 4 - Data restructuring and analysis

PROJECT WORKFLOW

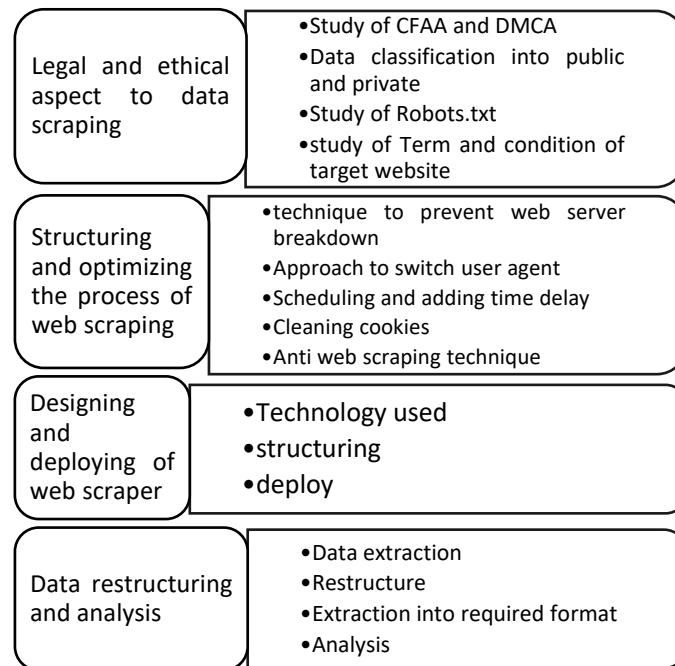


Figure 1.1 Workflow

Work in Phase – 1

Understanding digital laws and contextual analysis to frame a legitimate and moral establishment and to set boundaries to ensure that information is used and prevent its misuse.

The scraping of web data on the Robot.txt record gives better results and reduces the chances of "Access denied" to the website.

A general review of the terms and conditions of similar websites can lead to a summarized set of rules for the web scraper.

Work in Phase – 2

The organization and streamlining procedures for the equivalent are set up as follows by achieving an establishment and many general guidelines in Phase-1 for information Scratching in stage 2.

Algorithms for time booking to add arbitrary postponements to prevent breakdown and over-burdening of the web server.

Way to deal with switching client experts and system game plan to convey the heap.

Cleaning treatments and building up responses to prevent the site from being hostile to web
Scratching strategies.

Work in Phase – 3

Understanding the planning standards and using python and Scrapy to organize the web scraper. Scrapy's venture structure, making arachnids (crawlers), compartments of items (storing scratched information), building pipelines, middleware, insect settings and organization.

Work in Phase – 4

Removing and rebuilding information. Json and .csv configuration to perform exams using matplotlib python library and additionally in RStudio to be used in API to set yields for universally useful.

CHAPTER 2

LITERATURE REVIEW

[1] Analysis of factors influencing office workplace planning and design in corporate facilities. Mohammad A. Hassanain, Journal of Building Appraisals, pp. 183-197

We referred to this paper and consider this as a base of our project. We learnt many aspects of our project from this research paper including legal as well as technical part.

[2] User Evaluation of the Work Environment by Jacquelin C. Vischer, Gustave Nicolas-Fischer, pp. 73-96

We referred to this paper and it is a building block of our project. We learnt how to move forward in the project and to follow what steps after reading this paper.

[3] An Overview of the Influence of Physical Office Environments Towards Employee. A.A. Saleh, Procedia Engineering, vol. 20, pp. 262-268

We referred to this paper and got to understand the influence of physical presence of employees in the offices and its effect on the work environment.

[4] Workplace environment, employee satisfaction and intent to stay. Deepak Bangwal, Prakash Tiwari, International Journal of Contemporary Hospitality Management, pp. 268-284

We referred to this paper and got to understand the positive side of a workplace and what kind of things make an employee stay in an organization.

[5] Towards an Environmental Psychology of Workspace: How People are Affected by Environments for Work. Jacquelin C. Vischer, Architectural Science Review, vol. 51, pp. 97-108

We referred to this paper and got to know what kind of attributes and results can be derived for our project.

[6] A model of workspace environment satisfaction by Ying Hua, International Journal of Facility Management, vol.1, No.2

We referred to this paper and got to know what kind of keywords we can set in our model to look for in our project.

Combined Summary of All the Research Papers Mentioned Above and How Are They Incorporated in the Project.

While research was being written, a few definitions of web scraping came up. Each of the three definitions introduced below notifies the extraction of information from different sources. As the underlying hotspots for the separated information, they differ.

It's necessary to collect data from sites designed for layman, not programmers, every now and then. "Internet scraping"[1] is the term for this method. The key term applies to knowledge sources that were designed with humans in mind. This description has been shown to be out of date. The creation of computerized techniques for extracting significant sources from programming[2] also provided an opportunity. Nonetheless, the distribution date of 2009 must be considered. API (Application Programming Interface) sources were limited at the time. Open API index accessible (Berlin, 2015) on Programmable Web Site had approx. Compared with 17175 recorded in 2017, 750 accessible sources are accessible[3].

Web scraping, also known as web extraction or reaping, is a method of extracting data from the World Wide Web (WWW) and storing it in a document framework or archive for later retrieval or examination[4]. When using the Hypertext Transfer Protocol (HTTP) or an internet browser, web information is often rejected. A client can do this manually or a bot or web crawler can do it automatically. Since the WWW offers a massive amount of heterogeneous information on a regular basis, web Scratching is generally recognized as an important and amazing method for gathering large data. [5]. The present situation is the continuous definition, where Web Scraping is referred to as one of the hotspots for large

information assortment, all the more unambiguously represented. Another term is also referenced in the definition.

Crawling the web is performed in a unique way, yielding unique results. The two exercises are depicted in Figure 1. The procedure steps on the left demonstrate that Web Crawling does not have a specific goal in mind and processes any available data without referring to specific data. The Web Scraper, on the other hand, receives, forms, and parses data from a predefined source in correlation. This text does not cover web crawling.[6]

The concept below does not refer to numerous subtleties. Nonetheless it briefly catches the Web Scratching exercises most accurately.

Purpose of Web Scraping

A Hypertext Mark-up Language (HTML) page's layout also contains massive measurements of source data from the World Wide Web. Computerized extraction is problematic since the intended peruse was a human. The motivation and explanation for data collection through Web scraping are presented in this chapter. The rapid growth of the World Wide Web has fundamentally altered how we provide, store, and distribute information. An enormous amount of data is stored on the web in both structured and unstructured systems. Concerns about knowledge scarcity and detachment at the same time, akin to beating the tangled masses of online information, have never been posed again in terms of questions or study topics. Since automated web scraping is available, these uses are only conceivable on a regular basis. It's impossible to collect the same amount of data over and over again in a reasonable amount of time without these procedures.

Analysis and research

One of the statistical survey strategies was to obtain information from online sources. It offers a much faster response, as opposed to studying an old style.

While it is known that it is best to use conventional reviews, Web-Scratching is considered to be a savvy support for such tools. Numerous sources should be used to get a far-reaching picture and to collect data on devices in business sectors.

In the online world, shoppers are dynamic and they offer their experience, disappointment, or inspiration. Online data wellsprings can be included by organizations that wish to gain more from shoppers. One of the techniques for collecting such data is web Scratching.

With refreshing indexes, directed e-shop data assortment and server promotion helps. Which depends every now and again on changing expenses. Computerized Web scratch processed files can provide interim updates to visit more and more.

With the increasing importance and accessibility of online costs that we see today, it is normal to ask whether the customer value list (CPI) forecast or related insights can be processed more easily every now and again than the current month-to-month plans take into account.

A small example of 338 Craigslist postings was used by Wegman and Chapple (2013) to consider the prevalence of optional dwelling units in the San Francisco Bay Area. Finally, to examine Seattle's lodging market, Feng (2014) web-scratched 6,000 Craigslist posts.

Social Mining and Sentiment Analysis

Another wellspring of data that is essentially unique in relation to ordinary ones is web based life. Internet-based life information indexes are largely produced by customers, and are huge, interlinked, and heterogeneous.

Web-based social networking data can be obtained from openly accessible sources using various techniques, such as Scratching, using apps provided by websites, and slipping. The chance of acquiring information about internet-based life makes it possible to examine information via web-based networking media. For example, in Twitter Data Analytics, four key advances are explained by a start-to-end process for Twitter information investigation: slipping, putting away, examining, and envisioning. Internet-based life data is much the same as regular data, which is a potential fortune find, but requires data mining to reveal hidden fortunes.

Automatic data collection

The evolving importance of web-based exchange requires a significantly greater web value range. Notwithstanding budgetary constraints, the additional remaining task at hand requires an increasingly competent organization of existing HR.

Naturally extracted information, such as physically collected information, on the ground that it considers a combination of the new technique into existing procedures for cleaning, altering and coordinating information.

LEGAL ASPECTS OF DATA EXTRACTION

2.1 LAW AREAS FOR WEB SCRAPER DEFENCE ACTION

A few site managers address the use of data gathered by Web Crawling. People or organizations that are dynamic have a basic legal resistance technique in Web Scraping. Sites are brimming with knowledge that has been disseminated away from the prying eyes of others.

For example, due to the site LinkedIn, the primary field-tested approach is to distribute profiles and decide a wage by charging a portion of the administrations provided to consumers. The target for customers is to open their CVs to an upcoming business. As a result, it would violate their tried-and-true policy of keeping sensitive information private, and any available information would be vulnerable to expansion. Web Scraping may violate a few conditions of laws or agreements. The table clarifies the legal zones in which a resistance activity should be set up for Web Scraper.

2.1.1 Website terms and conditions

Numerous organisations specifically reject Scratching under the terms and conditions of their site. Regardless of if they are already muddled but depending on criteria to authorize those terms, an argument for penetrating the arrangement is imaginable.

2.1.2 Copyright, Intellectual Property Rights

Because scraping entails duplicating, an argument for copyright invasion may be triggered. Regardless on if such a situation has any drawbacks, the particular provisions would be contingent on the fact that not all scratched content satisfies all copyright insurance standards.

Copyright-assured works include special scholarly and creative works, such as PC systems, web designs, and images.

2.1.3 Database rights

When any or a generous portion of a database is separated or re-used without the permission of the proprietor, a database right is breached. Likewise, the rehashed extraction or re-use of meagre pieces of a database that conflicts with the ordinary use of the database can invade database rights. When Scratching catalogs or posts from outside websites, violation of database privileges can also apply if the owner has expended costs in producing and caring for them.

2.1.4 Trademarks

Under the unlikely possibility that the scraper copies the (enrolled or unregistered) trademarks of a site owner without their permission, the site owner may make a motion ensuring invasion of the trademark as well as moving abroad. Going off prevents an outsider from marketing goods or carrying on business under a name, label, representation, or in any other manner that is likely to misdirect, beguile or mislead individuals in general into accepting that getting a position with the brand owner is the product or organization.

2.1.5 Data protection

Organizations wishing to use mechanized Scratching systems to collect data about individuals should realize that if they collect "individual information" they can violate the nearby information insurance rule (that is, any data that recognizes a living person). The focal problem is whether individuals have committed to the processing of their own information. Despite the fact that information obtained from one location may not be similar to home information in seclusion, as it is gathered from several sites, without the permission of the person involved, a corporation may inadvertently end up holding individual information. Using such close-to-home data would break laws on privacy assurance.

2.1.6 Criminal Damage

It is a crime to do criminal damage to a Computer (counting damage to data) or to use a PC to get information without authorization. In the same way, Scratching details may be a criminal

offence since access to the database has not been authorised by the site owner.

2.2 General Legal Aspects

There is no specific bit of enactment in the outline that disables Web Scraping to accumulate knowledge. Nonetheless, under covered invention legislation and partnership law, the site owners could have legitimate rights against the company. However each case will transform on its own realities and there is a great deal of ward on what material is scratched off the sites. Organizations should be wary of authoritative agreements they have consented to with respect to the terms of use of a site, which may prohibit the user from getting the details off the site and using it. The only way to be very sure that a site owner's rights have not been violated is to get their express consent to scrape the screen and then use the info. The general outcome of unauthorized web scraping could be detrimental for the owner of the website. The likelihood of fewer visitors, fewer connections from news search sites, and lower ad revenue. As a result, where the scrubber poses a threat to the information host's centre business and the information host has a sufficient argument to prevail legitimately against the scrubber, data hosts should consider using legal activities against the scrubber. It is necessary to change the terms of use on the pages from a legal standpoint.

2.3 Conclusion

Different Web Scraping techniques can be used, calculated by details, periodicity and the appropriate outcome can be taken into consideration depending on the main function. Web scrubbers have an extensive range of tools to choose from. The venture should not only consist of specialized plans and executions. The valid part of the specific job examined and important advancements separated by all Web Scraping activities should be the legal part of the specific job. Data hosts can regularly survey the benefit scrubbers will offer individuals who scratch their data and follow a sober-minded approach. Site Scrapers can retain the Data Association and allow the Data Host ID as a wellspring of implemented data.

CHAPTER-3

METHODS OF WEB SCRAPING & AVAILABLE SOFTWARE TOOLS

3.1 Methods of Web Scraping

Along with the World Wide Web, the methods for Web Crawling have evolved. Not every single technique documented was accessible at the beginning. There are two books to refer to on the basis that they are by and by the most used methods. The Document Object Model (DOM) has been increasingly prevalent in DHTML since 2000. A more detailed recognition later allowed the HTML Parsing method to evolve to DOM Parsing. Application Programming Interfaces are the second model (APIs). The production of usable material APIs is dated from 2005 and this technique is the most youthful in the overview. The quantity of APIs created within 8 years from 0 to 10302 requires agreeing to ProgrammableWeb.com.

3.1.1 Manual Scraping

In specific circumstances, manual scraping is still an option. When the amount of information to be scratched is insignificant, when the information to be scrapped does not require a redundant assignment, and when setting up robotized Scratching will take longer than the information collection itself. Mechanized techniques may be prohibited due to site protection measures or explicit site characteristics.

3.1.2 Parsing HTML

In general, sites do not allow them to confine appropriate organizations, such as .csv or .json archives. The server renders HTML pages as a response to the demand of a client. Server programming is not important now but the program's yield is large. An analysis of the HTML layout (the simple page test given in Figure 3.1) would reveal rehashed components on the given page. The pages having comparative example can be used as a data hotspot for programming language content or Web Scraping.

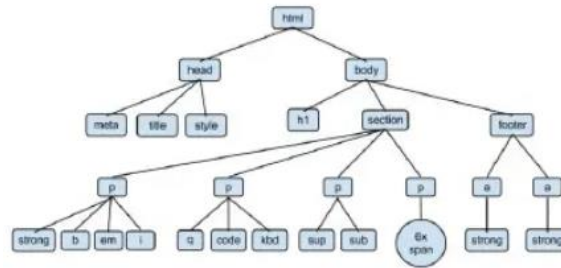


Figure 3.1 HTML Parsing (source: Google images)

3.1.3 Parsing DOM

Document Object Model (DOM) Parsing is a method in Parsing of HTML focused on language and program enhancements that add to the Document Object Model presentation. For Cascading Stylesheets (CSS) and JavaScript, DOM is used vigorously. Additional possibilities for tending to certain unique sections of the website page were discovered by the incorporation of DOM. Compartments of their own DOM addresses are shown in figure 3.2. These are used for smoother pathways across website material in Web Scraping.

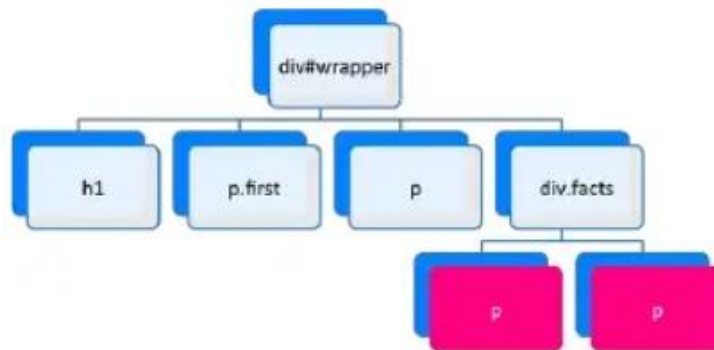


Figure 3.2 Dom Parsing (source: google images)

3.1.4 XPath

Comparing propensity to possibility as XPath offers DOM (XML Path Language). For XML files, the name indicates a usage. It is also applicable to HTML architecture. Way needs a site page that is more precisely structured than DOM and has a similar probability of addressing

fragments inside the page of the website. The report structure as deciphered in the XPath as seen in Figure 3.3.

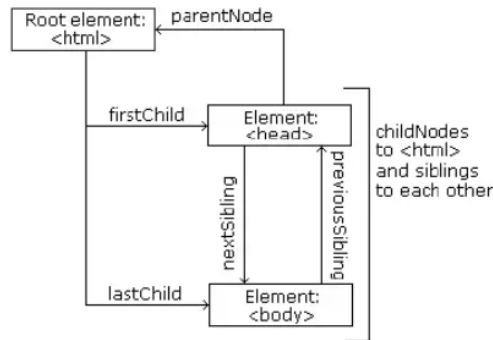


Figure 3.3 XPath (Source:Google images)

3.1.5 API's

Application Programming Interface (API) expects an application as a correspondence accomplice, while the previous methods work to scratch intelligible yields. APIs are often called as machine-discernible interfaces in this way (versus intelligible). Indeed a lot later than the WWW, also APIs were presented, and their creation was fast. It splits the world with APIs. API Directories were made for a fundamental analysis and direction. The vast majority of available APIs are reported and represented in the index with substantial source ties. Programmable Web (<https://www.programmableweb.com>) and APIs (<https://apis.guru/>) are two examples of such indexes. API repositories also have their own API, which enables clients to check for API sources throughout their directory. A typical HTTP request sent to the endpoint of an API restores a server response. Each Apus has its own commitment and choices. In the solicitation, the required answer structure may be set as an alternative. JSON is the most commonly used format for API communication.

3.2 Available Software Tools

This segment will address a few distributed Site Scraping programming units. We should agree that there are more devices that have been manufactured by individuals and organizations for internal purposes that have not been circulated along these lines. Along these lines, because of the multiplication of uses and steps, this rundown eliminates all usable programming. Areas that depend on the form of programming are formed according to:

- Cloud based Software
- Desktop Software
- Programming libraries in both python and R
- Browser Extensions available for google chrome

3.2.1 Cloud Software

Cloud arrangements, although the framework backend remains on the cloud platform, offer a user experience with a Web program. Such a specification limits the specifications of consumer end equipment to its foundations. Without getting your own PC on, large volumes of data can be scratched. Without additional equipment or extended web transmitting power, larger activities can be performed. In some cases, where the cloud is located is relevant. For ex, certain website pages with their full spectrum may not be accessible from Europe. A cloud agreement connected to a US-based ISP can have the option of going to various parts of such a platform.

3.2.1 Desktop Software

Site content is locally copied, parsed and stored. Work Field Systems include a web alliance with broadband. The PC's network access would legitimately affect the Web Scraping task planning period. Further built equipment is required in contrast to Cloud Computing for Desktop Software.

3.2.1 Programming libraries

Information on the webpage is downloaded, parsed and saved locally. Job territory Applications include an association with the broadband web. The PC's web association can really impact the Web Scraping task's getting ready period. In comparison to Cloud Computing for Desktop Software, more equipment produced appears differently. At least a Random-Access Memory (RAM) workstation with a capacity greater than 8 GB is suggested.

3.2.1.1 Scrapy

Scrapy is one of the network scraping programs accessible by propulsion. In Python, the Framework is written. It is an application system that offers various orders for Web Scraping projects to make their own software and use them.

Scrapy is a slippery network framework for programmers to write bug-making code that

characterizes how a single website (or destination collection) would be refused. The main factor is that it is built on Twisted, an offbeat device management library, so Scrapy is executed using a simultaneous non-blocking (otherwise known as non-concurrent) language which makes the execution of insects amazing.

To assist engineers who wish to coordinate it into their own programs, broad documentation is available. It is not a limitation to distinguish between a local desktop program or a cloud solution, both can be fueled by Scrapy javascript. One of the Cloud Technologies launched encourages the customer to transfer their Scrapy software to the cloud foundation.

3.2.1 Web Scraper

Web scraper is an optimization software available for Google Chrome that can be used for web scraping. The sitemap explains how a platform must be traversed and what data should be separated. At the same time, it can scratch different pages and even has dynamic capabilities for extracting information. Web scrapers will work with both JavaScript and Ajax pages as well which makes it more amazing. The instrument deletes information from a CSV or CouchDB record. The designer provides a cloud rendition for broader Web scraping occupations. The Google Chrome extensionist is often used as a research domain until the cloud version is reduced to scraping occupations.

3.3 CONCLUSION

The documented devices can hardly be compared with each other. Some are meant for professionals and others are suitable for companies. Expansions in the curriculum include a place in the expert class. These instruments are all that might probably be expected to gather data from a few places for a quick analysis or venture. Job Field Systems have a comprehensive collection of features. That is reflected in the considerable cost. A comparative power system is provided by cloud agreements. In comparison to the Desktop Computing Cloud Approach, there are 3 basic desirable conditions:

- Monthly billing based on asset use rather than an initial high-risk, low-equipment requirement on your own Server/PC.
- Own data transmission is used solely for job scheduling and cloud management. Software is currently the most accessible way to begin and test web scraping alternatives.
- Systems are the most perplexing option. For system engineers, the terminology used and the accessibility of documentation are critical.

CHAPTER-4

HANDS ON WEB SCRAPING & IMPLEMENTATION

The described topic of Web Scraping might not be clear from the description only. Therefore, this chapter offers a guide through a practical presentation.

4.1 Task

For the demonstration purposes, Top-Rated Workplace: Tech in India the best technology companies to work for in India in 2019, based on employer ratings and reviews on Indeed.

Target location. Information related is shown below in the figure 4.1.

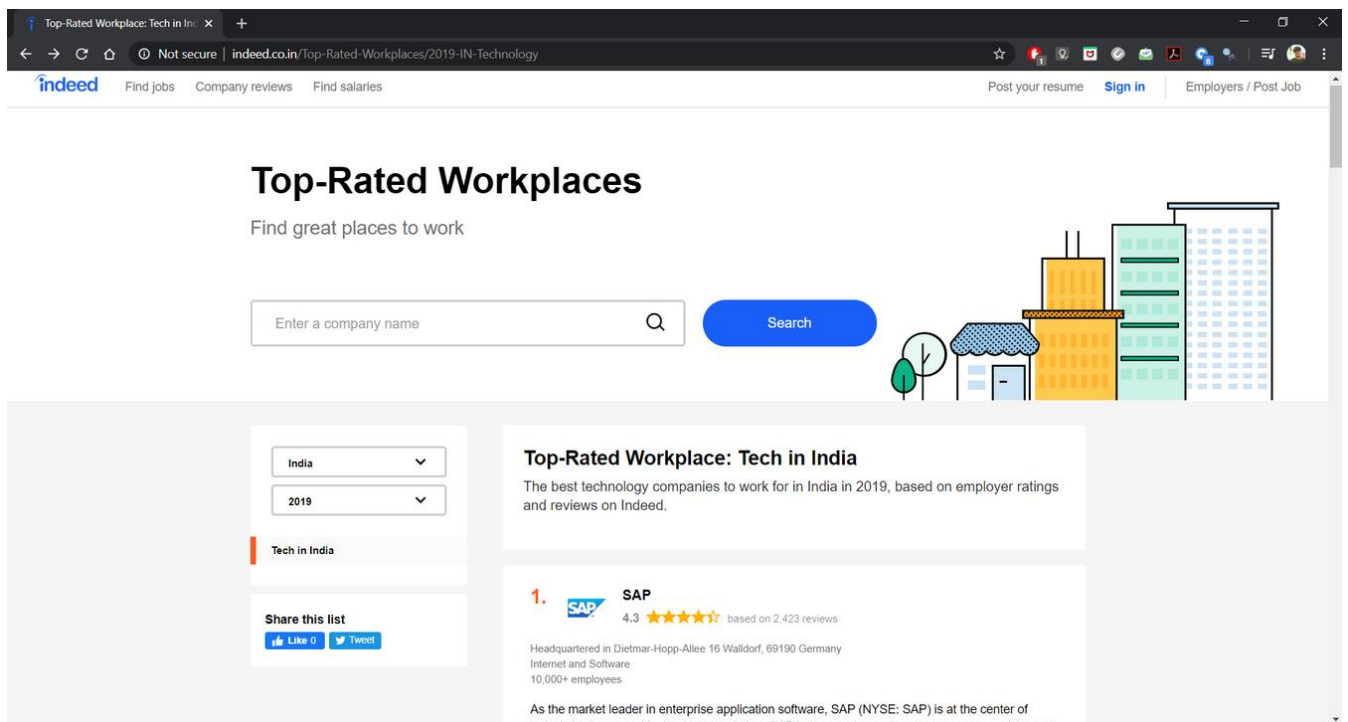


Figure 4.1 Target website (Source : Indeed.com)

DATA RESTRUCTURING FORMAT

Data restructuring is shown in the figure 4.2.

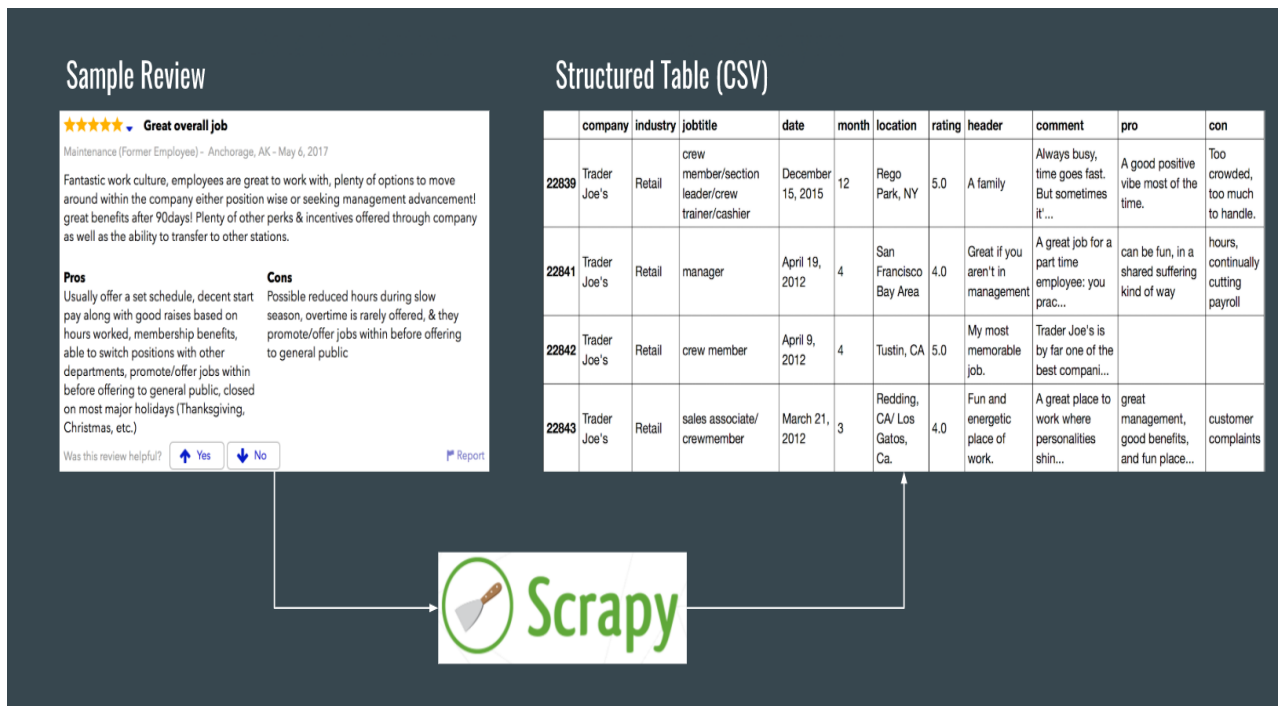


Figure 4.2 Data restructuring format (Source : Towards Datascience)

4.2 Prerequisites

```
[4] !pip install scrapy
```

Figure 4.3.1

Installing and Importing Important libraries

```
[5] import scrapy
```

Scrapy is a free and open-source web-crawling framework written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler. It is currently maintained by Scrapinghub Ltd., a web-scraping development and services company.

Figure 4.3.2

CODE Importing scrapy

```
[6] !scrapy startproject indeed

[ ] New Scrapy project 'indeed', using template directory '/usr/local/lib/python3.6/dist-packages/scrapy/templates/project',
/content/indeed

You can start your first spider with:
cd indeed
scrapy genspider example example.com
```

Figure 4.3.3
CODE starting scrapy project

4.4 Preparation

```
[18] !scrapy genspider indeed_review https://www.indeed.co.in/Top-Rated-Workplaces

[ ] Created spider 'indeed_review' using template 'basic' in module:
indeed.spiders.indeed_review
```

Figure 4.3.4
CODE creating scrapy spider

```
indeed_review.py X
# -*- coding: utf-8 -*-
import scrapy

class IndeedReviewSpider(scrapy.Spider):
    name = 'indeed_review'
    allowed_domains = ['https://www.indeed.co.in/Top-Rated-Workplaces']
    start_urls = ['http://https://www.indeed.co.in/Top-Rated-Workplaces/']

    def parse(self, response):
        pass
```

Figure 4.3.5
CODE parsing responses

4.5 Execution

```
indeed_review.py X
# -*- coding: utf-8 -*-
import scrapy
from ..items import IndeedItem

class IndeedReviewSpider(scrapy.Spider):
    name = 'indeed_review'

    start_urls = ['https://www.indeed.co.in/Top-Rated-Workplaces/2019-IN-Technology']

    def parse(self, response):
        items = IndeedItem()

        cmp_name = response.css('div.cmp-company-tile-name > a ::attr(title) ').extract()
        cmp_review_count = response.css('div.cmp-tile-footer-element:nth-child(2) > a::text').extract()
        cmp_review_hyperlink = response.css('div.cmp-tile-footer-element:nth-child(2) > a ::attr(href)').extract()

        items['cmp_name'] = cmp_name
        items['cmp_review_count'] = cmp_review_count
        items['cmp_review_hyperlink'] = cmp_review_hyperlink

        yield items
```

Figure 4.3.6
CODE spider indeed review

```
items.py X
# -*- coding: utf-8 -*-

# Define here the models for your scraped items
#
# See documentation in:
# https://docs.scrapy.org/en/latest/topics/items.html

import scrapy

class IndeedItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    cmp_name = scrapy.Field()
    cmp_review_count = scrapy.Field()
    cmp_review_hyperlink = scrapy.Field()

    pass
```

Figure 4.3.7 CODE items.py

4.6 Results

Table 4.1: TABLE Company_profile.py output

Rank	Company	Industry	Overall
1	Adobe	Internet and Software	4.3
2	Facebook	Internet and Software	4.2
4	Live Nation	Media, News and Publishing	4.1
7	Delta	NA	3.9
8	eBay Inc.	Internet and Software	3.9
9	Microsoft	Internet and Software	4.2
11	Bristol-Myers Squibb	Pharmaceuticals	4.2
12	Salesforce	Internet and Software	4.2
13	Fannie Mae	Banks and Financial Services	4
14	Eli Lilly	Pharmaceuticals	4.2
15	JetBlue Airways Corporation	NA	4.1
16	Freeport-McMoRan	Agriculture and Extraction	4.1
17	Fluor Corp.	Construction	4.1
18	Apple	Computers and Electronics	4.2
19	Cisco	Internet and Software	4.1
20	Capital One	Banks and Financial Services	4
22	Amgen	Health Care	4.1
23	Booz Allen Hamilton	Consulting and Business Services	3.9
24	Charles Schwab	Banks and Financial Services	4
25	Viacom	Media, News and Publishing	4

26	Southern Company	Energy and Utilities	4
27	NextEra Energy	Energy and Utilities	4
28	NA	NA	4.1
29	Land O'Lakes, Inc.	Agriculture and Extraction	3.7
30	Motorola Solutions	Telecommunications	4.1
31	Pfizer Inc.	Health Care	4.2
32	Lockheed Martin	Aerospace and Defense	4
34	Merck	Health Care	4.1
35	ConocoPhillips	Agriculture and Extraction	4.1
36	American Express	Banks and Financial Services	4.1
37	Applied Materials	Computers and Electronics	3.9
38	DTE Energy	Energy and Utilities	4
40	Boston Scientific	Health Care	4
42	Discover Financial Services	Banks and Financial Services	3.9
43	BlackRock Inc.	Banks and Financial Services	3.8
44	Darden Restaurants	NA	3.9
45	MGM Resorts International	NA	3.9
46	Hilton	Restaurants, Travel and Leisure	4
47	Edward Jones	Banks and Financial Services	3.8

4.7 Conclusion

The viable model has appeared, how to utilize changed site pages as information hotspots for own dataset. The reason for representative audit assortment expects an occasionally updates of the web Scratching task. Hence. It would require a manual activity to keep the information source refreshed. Appropriate arrangement would be a cloud application for such case. In the model the sites terms of utilization were not analysed. This would be essential if such errand would be performed for a genuine reason. In the following section we will pre-process this information and do the cleaning and change.

CHAPTER-5

DATA PRE-PROCESSING

5.1 Information into data

5.1.1 Data Cleaning

Identify and handle missing values

Identify missing values

Convert "?" to NaN

```
[8]: import numpy as np

[9]: # replace "?" to NaN
     df.replace("?", np.nan, inplace = True)

[10]: df.head(5)
```

Figure 5.1

CODE Convert "?" to NaN

Evaluating for Missing Data

The missing values are converted to Python's default. We use Python's built-in functions to identify these missing values. There are two methods to detect missing data:

.isnull()

.notnull()

The output is a boolean value indicating whether the value that is passed into the argument is in fact missing data

```
[11]: missing_data = df.isnull()
      missing_data.head(5)
```

Figure 5.2

CODE Evaluating for Missing Data

Count missing values in each column

Using a for loop in Python, we can quickly figure out the number of missing values in each column. As mentioned above, "True" represents a missing value, "False" means the value is

present in the dataset. In the body of the for loop the method ".value_counts()" counts the number of "True" values.

```
] : for column in missing_data.columns.values.tolist():  
    print(column)  
    print (missing_data[column].value_counts())  
    print("")
```

Figure 5.3

Deal with missing data

drop data

- a. drop the whole row
- b. drop the whole column

replace data

- a. replace it by mean
- b. replace it by frequency
- c. replace it based on other functions

Entire sections ought to be dropped just if most passages in the segment are unfilled. In our dataset, none of the sections are sufficiently unfilled to drop totally. We have some opportunity in picking which technique to supplant information; in any case, a few strategies may appear to be more sensible than others. We will apply every technique to a wide range of sections.

Finally check if the data is in the correct format (int, float, text or other).

As the Pandas library is being used, therefore following will be used to perform actions on data types

- .dtype() to check the data type
- .astype() to change the data type

```
] : df.dtypes
```

Figure 5.4

Convert data types to proper format

Syntax: `df[["", ""]] = df[["", ""]].astype("")`

Data Transformation

Information is typically gathered from various organizations with various configurations. (Information Standardization is likewise a term for an information standardization, where we deduct the mean and partition by the standard deviation)

Information standardization

Standardization is the way to change the estimations of some factors into a range that can be compared easily. Commonplace normalizations incorporate scaling the variable so the variable normal is 0, scaling the variable so the change is 1, or scaling variable so the variable qualities go from 0 to 1

Binning

For collected inquiry, Binning is a method of changing recurring numerical variables into discrete unmitigated 'containers.'

For clustered analysis, binning is a method of converting continuous numerical variables into discrete categorical 'bins.'

5.1.3 Data Reduction

The data reduction method may result in a simplified description of the original data that is much smaller in size but retains the original data's quality

CHAPTER-6

UNDERSTANDING EMPLOYEE REVIEWS USING SENTIMENT ANALYSIS

6.1 Topic Modelling

To visualise the feelings through representative surveys, an estimation analysis was completed using R's Syuzhet library. Each diagram in the delineation below represents the intensity of a feeling from January to December on a positive scale.

By and large, the feelings observed correspond to our perception that the majority of worker surveys are reliable, as we are primarily focused on analysing representative audits among the top 8 corporate retail bosses, as indicated by Indeed's positioning. On a side note, Build-A-Bear Workshop and Trader Joe's had more unpredictability in their negative feelings scoring (for example negative, bitterness, dread, nauseate, outrage).

6.1.1 Identify Patterns Among Positive and Negative Employee Reviews

Point displaying is a form of solo AI that can help us differentiate designs by grouping similar objects into groups. pyLDavis, a Python library, was used to break down the groups of positive and negative audits in this study (utilizing upsides and downsides of the representative surveys). Essentially, the library recognises a list of archives as data sources (for example, a list of audits) and attempts to assemble the surveys based on common combinations of catchphrases appearing in each of these surveys.

The downside of topic demonstrating, as with other bunching methods, is that the client must decide the number of groups into which the documents should be grouped. The underlying aim is to comprehend the number of themes or groups within your sources of information, which poses a difficult situation problem. Given this challenge, the client must use their business judgement, as well as run several preliminary analyses with varying numbers of subjects as contributions, to come up with findings that are generally interpretable.

6.2 Observed Topics

Expert audits and con surveys were used to guide two theme demonstrating examinations. The following are the most well-known word combinations that appear in each of the gatherings for each review. It's not surprising that the gatherings cover the common terms recognised among them, given that we've chosen a reasonably high number of points. In any case, the findings provide a useful starting point for figuring out what employees are thinking.

6.2.1 Observed Topics Among Pros and Cons in Employee Reviews

TABLE 6.1

Pros	Cons
free, lunch, holiday, store, bonus	break, short, work, lunch, weekend
flexible, hour, schedule, time, break	management, work, poor, business, lack
discount, employee, food, care, break	hour, work, long, stress, weak
membership, job, bonus, healthcare, wage	season, hard, pay, job, benefit
benefit, pay, great, good, worker	time, bad, position, management, season
health, benefit, people, day, meet	customer, shift, lot, low, time
good, benefit, company, fast, pace	schedule, change, work, worker, day
work, great, fun, environment, nice	rude, management, promotion, health, night
pay, good, friendly, work, customer	employee, time, supervisor, discount, college

6.3 Visualizing

The graphical representation of information and data is known as data visualisation. Data visualisation tools make it easy to see and understand trends, outliers, and patterns in data by using visual elements including charts, graphs, and maps.

CHAPTER-7

CONCLUSION & FUTURE SCOPE

7.1 Conclusion

We used ethical web scraping to study and collect data, which we then analysed to gain a deeper understanding of the company and their working environment.

To do so, we built a web scraper that automates the process of extracting data and information from Indeed.com, as well as analysing and visualising the extracted text.

The following are the outcomes of this project:

- A working knowledge of Internet data, web scraping techniques, web scrapers, and data processing.
- A web scraper to hack through and retrieve data from Indeed.com reviews.
- Keeping track of the results in a database so they can be manipulated later.
- Conducting studies on the collected data and visualising the results

The following are the project's deliverables:

- Web scraper
- Database
Indeed_cmp_profile.csv, Indeer_cmp_review.csv
- Analysis and Visualized outputs

7.2 Future Scope

Examine the reputation of Indeed's Top 50 Corporate Employers by comparing them to similar companies that did not make the list.

Understand why positive and negative reviews are concentrated in the months in which they were observed, and determine if this behaviour is common in other industries.

To test the relationships between variables, use statistical methods.

REFERENCES

1. Mohammad A. Hassanain, Analysis of factors influencing office workplace planning and design in corporate facilities. Journal of Building Appraisals, pp. 183-197
<https://link.springer.com/article/10.1057/jba.2010.22>
2. Jacquelin C. Vischer, Gustave Nicolas-Fischer, pp. 73-96
<https://www.cairn.info/journal-le-travail-humain-2005-1-page-73.htm>
3. A.A. Saleh, An Overview of the Influence of Physical Office Environments Towards Employee. Procedia Engineering, vol. 20, pp. 262-268
<https://www.sciencedirect.com/science/article/pii/S1877705811029730>
4. Deepak Bangwal, Prakash Tiwari Workplace environment, employee satisfaction and intent to stay. International Journal of Contemporary Hospitality Management, pp 268-284
<https://www.emerald.com/insight/content/doi/10.1108/IJCHM-04-2017-0230/full/html>
5. Jacquelin C.Vischer, Towards an Environmental Psychology of Workspace: How People are Affected by Environments for Work., Architectural Science Review, vol. 51, pp 97-108
<https://www.tandfonline.com/doi/abs/10.3763/asre.2008.5114>
6. Ying Hua, International Journal of Facility Management, vol.1, no.2
https://community.ifma.org/cfs-file/_key/telligent-evolution-components-attachments/13-465-00-00-01-05-76-92/2010_5F00_A-Model-of-Workplace-Environment-Satisfaction-Collaboration-Experience_5F00_Article.pdf
7. EEE521 Evan Gallagher B00642761 BSc Hons Computer Science Scraping Websites for Law Enforcement.
8. medium.com/@msalmon00/web-scraping-job-postings-from-indeed
9. John J. Salerno, D. M. B., 2003. Method and apparatus for improved web scraping. United States of America, Patentor. US 7072890 B2.
10. www.towardsdatascience.com
11. Stack overflow community
12. Geeksforgeeks community
13. <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>
14. Ryan Mitchell – Web Scrapping Using Python, First Edition, Orilley, June 2015
15. <https://www.quora.com/What-is-the-legality-of-web-scraping>
16. Google images