Degree Project

Level: Second-Cycle

Creating Financial Database for Education and Research: Using WEB SCRAPING Technique

Author: Lanny Anthony Rodrigues & Srujan Kumar Polepally

Supervisors: Kenneth Carling & Asif M Huq

Examiner: Siril Yella

Subject/ Main Field of Study: Microdata Analysis

Course Code: MI4002

Credits: 15 ECTS

Date of Examination: 20th January 7, 2021

Dalarna University – SE-791 88 Falun – Phone +4623-77 80 00

TABLE OF CONTENTS

HÖGSKOLAN
DALARNA

PROJECT TITLE

Creating Financial Database for Education and Research: Using WEB SCRAPING Technique

ABSTRACT: Our objective of this thesis is to expand the microdata database of publicly available corporate information of the university by web scraping mechanism. The tool for this thesis is a web scraper that can access and concentrate information from websites utilizing a web application as an interface for client connection. In our comprehensive work we have demonstrated that the GRI text files approximately consist of 7227 companies; from the total number of companies the data is filtered with "listed" companies. Among the filtered 2252 companies some do not have income statements data. Hence, we have finally collected data of 2112 companies with 36 different sectors and 13 different countries in this thesis. The publicly available information of income statements between 2016 to 2020 have been collected by GRI of microdata department. Collecting such data from any proprietary database by web scraping may cost more than $ 24000 a year were collecting the same from the public database may cost almost nil, which we will discuss further in our thesis.

In our work we are motivated to collect the financial data from the annual financial statement or financial report of the business concerns which can be used for the purpose to measure and investigate the trading costs and changes of securities, common assets, futures, cryptocurrencies, and so forth. Stock exchange, official statements and different business-related news are additionally sources of financial data that individuals will scrape. We are helping those petty investors and students who require financial statements from numerous companies for several years to verify the condition of the economy and finance concerning whether to capitalise or not, which is not possible in a conventional way; hence they use the web scraping mechanism to extract financial statements from diverse websites and make the investment decisions on further research and analysis.

Here in this thesis work, we have indicated the outcome of the web scraping is to keep the extracted data in a database. The gathered data of the resulted database can be implemented for the required goal of further research, education, and other purposes with the further use of the web scraping technique.

Keywords: Web Scraping, GRI Text Files, Proprietary Database, Public Database, Financial Report, Financial Statement, Financial Data, Extracted Data.

I. INTRODUCTION

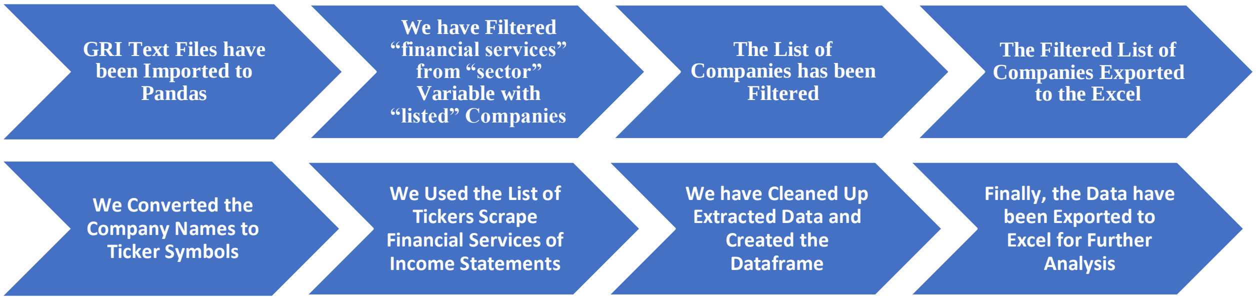
Extracting data from a website to utilize it in an alternate setting could be characterized as web scraping. It is otherwise called web harvesting or web data extracting. The strategy is worked out in collection of data from the annual financial report of a set of individuals, group of professionals as well as number of concerns or conglomerates by applying as sequence of instructions, net income adjustment method and also cost modification procedure against the assembled data. By utilizing web scraping strategies client can remove data accessible on various website into a solitary database as well as in a spreadsheet (Mitchell, R, 2018).

Since we have expressed in the starting point web scraping is the cycle to extricate data from a data to use it in an alternate setting, the question may arise, “What are we collecting through this unique technique and why are we collecting it?” In this thesis, we set out to collect and organize all financial data for all firms around the world to create an internal microdata database of publicly available corporate information. Briefly implying, the execution our web scraping technique is motivated by both educational and research needs at the educational and research institutions in order to use them for both educational and research purposes i.e., master’s thesis, bachelor thesis, dissertations, other course work or projects, training and execution of further research and analysis purposes. Collection of data by using web scraping rather than using the conventional data collection method is that it is financially highly expensive to extract data from the databases such as Bloomberg, Asset4, Kitco, SEC, XE etc. Although we need to refer to the number of positive characteristics related to complete data when scraped as opposed to subscribing it from Bloomberg which is quite a massively broad coverage, and eventually the data is perhaps more dependable than Yahoo or Google, which also do fail to perform precisely all the time. The stock market data scraping with the financial information, working as a business directory, by providing with the contact information of the financial entities makes Bloomberg extremely superior in the field of data providers. The efficiency and customized web scraper of Bloomberg gives the clients proper financial information to take robust decisive actions for their companies. More of all Bloomberg offers the preprocessed database which has the advantage to use as a finished product whereas the data collected from public database like Yahoo Finance has to be processed and then utilized. Still the grieving factor lies with the extremely high maintenance cost which cannot be ignored by

Bloomberg. The expenses can be faltering for non-profit educational organizations and government offices, and furthermore for independent companies. Bloomberg is the costliest among the financial data providers as they are charging $24,000 a year for a single terminal subscription. (Bloomberg, Michael R. August 27, 2001). On the other hand, Web scraping from the Yahoo Finance, Google Finance etc. offer a basic assistance at very low cost which is almost one fifth of the cost of Bloomberg; sometimes almost zero. The information has been collected from websites and routinely examined with the goal of web capacities. Web scraping technique figure out how to do this in an economic and budget-friendly way. It also has a low maintenance cost in comparison to the other conventional method applied on the proprietary databases. In that manner, it helps the budget plan to act precisely (Patel Jay M, 2020).

Our general contribution is that we have demonstrated how an inexpensive database of publicly available corporate information can be created by collecting data from “high end” websites such as Yahoo Finance. We also Highlight some of the challenges’ users might face for such kind of data collection and explain how those can be overcome. Our thesis work contributes to collect publicly available information from the global reporting initiative data sources. Collecting GRI's companies’ financial data and creates a micro database. Financial data provides with accessibility to the repository of information on the measures that financial services companies are undertaking in order to make their user services accessible and friendly to users of all potentials. The financial database is open data and content can be freely used, modified, and shared by anyone and for any purpose. Open data is the idea that some data should be freely available to everyone to use as they wish, without restrictions from any type of control. In this thesis we are collecting financial data by using web scraping technique through pre-processing, company name to ticker symbol, data extraction of web content, data cleansing and building excel database. The challenges of the process of data collection are inconsistent data collection, complexity of data, Fragmentation of Downloading Method from those Sources, and data quality. Overcome and rectifying these challenges by the script algorithms with web scraping technique and outcomes of data consist a quality of data. These high complexity, and huge amounts of data stores in excel with different formats for further implementation. The management of big financial data refers to the petabytes of structured data that can be used to anticipate user behaviour and create strategies for academic and financial institutions.

Now, how did we collect the financial data or through which process did we collect them? In order to achieve the goal to obtain the publicly available financial data, we implemented the web scraping tools. We have followed some steps to collect the financial data. The collection process of our financial data can be presented through a flowchart:


GRI Text Files have been Imported to Pandas
We have Filtered "financial services" from "sector" Variable with "listed" Companies
The List of Companies has been Filtered
The Filtered List of Companies Exported to the Excel
We Converted the Company Names to Ticker Symbols
We Used the List of Tickers Scrape Financial Services of Income Statements
We have Cleaned Up Extracted Data and Created the Dataframe
Finally, the Data have been Exported to Excel for Further Analysis

In our thesis work we aimed to create an internal microdata database of publicly available corporate information which we have already mentioned in the previous section. The database comprises of the vital data of the companies around the world which can be the historical financial and non-financial information of the companies. The financial information can be historical balance sheet data, stock market data, analyst forecast data (company ratings & text), stakeholders review (ratings & text), etc. The data will be used for research and educational purposes at the academics.

At the end of the discussion, we can briefly imply that, we are collecting data by automated process and creating dataset by using web scraping technique. The thesis work basically centers around web scraping method that can be utilized to get enormous amounts of financial information continuously. Automatization of fiscal data extraction framework is one of the fundamental objectives pursued in this work, quickly followed by the improvement of a system for perceptive modelling, applying neural networks and thoughtful learning techniques to the data achieved through web scraping.

II. RATIONALITY OF WEB SCRAPING

Web Scraping is characterized as utilizing technology tools for programmed extraction and association of data from the Web with the objective of additional analysis of this information

(Krotov and Tennyson 2018). Web Scraping comprises of the accompanying fundamental, entwined stages: website investigation, website creeping, and information association.

Since the website analysis requires analyzing the fundamental structure of a website or a web repository (for example an online database) to see how the required information is archived we have done exactly according to the requirement in our work by providing with the fundamental comprehension of the World Wide Web engineering and languages (for example HTML, CSS, LXML, and so forth). In this thesis work we have included creating and running the content that automatically browses the website and recovers the required data. We have used Python as programing language to extract the financial data from the websites mainly provides with the financial statements of financial and non-financial concerns.

To the best of our knowledge similar work that match with our topic is almost nonexistent. It is nonexistent as web scraping as our method have been developed recently. hence, we have taken a conceptual and analytical approach to fill the gap of the research. Here we have discussed mainly about the financial statements, challenges, and process to collect the data through Web Scraping technique and how did it fulfil our objective of the thesis work.

Although as mentioned above the past works by the researchers have a very little relevance with our work, we have gained enough knowledge and courage to pursue our thesis work from the books, blogs, journals, research papers etc by a number of authors as well as data researchers. Their works were not limited to the authenticity of the data published; importance of the data scraped for as well as data governance helped us a lot to create a significant impact of our objective. We are trying to focus a little on those works before we discuss about the rationality of our thesis.

The authenticity of the published data likewise sees the exactness and uprightness in comparable lights independently, during research and analysis. The authenticity of published information is checked appropriately. Once the above variables have been ensured the cycle of analysis can be effectively done. With regards to statistical surveying, time is a vital factor. With time the pattern changes quickly. So, one does this statistical surveying is to comprehend the current trend. The strategy planning hampers with the less authenticity of the collected data. Hence, we can say that the market research depends a lot on the authenticity of the data. (McCandless, David. 2012)

There is massive importance of the user case for which data is being scraped for. Definitely it took a lot for web scraping to provide the organizations with quite useful data. Mainly the business institutions have benefited from this meticulous mechanism. In order to change in price, setting up competitive price, inauguration of a new item in the market as well as running promotional activities, the companies use web scraping to send information to the customers. The banking industry also gets the benefit by upgrade the information of stock exchange rates as well as the interest rates automatically in customer's catalog. Collection of accurate and needed data from the websites, retrieval of the static and dynamic pages and very importantly, the data automation factor aids the user quite a lot which is possible by the web scraping technique. *(*Patel Jay M. 2020*)*

The complete management of data accessibility, significance, serviceability, reliability, and protection in a company is known as the data governance. It helps companies to control their data experience. It guarantees the data channel is prepared to aid the company catalog, safeguard and control the confidential data, locate data extraction, share data, regulate and overall control data to be protected against mismanagement. (Dryer, A.J., and Stockton, J. 2013). So, when we discuss about the impact of web scraping and the research objectives, we must also keep the traits of data governance in our mind.

In this thesis we have concentrated on the problem of compilation of the public data which can be quite conveniently collect through the modern and sophisticated process of web scraping rather than using the manual method which might be time consuming and monotonous. Here we will emphasize on scraping the data from the financial statement of the different institutions. But before that we must also understand what a Financial Statement is and what are the contents in it.

Financial statements are formal records of the financial activities of a business, person, or other entity. Financial statements provide an overview of a business or person's financial condition in both short and long term. All the relevant financial information of a business enterprise presented in a structured manner and in a form easy to understand is called the financial statement. Income statement is also referred to as Profit and Loss statement (or "P&L"), reports on a company's income, expenses, and profits over a period of time. Profit & Loss account provide information on the operation of the enterprise. These include sale and the various expenses incurred during the processing state (Ittelson, Thomas 2009). The public data can be amassed either manually or by web scraping which is collected through extraction. The collection of the historical financial and non- financial information of the concern becomes quite challenging. The prevailing challenges to collect the data are:

- Massive Quantity
- Acquiring Trustworthy Source(s)
- Fragmentation of Downloading Method from those Sources.
- Indexing and Merging Data from Multiple Sources
- Unstructured Data
- Reliability of the Data

This is the reason why in this thesis we have we have encountered and overcame these challenges to collect the publicly available data of financial and non-financial institutions' financial statements and the financial data through the highly unique, incredibly less expensive, structured, and systematic as well as trustworthy technique of web scraping.

The information accessible on the Web involves, organized, semi-organized, and unorganized quantitative and subjective data distributed as web pages, HTML tables, web information bases, emails, tweets, blog entries, photographs, recordings, and so on (Watson 2014). Controlling Web data requires tending to various specialized technical issues identified with volume, diversity, speed, and accuracy of data on the Web (Goes 2014).

In the first place, the information on the Web is routinely depicted by enormous volume assessed in Zettabytes (billions of gigabytes) (Cisco Systems 2016). Secondly, these huge data archives accessible on the Web, arrive in an assortment of configurations and depend on an assortment of technological and administrative norms (Basoglu and White 2015). Third, the information on the Web is not static; it is created with extraordinary speed. The final trait of Big Data is its veracity (Goes 2014). Because of the open, willful, and frequently unknown contacts on the Web, there is a fundamental ambiguity associated with availability and quality of Web data. A researcher can never be totally certain whether the required information is or will be accessible on the Web and whether this information is dependable enough to be utilized in research (IBM 2018).

Given the volume, assortment, speed, and accuracy of Big Data accessible on the Web, variety and association of this information should barely be possible physically by singular researchers or even enormous research groups (Krotov and Tennyson 2018). Therefore, researchers regularly resort to different technologies and devices to automate a few parts of data collection and organization. This arising practice of utilizing technology for gathering information from the Web is regularly suggested to as Web Scraping (Landers 2016).

Our objective has been accomplished through web scraping and in this respect, it includes retrieving the information from the web as well as, when required, turn unstructured data into a clean and appropriately orchestrated dataset. Finally, as said previously, the entire web scraping method is completely automatized and regenerated spotting updates and generating real time results to scrap the so far legalized publicly available financial information which is the core objective of this thesis work.

III. METHODS AND MATERIALS

Methodology:

In total 7227 firms, 36 different markets and 13 different countries worldwide have been using the GRI standard for their sustainability. The GRI monitoring mission is to inspire decision-makers everywhere to take decisions through the sustainability standard and inter network of GRI. Action for an economy and an environment that is more prosperous. Yahoo Finance is a media property that offers different financial news and statistics, including descriptions of stocks, quotes, press releases, financial reports of GRI companies. It also provides some personal financial management online resources. In this thesis scrape income statements of each GRI company by using web scraping method. Start by putting the code to scrape the data into a function which scrapes Income Statement data for a single symbol and puts it all into a single data frame and export the data to Excel for further analysis.

The international non-profit body is the Global Reporting Initiative (GRI) with a network-based structure. GRI provides free sustainability reporting guides that promote the posting of fiscal, environmental, social and governance outcomes for both corporations and companies. The mixture of drivers such as strategies and effects such as an organization's influence or impact renders the GRI system a systematic way of evaluating a company's sustainable development. In addition, the GRI concept is used by many companies from all over the country and is generally recognized in the commercial world since it includes the financial. The GRI data consists of a number of companies with different sectors and different countries. The data is filtered into “financial services” from the “sector” data with “listed” companies. The GRI of microdata department in from 2016 started to current year collect publicly information of income statements. The “Name” variable is the list of company names. Using the names of companies for collecting the data of financial income statements through Yahoo Finance with web scraping automated process (Boritz, J. E., & No, W. G. 2013).

To the best of our knowledge, we have two publicly available databases, the Yahoo Finance, and the Google Finance. Yahoo Finance provides news, all kinds of financial data, updates, stock quotes, sources of portfolio management and more as a source of information. In an identical framework, different company-profile pages are generated, and all this data as freely accessible sources without any subscriptions as cost. If you are dealing with stock market data, as well as needing a clean, free, and reliable resource, the best option may be Yahoo Finance. The API of Yahoo Finance functions perfectly within the scope of the library of pandas-datareaders. The best overall replacement for Google Finance is the Yahoo Finance API. It's hard to find any information published by Google Finance about it. Yahoo Finance has much more content and services for sustainability than Google Finance. The Yahoo Finance website allows users to scrape the data of financial statements of each company by using ticker symbols. Google Finance does not allow you to scrape financial statements. In this thesis prefers Yahoo Finance website to scrape income statements (Boritz, J. E., & No, W. G. 2013). Yahoo Finance delivers frequent financial information for companies, as well as their stock prices, balance sheet statements, etc, and they are web-scraping approachable. By working on Python, operating the code through this technique drags all of the information and collects it as if it were a static website. This is essential for dragging the stock price, as those are dynamic objects on the webpage and can refresh/update at frequent times. Hence by using Yahoo Finance’s link on the balance sheet or statement of cash flows, the code generates everything properly and keep the statement intact. (Zhu, Zijing, 2020).

Legality and Ethics Framework of Web Scraping:

Web scraping has both positive and bad components, including legal and illegal aspects. The number of tools and technology available on the market for scraping information, but also legality and ethics on web scraping in the 'grey' field. Scrapers collecting information is legal because they access public user data. Currently, no legislation specifically addressing web scraping issues is available. The method remains fraught with a plethora of legal instructions, though web scraping is prevalent. As of now web scraping has guided a set of rules and instructions while scraping information such as " copyright infringement ", "Conditions of use of data", "contract violation", the “Computer Fraud and Abuse Act (CFAA)”, and "chattel infringement", " legal and illegal access data ", ”Trespass to Chattels”. This thesis is following the list of particular conditions and instructions that contribute to a legal and ethical web scraping that before creating a database.

- Does the website's "Terms of Use" policy expressly forbid Web Crawling or Web Scraping?
- Was the thesis material on the website expressly copyrighted?
- Will they involve using the data unlawfully or fraudulently?
- Scraping can create material harm to the website and web server Website Hosting?
- Can the data collected from the website violate individual privacy, research subjects' rights, or the standards of non-discrimination?
- Will the data collected from the database disclose sensitive details about website-affiliated organizations?

The above guidelines are legal literature, along with an analysis of ethics and privacy information systems, together with a collection of specific issues to define broad areas of concern. In this thesis, online search engines for data collection and data scraping are used to discuss and pursue them. In addition to research on ethics and privacy information systems, and a collection of conditions for scraping, we have presented with the study of constitutional literature, using web scraping techniques that need to comply with legal and ethical requirements. This study accessing or strictly observing the requirements of Yahoo Finance services resulting from the use of services or infringement of terms (they are including any liability, website damages, suits and judgements). Consent to the collection, sharing, transfer and use of financial information as defined in the privacy policy is provided when communicating with Yahoo Finance services. Yahoo Finance is copyrighted and open-source for extracting data or scraping, so this data extraction is not detrimental to the hosting owner of the Yahoo Finance website. Legally and without fraudulently using scraped data for various purposes (Krotov, V., & Silva, L. 2018).

Implementation of Web Scraping Technique: Enormous amounts of source information, available on the World Wide Web, are still in the format of a Hypertext Mark-up Language (HTML) page. Automated extraction is difficult because the intended reader was a human. This chapter introduces the motivation and purpose of Information extraction through Web Scraping. Rapid growth of the World Wide Web has significantly changed the way we share, collect, and publish data. Vast amount of information is being stored online, both in structured and unstructured forms. Regarding certain questions or research topics, this has resulted in a new problem—no longer is the concern of data scarcity and inaccessibility but, rather, one of overcoming the tangled masses of online data. These utilizations are often only possible because the existence of automated Web Scraping. Without these techniques, it would be impossible to collect the amount of data repeatedly and in reasonable time (Krotov, V., & Silva, L. 2018).

There are various kits for web scraping. They are Scrapy and BeautifulSoup are potentially two of the biggest libraries known to exist. Scrapy, based on an asynchronous networking library, is a powerful idea that makes it very successful. It also has many characteristics to avoid conventional scraping problems. These include, for instance, subdirectories, requests to be retried, server overload protection, etc. However, owing to the difficulty, the active learning for Scrapy is also considerably greater than for BeautifulSoup. The scrapy library is used mostly for translating and retrieving information from HTML. Both exceptions and complications that arise while scraping must be understood by the developer and taken care of in coding (Miguel Reboiro-Jato, Florentino Fdez-Riverola).

The web application scraper, a lawless text-based Internet control message protocol that manages request-response interactions between a client, typically a Search engine, and a database server, maintains communication with the targeted Web site. The most popular request methods in HTTP are GET, used in resource queries, and Send, used in document verification and file uploading. Until the HTML page is acquired, the Site data scraper can remove the content of interest. For this purpose, pattern matching, in combination with additional logic, is widely implemented.

Alternatively, HTML parser libraries (using the Data Type layout of Web pages) and user defined languages such as XPath and CSS selector syntax are available.

Web Scraping Software Tools: A web scraping tool is a system for automatically extracting from the server. In other words, you cannot expect the original data to be easily retrieved and the exact route. This is where a web scraping tool can be a great help. This allows you to scrap an automatic in an accurate and quick way, process whatever information you need. We receive the final data (pictures, messages, At the end of the table data, etc.) in standardized formats such as CSV, JSON, EXCEL or XML Of the tools for web scraping methods. There are several software tools for web scraping available in the market. Let us explore tools for web scraping. They are Import.IO, Dexi.io, Ripper of the Visual Network, BeautifulSoup, Octoparse, Parsehub, Crawly, Screen.Scraper, Simple Web Extract, OutwitHub.

The web site scraping tools are Oversight Responsibilities Ripper, Screen Scraper, OutWit Hub, Web Material Extractor, and Simple Web Extract. Screen Scraper's basic edition is free, and OutWit Hub has a free Hit app, and all others have a free version of a trial. WebExtractor 360 and Scrapy are open access web scraping tools. A web-based scraper for online use is Import io. In the Mozenda screen scraper app, the primary difference from other scrapers is that it operates your web scraping tasks (Sirisuriya, D. S. 2015).

Web Scraping Software	**Operating System**	**Data Exports Formats**
Screen Scraper	Mac, Unix/Linux Text, Win.	HTML, XML file, HTTP submit form, SQL Script File, MySQL Script File
Web Content Extractor	Win	XML file, HTTP submit form, ODBC Data source, Excel, text, HTML, MS Access DB, SQL Script File, MySQL Script File
Easy Web Extract	Win	MySQL Script File, XML file, HTTP submit form, ODBC Data source, Excel (CSV, TSV), text, HTML, MS Access DB, SQL Script File.
Visual Web Ripper	Win	SQLite, Oracle and OleDB, Customized C# or VB script file output, CSV, Excel, XML, SQL Server, MySQL.
OutWit Hub	Win, Mac OS X, Linux	CSV (TSV), HTML, Excel, or SQL script.

Table 1. Web Scraping Software Tools

As per Table 1, Versions of windows support Site scraping tools for each method, and CSV, XML, and TXT files are standard data formats to be exported. The common features for separate data selection, primary email extraction, remote site, Mac address extraction, mobile number extraction, mobile number retrieval is Import io, Visual Web Ripper, Quick Web Extract, and Web Data Extraction. This research study proposes which according to this comparative study, most web scrapers are mostly very basic and often designed to perform common, specific tasks. They cannot in other words seem to be as versatile and scalable. Universal, as you would predict. Both web scraper developers attempt to crawl all types of web pages for their apps, but this thesis shows that some web scraping programs are appropriate for one task type, and some are perfect for one task type (Sirisuriya, D. S. 2015).

Packages Used:

Python Libraries or Bundles for Web Data Extraction and Scraping: Python is common for being a language at a large standard, but then with a clear stream and a comprehensible style of coding. Python continues to maintain the confidence of a few leading experts in the field of data gathering, retrieval, web data mining and web crawling, considering its extensive and also well archives and a wide variety of applications, including web creation and machine learning, and Robust Object -oriented programming support (Mahito Sugiyama, M Elisabetta Ghisu).

BeautifulSoup: Beautiful Soup is a Python library for getting data out of HTML, XML, and other mark-up languages. A package for parsing HTML and XML reports from the site is Excellent Soup. It parses and sits on HTML or XML parsers such as html5lib and lxml to quickly remove data from HTML.

Selenium: Selenium is a Python system that shows on a webpage like a web driver, opening applications, executing taps, filling structures, looking over and more. The Selenium framework is commonly used in the automatic testing of web apps, but an algorithm for automated data extraction has been found to be useful. We can visit sites and links using web drivers like Chrome Driver for Chrome, and Selenium optimizes the Python loop in a separate Python environment (Mahito Sugiyama, M Elisabetta Ghisu).

Lxml: The lxml collection of Python allows to link the two languages in the analysis, removal, and processing of XML and HTML pages with the same creation in Python and XML. In comparison to Beautiful Soup, lxml offers a superior parsing of Xml with more remarkable speed and consistency, but still works by building and parsing tree architectures of XML hubs. Link to such hubs helps to create relationships between parent and child and applications such as the e-tree framework (ElementTree API). Like Selenium, lxml also reinforces XML Route or XPath, making it easier to parse complicated XML page structures on the web. The substantial efficacy of Beautiful Soup with lxml, however can be consolidated as the two of them sustain and are viable with each other; Beautiful Soup uses it as a parser (Sugiyama, M., Ghisu, M. E., Llinares-López, F., & Borgwardt, K. 2018).

Scrapy: Scrapy is a Python framework for web scraping technique that provides a complete package for developers. It is a powerhouse for web scraping and offers a lot of ways to scrape a web page which requires more time to learn the process of making web crawlers and running them from just one line command by becoming an expert in scrapy functionalities. The self.parse for HTTP specifications, the spider is rendered using a bunch of commands and a goal webpage. When the software for the spider is performed, the offered website page is asked to receive the HTML for the key URL from the given rundown of URLs and decode it according to its limits. Scrapy's system for organizing and encoding and tagging CSS to pass into the HTML tree alongside XPath (Sugiyama, M., Ghisu, M. E., Llinares-López, F., & Borgwardt, K. 2018)

Requests: In order to scrape a website, the first thing we do is to download the page. Python's library of queries enables one to do so exactly. The library uploads a GET information to the web server for us to download the HTML part of the website. There are a few special kinds of requests we can make using specifications, of which only one is GET (Sugiyama, M., Ghisu, M. E., Llinares-López, F., & Borgwardt, K. 2018).

Automated Data Collection of Financials Statements:

Automated Data Collection: Automated data collection refers to the process of automatically extracting data from paper forms using software. Workflow tends to slow down wherever paper is involved because even if documents are meticulously organized and stored, time still has to be spent in physically looking for them. In addition to this, paper documents can be lost or damaged. using automated data collection is managed to overcome all these problems (Landers, R. N., Brusso, R. C).

What Does Automated Data Collection Do?

Any office that scans their documents is already well on their way to automated data collection. The overall transition is one from physical paperwork to electronic, digitized files. When you automate tasks such as invoice and forms processing you greatly improve the efficiency of the office. Businesses expend valuable time, money, and employees on manual data entry. Employees must go through every paper form, identify data, and then transfer it to other forms. With automated data collection, software automatically performs all these tasks (Landers, R. N., Brusso, R. C).

Process of Financial Data Collection:

The GRI text files consists of a list of companies with different sectors and number of countries. Scraping financial income statements by using GRI text files of list companies. Below is the steps of data collection:

Step 1. Pre-processing

- From the GRI text files firstly importing into pandas.
- Then filtered into “financial services” from the “sector” data with “listed” companies. The “Name” variable is the list of company names.
- Using the names of companies for collecting the data of financial income statements through Yahoo Finance with automated process.

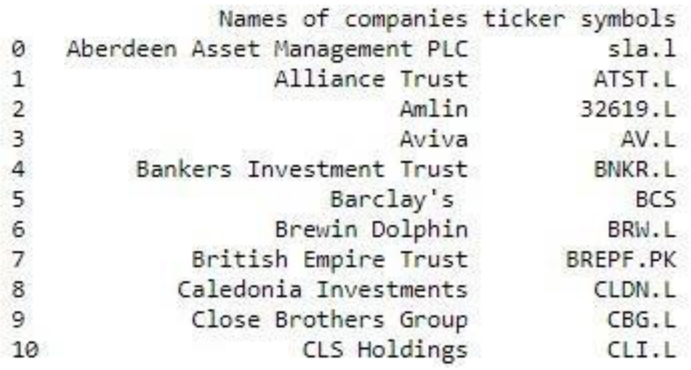
Step 2: Company Name to Ticker Symbol

Yahoo Finance does not allow for the searching of information using names, only tickers. Using python packages fetching into yahoo finance official website and extract that ticker symbol of name of the company. The python allows that extract number of companies’ ticker symbol at a time.

Steps of Company Name to Ticker Symbols:

- List of company names import to search with google search library in yahoo finance of official website.
- Yahoo Finance does not allow for the searching information using name of companies, it allows by only ticker symbols.
- Using company names that yahoo finance is converts to ticker symbol and search for the company information.
- We are scraping the ticker symbol of company name from the url’s links.
- Every url links are splitting into the backslashes (/) that ticker symbols of company name available at the before end of url link.

- Scraping ticker symbols and creates dataframes using python algorithms.



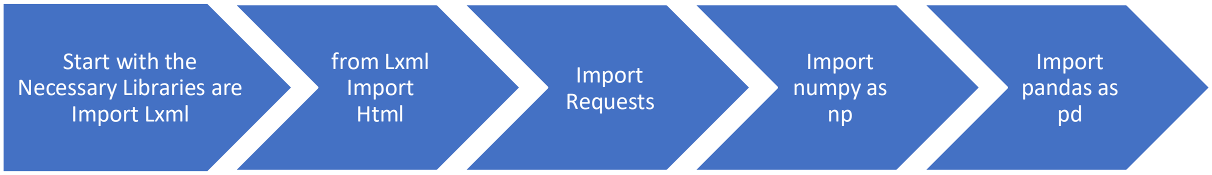
```
              Names of companies ticker symbols
0   Aberdeen Asset Management PLC        sla.l
1                  Alliance Trust       ATST.L
2                           Amlin      32619.L
3                           Aviva         AV.L
4         Bankers Investment Trust      BNKR.L
5                       Barclay's          BCS
6                  Brewin Dolphin        BRW.L
7            British Empire Trust      BREPF.PK
8           Caledonia Investments       CLDN.L
9             Close Brothers Group       CBG.L
10                   CLS Holdings        CLI.L
```

Figure 1: Dataframe of Company Name to Ticker Symbols

Step 3: Data Extraction

Let us extend the code to support scraping multiple symbols. Start by putting the code to scrape the data into a function which scrapes Income Statement data for a single symbol and puts it all into a single data frame.Web sites are written using HTML, which means that each web page is a structured document.

Steps to follow to Import the Libraries:



- Lxml is an extensive library written for parsing XML and HTML documents very quickly, even handling messed up tags in the process.
- Next, we will use requests.get to retrieve the web page with our data. Parse the page with lxml, so that we want start doing some XPATH queries to extract the data that we want.
- That extracted information stores in empty list, creates data frame

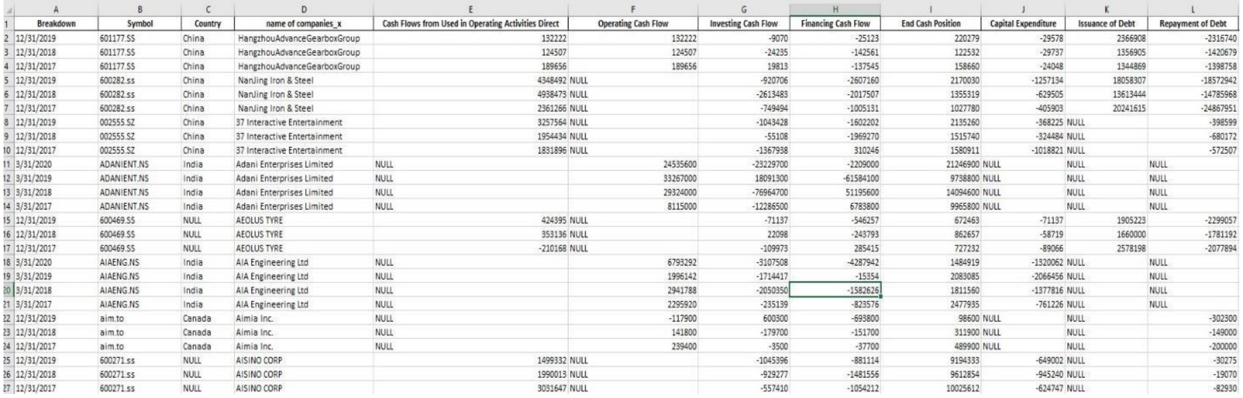
Step 4: Data Cleaning

- Set the index to the first column: ‘period ending’.
- Drop the null values and transpose the data frame.
- Remove “, ” , “ - ” thousands of operators and convers columns to the float64.

As per clean up data steps is judging quality of data or good data requires an examination of its characteristics are reliability, consistency, and completeness. Now Data frame consists quality of data or good data.

Step 5: Building Excel Database

- Now that we have successfully scraped the Income Statement.
- We export the data to Excel and create a database for academic, scientific research and further analysis. It is possible to export the Pandas data Frame to Excel via Excel Writer.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Breakdown	Symbol	Country	name of companies_x	Cash Flows from Used in Operating Activities Direct	Operating Cash Flow	Investing Cash Flow	Financing Cash Flow	End Cash Position	Capital Expenditure	Issuance of Debt	Repayment of Debt
2	12/31/2019	601177.SS	China	HangzhouAdvanceGearboxGroup	132222	132222	-9070	-25123	220279	-29578	2366908	-2316740
3	12/31/2018	601177.SS	China	HangzhouAdvanceGearboxGroup	124507	124507	-24235	-142561	122532	-29737	1356905	-1420679
4	12/31/2017	601177.SS	China	HangzhouAdvanceGearboxGroup	189656	189656	19813	-137545	158660	-24048	1344869	-1398758
5	12/31/2019	600282.ss	China	NanJing Iron & Steel	4348492	NULL	-920706	-2607160	2170030	-1257134	18058307	-18572942
6	12/31/2018	600282.ss	China	NanJing Iron & Steel	4938473	NULL	-2613483	-2017507	1355319	-629505	13613444	-14785968
7	12/31/2017	600282.ss	China	NanJing Iron & Steel	2361266	NULL	-749494	-1005131	1027780	-405903	20241615	-24867951
8	12/31/2019	002555.SZ	China	37 Interactive Entertainment	3257564	NULL	-1043428	-1602202	2135260	-368225	NULL	-398599
9	12/31/2018	002555.SZ	China	37 Interactive Entertainment	1954434	NULL	-55108	-1969270	1515740	-324484	NULL	-680172
10	12/31/2017	002555.SZ	China	37 Interactive Entertainment	1831896	NULL	-1367938	310246	1580911	-1018821	NULL	-572507
11	3/31/2020	ADANIENT.NS	India	Adani Enterprises Limited	NULL	24535600	-23229700	-2209000	21246900	NULL	NULL	NULL
12	3/31/2019	ADANIENT.NS	India	Adani Enterprises Limited	NULL	33267000	18091300	-61584100	9738800	NULL	NULL	NULL
13	3/31/2018	ADANIENT.NS	India	Adani Enterprises Limited	NULL	29324000	-76964700	51195600	14094600	NULL	NULL	NULL
14	3/31/2017	ADANIENT.NS	India	Adani Enterprises Limited	NULL	8115000	-12286500	6783800	9965800	NULL	NULL	NULL
15	12/31/2019	600469.SS	NULL	AEOLUS TYRE	424395	NULL	-71137	-546257	672463	-71137	1905223	-2299057
16	12/31/2018	600469.SS	NULL	AEOLUS TYRE	353136	NULL	22098	-243793	862657	-58719	1660000	-1781192
17	12/31/2017	600469.SS	NULL	AEOLUS TYRE	-210168	NULL	-109973	285415	727232	-89066	2578198	-2077894
18	3/31/2020	AIAENG.NS	India	AIA Engineering Ltd	NULL	6793292	-3107508	-4287942	1484919	-1320062	NULL	NULL
19	3/31/2019	AIAENG.NS	India	AIA Engineering Ltd	NULL	1996142	-1714417	-15354	2083085	-2066456	NULL	NULL
20	3/31/2018	AIAENG.NS	India	AIA Engineering Ltd	NULL	2941788	-2050350	-1582626	1811560	-1377816	NULL	NULL
21	3/31/2017	AIAENG.NS	India	AIA Engineering Ltd	NULL	2295920	-235139	-823576	2477935	-761226	NULL	NULL
22	12/31/2019	aim.to	Canada	Aimia Inc.	NULL	-117900	600300	-693800	98600	NULL	NULL	-302300
23	12/31/2018	aim.to	Canada	Aimia Inc.	NULL	141800	-179700	-151700	311900	NULL	NULL	-149000
24	12/31/2017	aim.to	Canada	Aimia Inc.	NULL	239400	-3500	-37700	489900	NULL	NULL	-200000
25	12/31/2019	600271.ss	NULL	AISINO CORP	1499332	NULL	-1045396	-881114	9194333	-649002	NULL	-30275
26	12/31/2018	600271.ss	NULL	AISINO CORP	1990013	NULL	-929277	-1481556	9612854	-945240	NULL	-19070
27	12/31/2017	600271.ss	NULL	AISINO CORP	3031647	NULL	-557410	-1054212	10025612	-624747	NULL	-82930

Figure 2: Micro Database of Financial Data

Test the Quality of Web Data: The consistency of the data being collected is the most significant factor of any web scraping result. Web scraping infrastructure would never be able to assist with academic and further deployments without reliable high-quality results (Kontokostas, D., Westphal).

Manual Testing: In order to scan manually, at least 10 records (list rows) against the source page, provide someone who is not the data set maker. Information is retained in the registry. This is really the safest method for the data extraction to spot any obvious problems. By using manual testing checks whether it is number of companies of financial income statements in the database.

Coverage Testing: Testing coverage is understood to be the scale of the dataset before collecting information scrapers. There is a proper number of companies in the data collection. Only claim that estimated size of the database (in memory or number of file records). By using coverage testing checking that size of the database.

Completeness Testing: Scraped data must verify the completeness of the data i.e., whether the data collected consists of null values, missing values that should be retrieved, and the source code must guarantee the information is accessible on the page. The collection of financial income statements is following that completeness testing with accuracy.

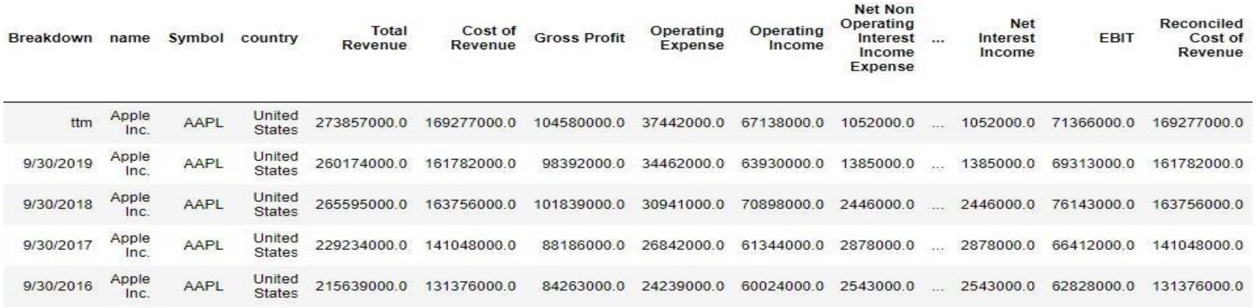
Analysis of Web Data Extraction: After completion and creating database of financial income statements used for data analysis is the process by which inspection, washing, transformation, and simulation analyse a collected set of data. To access and interpret rich types of data, data analysts use data analysis techniques to extract valuable information or observations from data in different ways. It is a process in which raw data is obtained and then turned into information useful for consumers to make decisions. The different components of data analytics are data acquisition, sorting, cleaning, analysis and modelling (S. S., & Bodke, K. R. 2018). Data collected can be gathered from multiple outlets, such as interviews, newspapers, journals, polls, the Internet, etc. The Web is one of the key means of static raw data acquisition that is readily available. Therefore, to retrieve data from the oddly structured web world, web crawling methods such as web wrappers, HTTP coding, etc. are used. This entails collecting website knowledge (Parvez, M. S., Tasneem, K. S. A., Rajendra).

Data Quality and Impact on Decision-Making:

The operational knowledge gained from site crawling and strategic decision-making for the organizations. Bad quality of data can lead to poor choices that result in poor productivity. Because of the authenticity of web data due to anonymous data created on the web, this can lead to inaccurate decision-making. Despite these issues, the need for market analytics create negative interest for such knowledge because of these sometimes-decision-making methods that can lead to erroneous, unreliable, incompleteness. Web scraping operations that gather and spread web pages of low quality are facilitated and promoted. Users often lack the capacity to modify aggregated and transmitted precision data to complicate these issues, contributing to the potential explosive spike in what this knowledge misrepresents. Decision Making techniques due to these concerns produce the miscalculation between true and inaccurate details on the site.

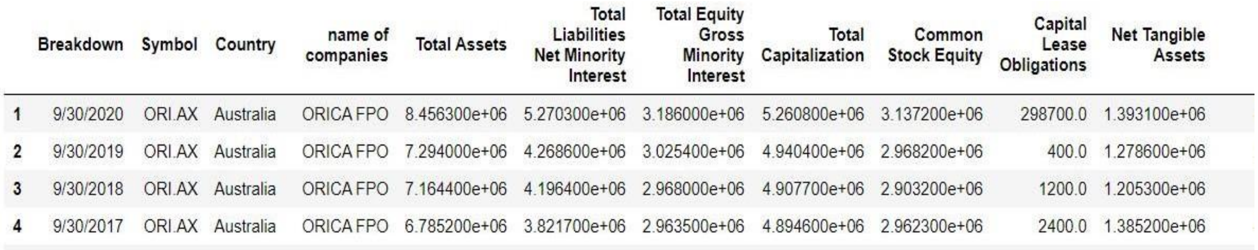
IV. RESULTS & DISCUSSION

As per the aim of our thesis, the objective is principally to collect financial income statements' data from different sectors and multiple countries with automation process from GRI's database. The GRI text files approximately consist of 7227 companies. From the total number of companies of the data is filtered with "listed" companies. They are totally 2252 companies, some of the companies they don't have income statements of data now finally in this thesis collecting data of 2112 companies, 36 different sectors(they are Agriculture, automotive, aviation, chemicals, commercial services, computers, conglomerates, constructions, construction materials, energy, energy utilities, equipment, financial services, food and beverage products, forest and paper products, healthcare products, household and personal products, logistics, media, metal products, mining, non-profit, other, public agency, railroad, real estates, retailers, technology hardware, telecommunications, textiles and apparel, tobacco, tourism, toys, universities, water management and water utilities) and 13 different countries(they are Australia, Japan, Germany, Canada, Russian Federation, Brazil, France, Spain, India, Italy, Mainland China, United Kingdom of Great Britain and Northern Ireland, United States Of America). The GRI of microdata department in from 2016 started to 2020 collect publicly information of income statements. Each sector consists of a number of companies and collecting financial data of all companies from the sector at a time. In the process of data collection handling high complexity data, huge amount of data and for completion of this collection process of duration time is thirty (30) minutes for each sector because of several steps in there. Firstly, GRI text files are filtered into 'listed' companies of each sector, from name of the company converts into ticker symbol, using ticker symbols fetching into yahoo finance official website and scraping name of company's financials income statements. After completion of data collection building a database. Data Processing is the method by which a gathered collection of data is evaluated by inspection, cleaning, transformation, and simulation. The quality of outcome of the database consists of a completeness with highly accurate. The quality test of database is checks by the manual testing, coverage testing and completeness testing.



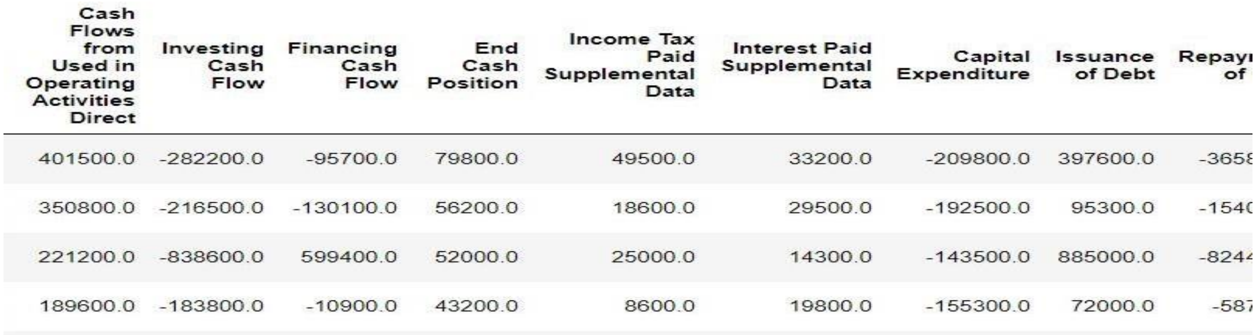
Breakdown	name	Symbol	country	Total Revenue	Cost of Revenue	Gross Profit	Operating Expense	Operating Income	Net Non Operating Interest Income Expense	...	Net Interest Income	EBIT	Reconciled Cost of Revenue
ttm	Apple Inc.	AAPL	United States	273857000.0	169277000.0	104580000.0	37442000.0	67138000.0	1052000.0	...	1052000.0	71366000.0	169277000.0
9/30/2019	Apple Inc.	AAPL	United States	260174000.0	161782000.0	98392000.0	34462000.0	63930000.0	1385000.0	...	1385000.0	69313000.0	161782000.0
9/30/2018	Apple Inc.	AAPL	United States	265595000.0	163756000.0	101839000.0	30941000.0	70898000.0	2446000.0	...	2446000.0	76143000.0	163756000.0
9/30/2017	Apple Inc.	AAPL	United States	229234000.0	141048000.0	88186000.0	26842000.0	61344000.0	2878000.0	...	2878000.0	66412000.0	141048000.0
9/30/2016	Apple Inc.	AAPL	United States	215639000.0	131376000.0	84263000.0	24239000.0	60024000.0	2543000.0	...	2543000.0	62828000.0	131376000.0

Figure 3: Data Frame of Financial Statement



	Breakdown	Symbol	Country	name of companies	Total Assets	Total Liabilities Net Minority Interest	Total Equity Gross Minority Interest	Total Capitalization	Common Stock Equity	Capital Lease Obligations	Net Tangible Assets
1	9/30/2020	ORI.AX	Australia	ORICA FPO	8.456300e+06	5.270300e+06	3.186000e+06	5.260800e+06	3.137200e+06	298700.0	1.393100e+06
2	9/30/2019	ORI.AX	Australia	ORICA FPO	7.294000e+06	4.268600e+06	3.025400e+06	4.940400e+06	2.968200e+06	400.0	1.278600e+06
3	9/30/2018	ORI.AX	Australia	ORICA FPO	7.164400e+06	4.196400e+06	2.968000e+06	4.907700e+06	2.903200e+06	1200.0	1.205300e+06
4	9/30/2017	ORI.AX	Australia	ORICA FPO	6.785200e+06	3.821700e+06	2.963500e+06	4.894600e+06	2.962300e+06	2400.0	1.385200e+06

Figure 4: Data Frame of Balance Sheet



Cash Flows from Used in Operating Activities Direct	Investing Cash Flow	Financing Cash Flow	End Cash Position	Income Tax Paid Supplemental Data	Interest Paid Supplemental Data	Capital Expenditure	Issuance of Debt	Repay of
401500.0	-282200.0	-95700.0	79800.0	49500.0	33200.0	-209800.0	397600.0	-3658
350800.0	-216500.0	-130100.0	56200.0	18600.0	29500.0	-192500.0	95300.0	-1540
221200.0	-838600.0	599400.0	52000.0	25000.0	14300.0	-143500.0	885000.0	-8244
189600.0	-183800.0	-10900.0	43200.0	8600.0	19800.0	-155300.0	72000.0	-587

Figure 5: Data Frame of Financial Cash Flow Statement

Financial Income Statements are formal documents conveying the operations of the corporation and the company's financial results. Financial statements, also known as profit and loss (p&l) statements, summarize both income and losses for a given period, including the cumulative effect of income, gain, expense, and Transactions for losses. Income statements are regularly published as quarterly and annual accounts, indicating funding. The aim of an income statement is to demonstrate the financial success of a business over a period of time. Although all financial information serves to build an image of the financial stability of a company, an income statement is one of the most valuable records an executive committee of a company may review and individual investors can review, since it provides a thorough analysis of income and expenditures over a reporting period. This includes:

- Net Revenue: During a tax period, the sum of money a company gets in.
- Operating Expenses: During a reporting cycle, the amount of capital an organization pays.
- Gross Profit: Total profits minus costs of goods sold.
- Operating Revenue: Sales benefit minus operating costs.
- Net Revenue: Less tax on pre-tax income.

Financial Balance Sheet is a financial statement that at a particular point in time, records the assets, liabilities and shareholder equity of a corporation and provides a framework for calculating return rates and assessing the capital structure. It offers a summary of what is owned and owed by a company, as well as the amount spent by shareholders. The balance sheet adheres to the following accounting equation, where assets on one side, and liabilities plus shareholders' equity on the other, balance out:

Assets = Liabilities + Shareholders' Equity

This formula is intuitive: a company must pay for all the things it owns (assets) by either borrowing money (taking on liabilities) or taking it from investors (issuing shareholders' equity).

Financial Cash Flow Statement contains information from its continuing operations and foreign investment sources on all cash inflows a business receives. The cash flow statement contains cash generated by the company, the amount of which is called net cash flow, by operations, acquisition, and financing. Cash flow from operations, which contains purchases from all operational business activities, is the first segment of the cash flow statement. The second part of the cash flow statement is cash flow from investment, which is the product of profits and losses from investment.

The final section includes an analysis of cash used by debt and equity, is the cash flow from funding.

VI. SUMMARY & FUTUREWORK

Conventional data collection frameworks are less pragmatic, being replaced with new systems that basically “reap” information rather than collecting. With the outbreak of www, and in this day and age of net-worked and circulated applications, the Internet is turning into the basic information source as for all intents and purposes everything is shared through it and, consequently, it contains array of data anticipating for retrieval. Web scraping permit researchers and organizations to screen data over the net and furthermore a computerized information assortment on a regular basis. They likewise grant an exceptionally adaptable way to deal with information extraction in a proficient and quick manner, visiting an enormous number of pages. The huge field of web scraping applications are truly amazing. Competitor price monitoring, market data, financial statements are only a couple instances of information that scrapers can collect over the net. This thesis work addressed the web scraping strategies which are confronting numerous difficulties as the extraction of the data are not excessively simple and the high volume of web scraping can cause regulatory damage for the concerns. Still, we cannot ignore the undeniable fact that web scraping method guarantee that the assembled data is exact, dependable, and having higher confidentiality as the data present is in enormous sum which is hard to oversee and keep up.

Here in this thesis, we have expressed that our motivation to accomplish the necessities of the financial and non-financial companies for their own research, educational and business purposes. The financial companies can also be immensely benefited to step up their financial activities with ease as well as effectiveness. We have discussed below those financial activities which will provide them with direct practical benefits:

Identification of Trend: In order to determine the trend in business, the business owner should prepare and compare financial statements over different periods. This helps the company to understand what segments are growing well, and which business segment needs further analysis and reinvestment or complete exit at once.

Benchmarking: Every company must have a vision. The company must have defined objectives and goals in order to plan a vision. By analysing the previous financial statements already prepared and audited, the objective of the financial statements is to prepare a blueprint for the future. Budgets help to keep costs in line with earnings and revenues. The financial statements present the company's short- and long-term obligations. If the owner seeks to increase his company, he must look at the financial position statements and deduce the logic about whether to reduce current liabilities in order to apply for further investment. Based on sales, assets, and assets, lenders look at the financial statements and assess the prospect of operation. The cash flow statement splits the declaration into operating, spending, and financial elements. A cash flow analysis allows one to understand whether the organization works within the framework of a cyclical revenue stream or a stable revenue model. It also allows the organization to manage and maintain the company's expenditures within the revenue model under which it works.

The web-based way to deal with data extraction method can improve the nature of our information system. This additional value is positively seen in the utilization of such techniques portrayed in this thesis work, as financial as well as the non-financial sectors can be taken as the clear-cut example of the phenomena where the operation provided by the traditional way of data collection is obsolete and the automated technique like web scraping has become the necessity as well as an apparent reality.

VII. Acknowledgement

We would like to thank our thesis supervisors Kenneth Carling and Asif M Huq for the help and support they have given us over the course of this work.

VIII. REFERENCES

✓ Auger, Alex-Adrien Concrete Example of Web Scraping with Financial Data, Sipios, Dec 28, 2018

✓ Basoglu, K. A., and White, Jr. C. E. 2015. “Inline XBRL versus XBRL for SEC Reporting,” Journal of Emerging Technologies in Accounting (12:1), pp. 189-199.

✓ Bloomberg, Michael R. (August 27, 2001) Bloomberg by Bloomberg Paperback.

✓ Buchanan, E. 2017. “Internet Research Ethics: Twenty Years Later,” in Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts, M. Zimmer and K. Kinder-Kurlanda (eds.), Bern, Switzerland: Peter Lang International Academic Publishers, pp. xxix-xxxiii.

✓ Calabrese, A., Costa, R., Ghiron, N. L., & Menichini, T. (2017). Materiality analysis in sustainability reporting: a method for making it work in practice. European Journal of Sustainable Development, 6(3), 439-439.

✓ Cisco Systems. 2016. “Cisco Visual Networking Index: Forecast and Methodology, 2014-2019,” White Paper. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

✓ Clarke, R. 2016. “Big Data, Big Risks.” Information Systems Journal (26:1).

✓ Cohen and Fan. Learning page-independent heuristics for extracting data from web pages. CN, 31(11-16), 1999

✓ Constantiou, I. D., and Kallinikos, J. 2015. “New Games, New Rules: Big Data and the Changing Context of Strategy,” Journal of Information Technology (30:1), pp. 44-57.

✓ Dryer, A.J., and Stockton, J. 2013. “Internet 'Data Scraping': A Primer for Counseling Clients,” New York Law Journal. Retrieved from https://www.law.com/newyorklawjournal/almID/1202610687621

✓ Zhu, Zijing. Nov 5, 2020. Web Scraping Yahoo Finance News

✓ Exploiting web scraping in a collaborative filtering- based approach to web advertising

✓ Faustina Johnson and Santosh Kumar Gupta.Web Content Mining Techniques: A Survey, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012

✓ Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797.

✓ Goes, P. B. 2014. “Editor's Comments: Big Data and IS Research,” MIS Quarterly (38:3), pp. iii-viii.

✓ Hirschey, J. K. 2014. “Symbiotic Relationships: Pragmatic Acceptance of Data Scraping,” Berkeley Technology Law Journal (29), pp. 897-927.

✓ http://web.archive.org/web/20201028171854/http://yej.hotelristorantelasiesta.it/gdp-analysis-in-python.html, 28 Oct 2020

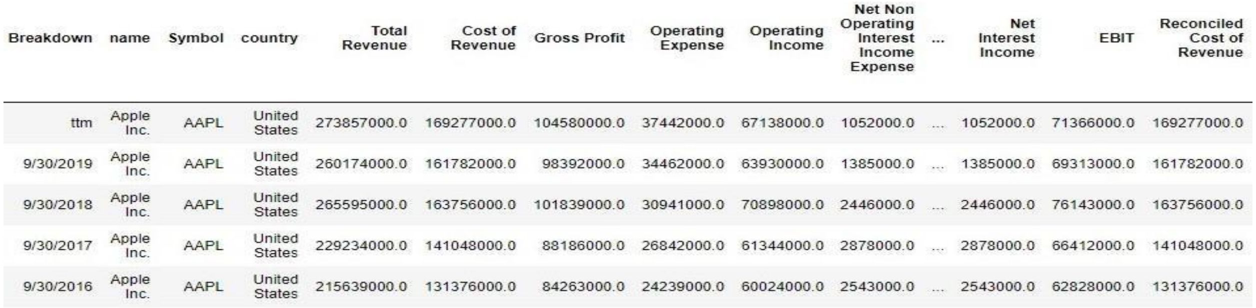
- ✓ https://www.import.io/post/how-to-test-the-quality-of-web-data/
- ✓ IBM. 2018. “The Four V’s of Big Data,” Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data
- ✓ Initiative, G. R. (2012). Global reporting initiative. Online at: https://www. Global reporting. org/Pages/default. aspx (20 Dec 2012).
- ✓ Ittelson, Thomas, (August 15, 2009) Financial Statements: A Step-by-Step Guide to Understanding and Creating Financial Reports Paperback.
- ✓ Ives, B., and Krotov, V. 2006. “Anything You Search Can Be Used Against You in a Court of Law: Data Mining in Search Archives,” Communications of the Association for Information Systems (18:1), pp. 593-611.
- ✓ Ives, B., Palese, B., and Rodriguez, J. A. 2016. “Enhancing Customer Service through the Internet of Things and Digital Data Streams,” MIS Quarterly Executive (15:4).
- ✓ Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping.
- ✓ Krotov, V., and Tennyson, M. 2018. “Scraping Financial Data from the Web Using the R Language,” Journal of Emerging Technologies in Accounting, Forthcoming
- ✓ Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychological methods, 21(4), 475.
- ✓ Landers, R. N., Brusso, R. C., Cavanaugh, K. J., and Collmus, A. B. 2016. “A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for use in Psychological Research,” Psychological Methods (21:4), pp. 475-492.
- ✓ Lee, C. F., Lee, J., Chang, J. R., & Tai, T. (2016). Data Collection, Presentation, and Yahoo Finance. In Essentials of Excel, Excel VBA, SAS and Minitab for Statistical and Financial Analyses (pp. 13-45). Springer, Cham.
- ✓ Light, B., & McGrath, K. 2010. “Ethics and Social Networking Sit es: A Disclosive Analysis of Facebook,” Information Technology & People (23:44), pp. 290-311.
- ✓ Lindgren, C, Huq, A, Li, Y, and Carling, K, (2019), “Current practices of CSR around the globe: An exploratory text mining study”, (in progress)
- ✓ List of Web Harvester, Data Scraper,Web Scraping Software and Tools, n.d.WebData Scraping. URL http://webdata-scraping.com/webscraping-software/
- ✓ Mahito Sugiyama, M Elisabetta Ghisu, Felipe Llinares-López, Karsten Borgwardt, graphkernels: R and Python packages for graph comparison, Bioinformatics, Volume 34, Issue 3, 01 February 2018, Pages 530–532, https://doi.org/10.1093/bioinformatics/btx602
- ✓ Mason, R. O. 1986. “Four Ethical Issues of the Information Age,” MIS Quarterly, (10:1), pp. 5-12.

- McCandless, David. 1 December 2012. Information is Beautiful (New Edition) HarperCollins Publishers; UK ed. edition (1 December 2012)
- Mitchell, R., 2018. Web scraping with Python: Collecting more data from the modern web. O'Reilly Media, Inc.
- Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. 2015. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining, Chichester, UK: John Wiley & Sons, Ltd.
- Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons, 2014.
- Parvez, M. S., Tasneem, K. S. A., Rajendra, S. S., & Bodke, K. R. (2018, January). Analysis of different web data extraction techniques. In 2018 International Conference on Smart City and Emerging Technology (ICSCET) (pp. 1-7). IEEE
- Patel Jay M (2020) Getting Structured Data from the Internet
- Paul, R., and Elder, L. 2006. The Thinker's Guide to Understanding the Foundations of Ethical Reasoning. Foundation for Critical Thinking.
- Perez, Martin, Digital Marketing Specialist, ParseHub. Is Web Scraping Legal?: Web Scraping Publicly Available Data, ParseHub; Oct 07, 2019
- R. Baumgartner, K. Fro¨schl, M. Hronsky, M. P¨ottler, and N. Walchhofer. Semantic online tourism market monitoring. Proc. 17th ENTER eTourism International Conference, 2010
- R. Baumgartner, W. Gatterbauer, and G. Gottlob. Web data extraction system. Encyclopedia of Database Systems, pages 3465–3471, 2009
- S.C.M. de S Sirisuriya,2015, A Comparative Study on Web Scraping. Proceedings of 8th International Research Conference, KDU.
- Schenker Jennifer L. (April 28, 2008); Investors want to know more about companies than everand they are increasingly turning to the nonfinancial metrics of Asset4
- Sirisuriya, D. S. (2015). A comparative study on web scraping.
- Snell, J., and Menaldo, N. 2016. “Web Scraping in an Era of Big Data 2.0,” Bloomberg BNA. Retrieved from https://www.bna.com/web-scraping-era-n57982073780/
- Watson, H. J. 2014. “Tutorial: Big Data Analytics: Concepts, Technologies, and Applications,” Communications of the Association for Information Systems (34:1), pp. 1247-1268.
- Web Data Extraction, Applications and Techniques: A SurveyEmilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartnerd
- WebSelF: A Web Scraping Framework Jakob Thomsen1, Erik Ernst1 , Claus Brabrand2 , and Michael Schwartzbach
- www.octoparse.com›blog›scrape-financial-data-without-python, Monday, August 31, 2020

IX. APPENDICES

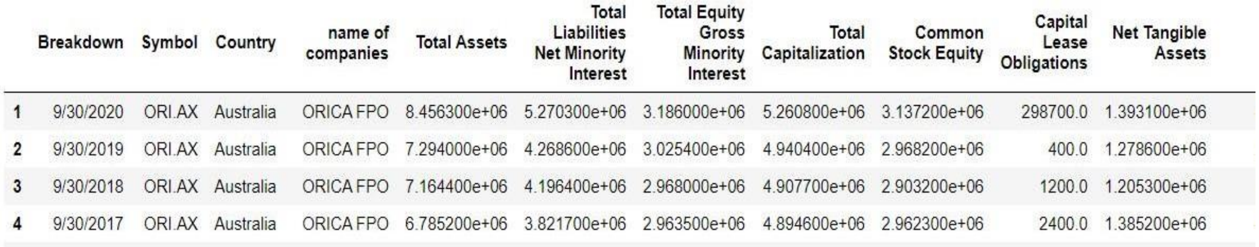
1. All the source code and the necessary data is available at **https://github.com/Srujanpolepally7/Financial_data.git**

2. The complete financial data of micro database:



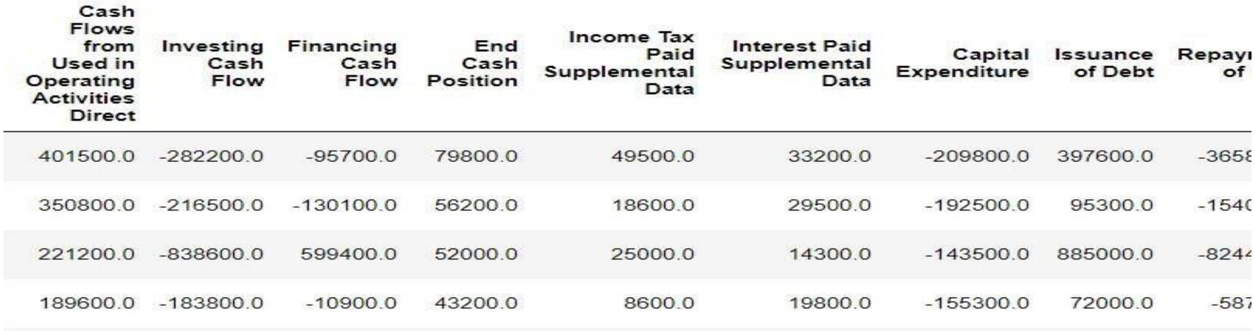
Breakdown	name	Symbol	country	Total Revenue	Cost of Revenue	Gross Profit	Operating Expense	Operating Income	Net Non Operating Interest Income Expense	...	Net Interest Income	EBIT	Reconciled Cost of Revenue
ttm	Apple Inc.	AAPL	United States	273857000.0	169277000.0	104580000.0	37442000.0	67138000.0	1052000.0	...	1052000.0	71366000.0	169277000.0
9/30/2019	Apple Inc.	AAPL	United States	260174000.0	161782000.0	98392000.0	34462000.0	63930000.0	1385000.0	...	1385000.0	69313000.0	161782000.0
9/30/2018	Apple Inc.	AAPL	United States	265595000.0	163756000.0	101839000.0	30941000.0	70898000.0	2446000.0	...	2446000.0	76143000.0	163756000.0
9/30/2017	Apple Inc.	AAPL	United States	229234000.0	141048000.0	88186000.0	26842000.0	61344000.0	2878000.0	...	2878000.0	66412000.0	141048000.0
9/30/2016	Apple Inc.	AAPL	United States	215639000.0	131376000.0	84263000.0	24239000.0	60024000.0	2543000.0	...	2543000.0	62828000.0	131376000.0

Data Frame of Financial Statement



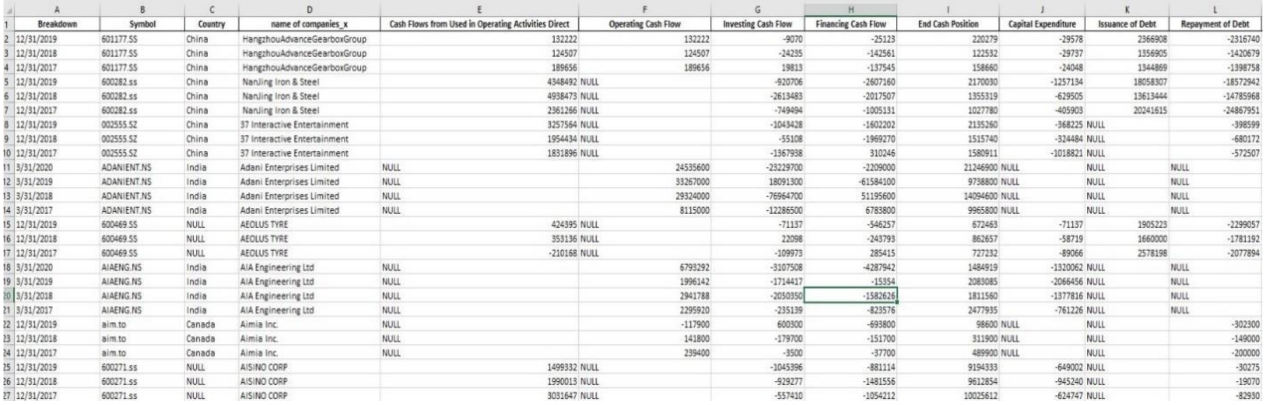
	Breakdown	Symbol	Country	name of companies	Total Assets	Total Liabilities Net Minority Interest	Total Equity Gross Minority Interest	Total Capitalization	Common Stock Equity	Capital Lease Obligations	Net Tangible Assets
1	9/30/2020	ORI.AX	Australia	ORICA FPO	8.456300e+06	5.270300e+06	3.186000e+06	5.260800e+06	3.137200e+06	298700.0	1.393100e+06
2	9/30/2019	ORI.AX	Australia	ORICA FPO	7.294000e+06	4.268600e+06	3.025400e+06	4.940400e+06	2.968200e+06	400.0	1.278600e+06
3	9/30/2018	ORI.AX	Australia	ORICA FPO	7.164400e+06	4.196400e+06	2.968000e+06	4.907700e+06	2.903200e+06	1200.0	1.205300e+06
4	9/30/2017	ORI.AX	Australia	ORICA FPO	6.785200e+06	3.821700e+06	2.963500e+06	4.894600e+06	2.962300e+06	2400.0	1.385200e+06

Data Frame of Balance Sheet



Cash Flows from Used in Operating Activities Direct	Investing Cash Flow	Financing Cash Flow	End Cash Position	Income Tax Paid Supplemental Data	Interest Paid Supplemental Data	Capital Expenditure	Issuance of Debt	Repay of
401500.0	-282200.0	-95700.0	79800.0	49500.0	33200.0	-209800.0	397600.0	-3658
350800.0	-216500.0	-130100.0	56200.0	18600.0	29500.0	-192500.0	95300.0	-1540
221200.0	-838600.0	599400.0	52000.0	25000.0	14300.0	-143500.0	885000.0	-8244
189600.0	-183800.0	-10900.0	43200.0	8600.0	19800.0	-155300.0	72000.0	-587

Data Frame of Financial Cash Flow Statement



	A	B	C	D	E	F	G	H	I	J	K	L
1	Breakdown	Symbol	Country	name of companies_x	Cash Flows from Used in Operating Activities Direct	Operating Cash Flow	Investing Cash Flow	Financing Cash Flow	End Cash Position	Capital Expenditure	Issuance of Debt	Repayment of Debt
2	12/31/2019	601177.SS	China	HangzhouAdvanceGearboxGroup	132222	132222	-9070	-25123	220279	-29578	2366908	-2316740
3	12/31/2018	601177.SS	China	HangzhouAdvanceGearboxGroup	124507	124507	-24235	-142561	122532	-29737	1356905	-1420679
4	12/31/2017	601177.SS	China	HangzhouAdvanceGearboxGroup	189656	189656	19813	-137545	158660	-24048	1344869	-1398758
5	12/31/2019	600282.ss	China	NanJing Iron & Steel	4348492	NULL	-920706	-2607160	2170030	-1257134	18058307	-18572942
6	12/31/2018	600282.ss	China	NanJing Iron & Steel	4938473	NULL	-2613483	-2017507	1355319	-629505	13613444	-14785968
7	12/31/2017	600282.ss	China	NanJing Iron & Steel	2361266	NULL	-749494	-1005131	1027780	-405903	20241615	-24867951
8	12/31/2019	002555.SZ	China	37 Interactive Entertainment	3257564	NULL	-1043428	-1602202	2135260	-368225	NULL	-398599
9	12/31/2018	002555.SZ	China	37 Interactive Entertainment	1954434	NULL	-55108	-1969270	1515740	-324484	NULL	-680172
10	12/31/2017	002555.SZ	China	37 Interactive Entertainment	1831896	NULL	-1367938	310246	1580911	-1018821	NULL	-572507
11	3/31/2020	ADANIENT.NS	India	Adani Enterprises Limited	NULL	24535600	-23229700	-2209000	21246900	NULL	NULL	NULL
12	3/31/2019	ADANIENT.NS	India	Adani Enterprises Limited	NULL	33267000	18091300	-61584100	9738800	NULL	NULL	NULL
13	3/31/2018	ADANIENT.NS	India	Adani Enterprises Limited	NULL	29324000	-76964700	51195600	14094600	NULL	NULL	NULL
14	3/31/2017	ADANIENT.NS	India	Adani Enterprises Limited	NULL	8115000	-12286500	6783800	9965800	NULL	NULL	NULL
15	12/31/2019	600469.SS	NULL	AEOLUS TYRE	424395	NULL	-71137	-546257	672463	-71137	1905223	-2299057
16	12/31/2018	600469.SS	NULL	AEOLUS TYRE	353136	NULL	22098	-243793	862657	-58719	1660000	-1781192
17	12/31/2017	600469.SS	NULL	AEOLUS TYRE	-210168	NULL	-109973	285415	727232	-89066	2578198	-2077894
18	3/31/2020	AIAENG.NS	India	AIA Engineering Ltd	NULL	6793292	-3107508	-4287942	1484919	-1320062	NULL	NULL
19	3/31/2019	AIAENG.NS	India	AIA Engineering Ltd	NULL	1996142	-1714417	-15354	2083085	-2066456	NULL	NULL
20	3/31/2018	AIAENG.NS	India	AIA Engineering Ltd	NULL	2941788	-2050350	-1582626	1811560	-1377816	NULL	NULL
21	3/31/2017	AIAENG.NS	India	AIA Engineering Ltd	NULL	2295920	-235139	-823576	2477935	-761226	NULL	NULL
22	12/31/2019	aim.to	Canada	Aimia Inc.	NULL	-117900	600300	-693800	98600	NULL	NULL	-302300
23	12/31/2018	aim.to	Canada	Aimia Inc.	NULL	141800	-179700	-151700	311900	NULL	NULL	-149000
24	12/31/2017	aim.to	Canada	Aimia Inc.	NULL	239400	-3500	-37700	489900	NULL	NULL	-200000
25	12/31/2019	600271.ss	NULL	AISINO CORP	1499332	NULL	-1045396	-881114	9194333	-649002	NULL	-30275
26	12/31/2018	600271.ss	NULL	AISINO CORP	1990013	NULL	-929277	-1481556	9612854	-945240	NULL	-19070
27	12/31/2017	600271.ss	NULL	AISINO CORP	3031647	NULL	-557410	-1054212	10025612	-624747	NULL	-82930

Micro Database of Financial Data