ICIC Express Letters
Part B: Applications
Volume **13**, Number **1**, January **2022**

ICIC International ©2022 ISSN 2185-2766
pp. **107–114**

WEBSCRAPING DATA LABELING SYSTEM ON LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY OF RODENT TUBER FOR EFFICIENCY OF SUPERVISED LEARNING PREPROCESSING

Iwan Binanto[1,2,*], Harco Leslie Hendric Spits Warnars[1]
Nesti Fronika Sianipar[3,4] and Widodo Budiharto[1]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[3]Food Technology Department, Faculty of Engineering
[4]Research Interest Group Biotechnology
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
spits.hendric@binus.ac.id; { nsianipar; wbudiharto }@binus.edu

[2]Informatics Department
Sanata Dharma University
Kampus 3, Jl. Paingan, Krodan, Maguwoharjo, Kec. Depok, Kabupaten Sleman
Daerah Istimewa Yogyakarta 55281, Indonesia
*Corresponding author: iwan@usd.ac.id

Received March 2021; accepted June 2021

Abstract. *Although it seems like a trivial process, data labeling is a cornerstone for supervised learning. It becomes problematic when the data is huge, like the data of Liquid Chromatography-Mass Spectrometry (LC-MS), especially manually worked by human as it is time-consuming and needs a high-accuracy. It is the main problem. In this study, the data source for labeling is an online database and does not provide an Application Programming Interface (API). We developed an automatic labeling model which connects to an online database to get the label, utilizing webscraping technique. It is usually utilized for market analysis purposes, but in this study we utilized to retrieve label data, which is the name of the chemical compound. The model successfully labeled data as desired and gets 91.6% time efficiency compared to manual. It cannot be 100% because the model restricts requests to the server from being considered an attack. The result of this labeling can be used for supervised learning.*

Keywords: LC-MS, Preprocessing, Rodent Tuber, Huge data labeling, Webscraping, Efficiency

1. **Introduction.** Supervised learning requires labeled data. It is not a difficult thing but it does in terms of thoroughness, labor-intensive task, and time-consuming when the data is huge. Cowie et al. said that work data labeling is the result of the tension between complexity and simplicity [1], and it is not a trivial process.

Raw data of Liquid Chromatography-Mass Spectrometry (LC-MS) is a huge data. It contains millions of data points, or after integration and peak extraction, there are hundreds of chromatographic peaks [2]. It provides advantages for highly complex biological samples [3].

LC-MS is widely used, especially in the interpretation or identification of the content of chemical compounds in a biological sample [3-7]. However, it is not trivial to processing this data to be useful information. Many techniques are used to process this data. One of them is supervised learning. This process requires data labeling as preprocessing.

LC-MS data consist of hundreds of thousands of mass per charge (m/z), retention time, and intensity [8]. Likewise, the LC-MS data used in this paper are LC-MS data of Rodent

DOI: 10.24507/icicelb.13.01.107

Tuber from the studies of Sianipar et al. [9-13]. One of the most important things to do with this data is labeling by giving all names of the chemical compounds. Currently, it is done manually by copying the m/z values and pasting them one by one to the massbank.jp site. It is tedious and time-consuming for human labor.

There are more than 700,000 records in each of the LC-MS data of Rodent Tuber. We have calculated, to get one compound name manually, at the fastest time it takes about 20 seconds. If there are about 700,000 records, it will take 20 $\times$ 700,000 seconds. By doing a simple calculation, it will take 3,888.89 hours or more than 5 months if doing this 24 hours per day to complete the labeling. If it takes more than 20 seconds to get a single compound name, it needs longer time to complete this job. This is the main problem. The answer to this problem is automation. However, there is another problem because massbank.jp does not provide an Application Programming Interface (API).

Massbank.jp is a website of an online database of compounds which is the official database of the Mass Spectrometry Society of Japan. It is a distributed database. Each research group provided data from its MassBank data server, which was distributed on the Internet [14,15]. Unfortunately, we do not have any information about the Application Programming Interface (API) on this website. Hence, we utilize webscraping technique to get chemical compound names from massbank.jp.

This paper describes an automated labeling system model with an online database utilizing webscraping technique to speed up the labeling process, although webscraping is usually used for business purposes. However, it can also be used to find specific information on a web page [16-18].

The academic contribution of this paper is to provide a new paradigm that webscraping technique can be utilized for data labeling so it can greatly speed up the process. In this study, 91.6% time efficiency was obtained compared to manual work. The practical contribution is to help scientists, especially in the pharmaceutical field, to speed up data labeling based on mass per charge (m/z).

This paper is categorized as follows: Section 2 describes the related works, Section 3 describes the preliminaries work for this study, Section 4 describes the developed model, Section 5 describes the result, and Section 6 focuses on conclusions and future work.

2. **Related Works.**

The webscraping technique is used to get the content from websites to analyze specific structured or unstructured data. It has been developed in the private sector for business purposes, especially in market analysis, but it offers substantial benefits to those searching for specific information [16-18].

Several studies have provided several methods and frameworks to labeling data. Yang et al. developed a game-based framework for crowdsourced label data which involved machine learning [19]. Tseng et al. developed a method for data labeling utilizing tri-training which used three classifiers [20]. Kamminga et al. [21] synchronized sensors and cameras for data labeling and compared two approaches, namely synchronization using visual key and synchronization using real-time clocks. These studies deal with huge data.

Other studies do not require data labeling, because using an unsupervised learning method, one of them is conducted by Albert et al. which is towards the application of unsupervised machine learning methods for analyzing protein conformational transitions to extract information about their structural similarity [22].

LC-MS has been widely used to determine the content in test solutions. Kharyuk et al. combined LC-MS and machine learning to train and validate plant species identification algorithms [2] as well as Nazarenko et al. [23]. Roux-Dalvai et al. also combined LC-MS and machine learning to identify bacterial species in urine specimens [24]. Planinc et al. combined LC-MS and PCA for a better detection of changes in N-glycosylation profiles of therapeutic glycoproteins [25].

3. **Preliminaries Work.** The data used here were obtained from studies of Sianipar et al. [9-13] which produce 10 datasets from the outputs of the LCMS instrument. They are proprietary raw data that only can be read by the instrument that produced them. We have to convert them to .xlsx files so they are human-readable and easier to analyze. To do this, there are two stages, 1) raw data conversion to open format which is .mzXML, utilized by an open-source software, namely Proteowizard version 3 for Windows, developed by Chambers et al. [26], and 2) .mzXML conversion to .xlsx, utilized by our developed software by Python programming. In this paper, only one dataset is used.

Using large native data like this would be costly. Therefore, data sampling will be carried out. This LC-MS data of Rodent Tuber is time-series data [27], so it required sample per period time to obtain a representative sample that represents the actual dataset. Linear systematic sampling is chosen because it is simple and can represent the actual dataset [28].

To obtain a systematic sampling, given that N is the number of population of elements, while n is the number of samples desired, so if N/n is an integer, then k = N/n; otherwise, let k be the next integer after N/n. After that, find a random integer R between 1 and k, defining the sample as the unit numbered R, R + k, R + 2k, and so on. The initial unit R selected is called "random start" and k is called "sampling interval" [28,29].

Based on that, we develop Algorithm 1. It is applied to one of ten of the LC-MS dataset of Rodent Tuber which will be used in this paper. In this case, it is applied to LC-MS dataset of Rodent Tuber which contains 985,924 records and resulting 18,271 records. If this sample is manually labeled as a simple calculation above, then 20 × 18,271 seconds, it will take about 101.5 hours or about 4 days if doing this 24 hours per day.

Algorithm 1. LinearSystematicSampling

```
listDF = read(Dataset)
Create list arrRT by pivot technique to get "Retention Time" without duplication.
arrSample = [ ]
arrTempRT = [ ]
arrTempI = [ ]
k = len(listDF)/18000
for n in range(len(arrRT)):
    arrTemp = [ ]
    arrTempInt = [ ]
    for i in range(len(listDF)):
        if listDF[i][2] == arrRT[n]:
            arrTemp.append(listDF[i][0])
            arrTempInt.append(listDF[i][1])
    indexes = np.arange(random.randint(0,k), len(arrTemp), step = k)
    for i in indexes:
        arrSample.append(arrTemp[i])
        arrTempRT.append(arrRT[n])
        arrTempI.append(arrTempInt[i])
```

To find the compound's name, m/z data is required to be queried to the MassBank [14]. First of all it gets a complete URL with its parameters. It is done by filling in provided forms which is the "Exact Mass". One of the m/z data copied from .xlsx and pasted in "Exact Mass" fills form's massbank.jp and is click the Search button. After the Search button is clicked, the complete URL with its options emerges in the browser's address bar as shown in Figure 1 and it is saved for further use. In Figure 1 there is an underlined

```
http://www.massbank.jp/Result.jsp?compound=&op1=and&mz=68.78610229&tol=0
.3&op2=and&formula=&type=quick&searchType=keyword&sortKey=not&sortAction
=1&pageNo=1&exec=&inst_grp=ESI&inst=CE-ESI-TOF&inst=ESI-ITFT&inst=ESI-IT
TOF&inst=ESI-QIT&inst=ESI-QTOF&inst=ESI-TOF&inst=LC-ESI-IT&inst=LC-ESI-I
TFT&inst=LC-ESI-ITTOF&inst=LC-ESI-Q&inst=LC-ESI-QFT&inst=LC-ESI-QIT&inst
=LC-ESI-QQ&inst=LC-ESI-QQQ&inst=LC-ESI-QTOF&inst=LC-ESI-TOF&ms=MS2&ion=0
```

FIGURE 1. The complete URL

and bold section, and the value of this section will be changed according to the data, as input for webscraping.

Frequently, there is more than one compound name with a different m/z value as the search's result. It is very rare to get a compound name with the same value as the input.

4. **Developed Model.** The preliminaries above are the preparation and initial step of our developed model, which are sample dataset in .xlsx format, and complete URL for webscraping. The developed model is shown in Figure 2.

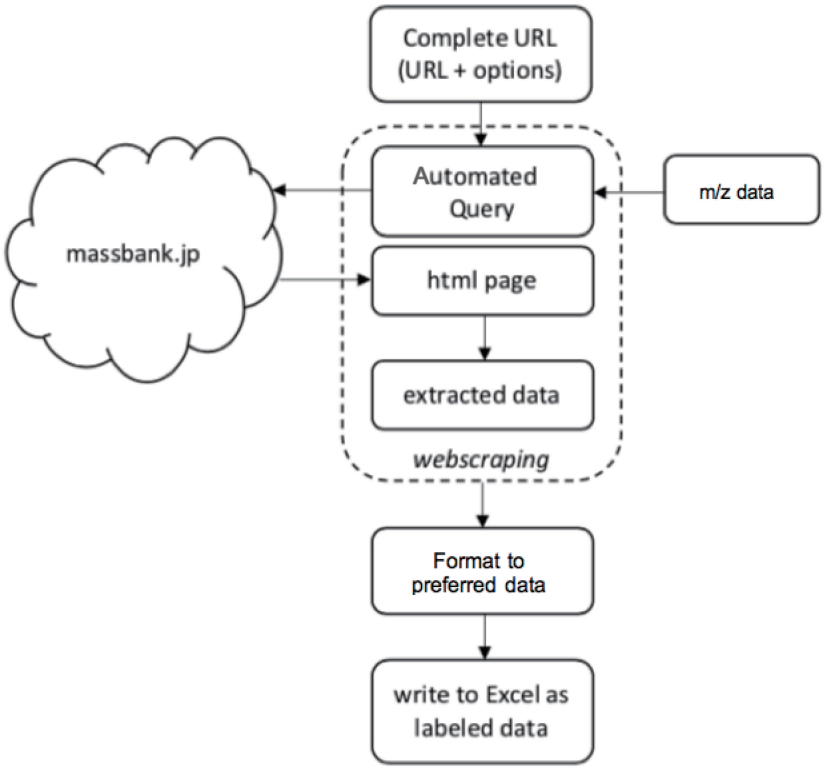

FIGURE 2. Developed model

The next step is "Webscaping" stage by developing Algorithm 2. It is implementing webscraping technique for gathering compound names from massbank.jp. It is implemented in Python programming by utilizing the library namely Requests [30] and BeautifulSoup [31].

There are more than one compound name per inputted m/z value with different exact m/z value as a result of webscraping technique or manually. It means more than one compound name in one record/row. Therefore, it is necessary to separate data to get one compound name in one row. We developed an algorithm to process this, and the impact has duplication of the m/z.

As a result, we have a .xlsx file containing the compound name with duplication of inputted m/z data, m/z data which is closest to the inputted m/z, and the "-" character because no compound name was found which correlated to the m/z. It is necessary

Algorithm 2. GetCompoundName

```
1: MAZZ = read(Dataset["m/z"])
2: arrNamaSenyawa = [ ]
   arrRealMassa = [ ]
3: for a in range(len(MAZZ)):
       mz = str(MAZZ[a])
       MZ = urllib.quote(mz)
       url = URL_lengkap
       try:
           r = requests.get(url,timeout = None)
       except:
           continue
       soup = BS(r.text,'html.parser')
       tabel = soup.find_all('td', class_ = 'treeLayout2', width = "142")
       if (len(tabel)) > 0:
           rumus = [ ]
           for r in soup.find_all('td', attrs = {'width': '142'}):
               rumus.append(r.get_text(strip = True))
           namaSenyawa = [ ]
           for nama in soup.find_all('a', attrs = {'class': 'noLinkImg'}):
               namaSenyawa.append(nama.get_text(strip = True))
           massa = [ ]
           for m in soup.find_all('td', attrs = {'width': '122'}):
               massa.append(m.get_text(strip = True))
           massa_float = [float(item) for item in massa]
           mz_float = float(mz)
           selisih = [ ]
           for x in massa_float:
               selisih.append(abs(mz_float-x))
           npSelisih = np.array(selisih)
           minSelisih = npSelisih.min()
           pos = [ ]
           for i in range(len(selisih)):
               if (round(minSelisih,10) == round(selisih[i],10)):
                   pos.append(i)
               i = +1
           arrNamaSenyawa.append(encoded_n[pos[0]])
           arrRealMassa.append(massa[pos[0]])
       else:
           arrNamaSenyawa.append("-")
           arrRealMassa.append(0)
       if (a % 3000 == 0):
           time.sleep(30)
       else:
           time.sleep(0.25)
       a = +1
4: Repeat point 3 until all list MAZZ processed
```

Algorithm 3. Cleaning "-"

```
df ← excel file
arr ← convert values df to array
arrNew ← create new array
while i < len(arr) do
    if arr[i] != "-" then
        arrNew.append(arr[i])
end while
dfNew ← arrNew
write_to_excel(dfNew, "cleanData.xlsx")
```

to clean up the data so that the "-" character is removed. This process is "Format to Preferred Data" stage and writes as Algorithm 3. The output is "Labeled Data".

5. **Result and Discussion.** In this paper, we developed a model for automated labeling LC-MS data of Rodent Tuber to speed up data labeling. This model utilized webscraping technique to gather information because of the absence of Application Programming Interface.

The model was implemented in Python programming with its library and success to gather information. However, we encounter obstacles that occur frequently disconnected from the online database. In our opinion, this happened because it was considered an attack to the server. To handle this problem, the queries to the server are slowed down by adding a delay to the query loop. In our experiment, for every 3000-4000 records/rows, it always disconnects from an online database, it seems because the server refuses the connection. In our opinion, it is caused by a fast continuous connection and it is considered as a Denial-of-Service attack [32]. So, we create some delays in limiting requests when making an HTTP request to the online database.

However, the long delay will impact the longer execution time will be. It is a tug war. In this study, we use delay 30 seconds when the records reached a multiple of 3000 and 0.25 seconds for the loop delay as shown in Algorithm 2.

After applying Algorithm 3 to the dataset that has 18,271 records, the time for labeling is 8.5 hours. It is very fast compared to manual labeling and obtains time efficiency. If this works manually, it will take time 101.5 hours. So, the efficiency we get is $101.5 - 8.5 = 93$ hours or 91.6% efficient.

6. **Conclusions.** Webscraping technique is used to get the content from websites and developed in the private sector for business purposes. We developed a model based on this technique for data labeling. This technique is utilized because we do not have any information about the Application Programming Interface (API) on the target. It is not an easy way, and it takes extra work to get the preferred data. Webscraping in our developed model is done by Request and Beautiful Soup libraries in the Python programming environment.

Overall, the developed model can perform as expected, although there are still shortcomings. This model is not perfect yet but perfectly applied to data in the range of 3000-5000 records or rows. We found incomplete compound names with an extra character "..." which means there is an additional compound name. It is from original content on massbank.jp which gathered. This compound name cannot be taken completely yet. However, the result still can be used as a labeled data and can be used for further processing, in this case for supervised learning.

In the future, we are planning to modify this model in the following two aspects: 1) fixing the frequent disconnects, and 2) completing the compound name which is still in the form "...".

REFERENCES

[1] R. Cowie, C. Cox, J. C. Martin, A. Batliner, D. Heylen and K. Karpouzis, Issues in data labelling, in *Emotion-Oriented Systems. Cognitive Technologies*, R. Cowie, C. Pelachaud and P. Petta (eds.), Berlin, Heidelberg, Springer, DOI: 10.1007/978-3-642-15184-2_13, 2011.

[2] P. Kharyuk et al., Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task, *Sci. Rep.*, vol.8, no.1, pp.1-12, DOI: 10.1038/s41598-018-35399-z, 2018.

[3] M. Brown et al., Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets, *Bioinformatics*, vol.27, no.8, pp.1108-1112, DOI: 10.1093/bioinformatics/btr079, 2011.

[4] M. Gerlich and S. Neumann, MetFusion: Integration of compound identification strategies, *J. Mass Spectrom.*, vol.48, no.3, pp.291-298, DOI: 10.1002/jms.3123, 2013.

[5] C. Guijas et al., METLIN: A technology platform for identifying knowns and unknowns, *Anal. Chem.*, vol.90, no.5, pp.3156-3164, DOI: 10.1021/acs.analchem.7b04424, 2018.

[6] B. Zhou, J. F. Xiao, L. Tuli and H. W. Ressom, LC-MS-based metabolomics, *Mol. Biosyst.*, vol.8, no.2, pp.470-481, DOI: 10.1039/C1MB05350G, 2012.

[7] J. Listgarten and A. Emili, Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry, *Mol. Cell. Proteomics*, vol.4, no.4, pp.419-434, DOI: 10.1074/mcp.R500005-MCP200, 2005.

[8] F. Fernández-Albert, *Machine Learning Methods for the Analysis of Liquid Chromatography-Mass Spectrometry datasets in Metabolomics*, Ph.D. Thesis, Universitat Politècnica De Catalunya, 2014.

[9] D. Laurent, N. F. Sianipar, Chelen, Listiarini and A. Wantho, Analysis of genetic diversity of Indonesia Rodent Tuber (Typhonium flagelliforme Lodd.) cultivars based on RAPD marker, *The 3rd International Conference on Biological Science 2013 (ICBS-2013)*, vol.2, pp.139-145, 2015.

[10] N. F. Sianipar, R. Purnamaningsih and Rosaria, Bioactive compounds of fourth generation gamma-irradiated *Typhoniumflagelliforme* Lodd. mutants based on gas chromatography-mass spectrometry, *IOP Conference Series: Earth and Environmental Science*, DOI: 10.1088/1755-1315/41/1/012025, 2016.

[11] N. F. Sianipar, R. Purnamaningsih, D. L. Gumanti, Rosaria and M. Vidianti, Analysis of gamma irradiated-third generation mutants of rodent tuber (Typhonium flagelliforme Lodd.) based on morphology, RAPD, and GC-MS markers, *Pertanika J. Trop. Agric. Sci.*, vol.40, no.1, pp.185-202, 2017.

[12] N. F. Sianipar and R. Purnamaningsih, Molecular detection of putative mutant clones of Rodent Tuber (Typhonium Flaelliforme Lodd.) CV. pekalongan using RAPD markers, *Malays. Appl. Biol.*, vol.47, no.2, pp.1-8, 2018.

[13] N. F. Sianipar, K. Assidqi, R. Purnamaningsih and T. Herlina, *In vitro* cytotoxic activity of Rodent Tuber mutant plant (Typhonium flagelliforme Lodd.) against to MCF-7 breast cancer cell line, *Asian J. Pharm. Clin. Res.*, vol.12, no.3, pp.185-189, DOI: 10.22159/ajpcr.2019.v12i3.29651, 2019.

[14] H. Horai et al., MassBank: A public repository for sharing mass spectral data for life sciences, *J. Mass Spectrom.*, vol.45, no.7, pp.703-714, DOI: 10.1002/jms.1777, 2010.

[15] R. Arakawa et al., Proposal: Recommendation on measuring and providing mass spectra as chemical information of organic molecules (secondary publication), *Mass Spectrom.*, vol.8, no.1, pp.1-6, DOI: 10.5702/massspectrometry.A0076, 2019.

[16] R. McAlister, Webscraping as an investigation tool to identify potential human trafficking operations in Romania, *Proc. of 2015 ACM Web Sci. Conf.*, DOI: 10.1145/2786451.2786510, 2015.

[17] M. Herrmann and L. Hoyden, Applied webscraping in market research, *The 1st International Conference on Advanced Research Methods and Analytics*, DOI: 10.4995/carma2016.2016.3131, 2016.

[18] M. Shreesha, S. B. Srikara and R. Manjesh, A novel approach for news extraction using webscraping technique, *The 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA2018)*, pp.359-362, DOI: 10.21467/proceedings.1.56, 2018.

[19] J. Yang, J. Fan, Z. Wei, G. Li, T. Liu and X. Du, A game-based framework for crowdsourced data labeling, *VLDB J.*, DOI: 10.1007/s00778-020-00613-w, 2020.

[20] C. M. Tseng, T. W. Huang and T. J. Liu, Data labeling with novel decision module of tri-training, *2020 2nd Int. Conf. Comput. Commun. Internet (ICCCI2020)*, pp.82-87, DOI: 10.1109/ICCCI49374.2020.9145968, 2020.

[21] J. W. Kamminga, M. Jones, K. Seppi, N. Meratnia and P. J. M. Havinga, Synchronization between sensors and cameras in movement data labeling frameworks, *Proc. of the 2nd Workshop on Data Acquisition To Analysis (DATA'19)*, no.11, pp.37-39, DOI: 10.1145/3359427.3361920, 2019.

[22] S. Albert, M. Teletin and G. Czibula, Analyzing protein data using unsupervised learning techniques, *International Journal of Innovative Computing, Information and Control*, vol.14, no.3, pp.861-880, 2018.

[23] D. V. Nazarenko, P. V. Kharyuk, I. V. Oseledets, I. A. Rodin and O. A. Shpigun, Machine learning for LC-MS medicinal plants identification, *Chemom. Intell. Lab. Syst.*, vol.156, pp.174-180, DOI: 10.1016/j.chemolab.2016.06.003, 2016.

[24] F. Roux-Dalvai et al., Fast and accurate bacterial species identification in urine specimens using LC-MS/MS mass spectrometry and machine learning, *Mol. Cell. Proteomics*, vol.5, 2019.

[25] A. Planinc et al., LC-MS analysis combined with principal component analysis and soft independent modelling by class analogy for a better detection of changes in N-glycosylation profiles of therapeutic glycoproteins, *Anal. Bioanal. Chem.*, vol.409, no.2, pp.477-485, DOI: 10.1007/s00216-016-9683-9, 2017.

[26] M. C. Chambers et al., A cross-platform toolkit for mass spectrometry and proteomics, *Nat. Biotechnol.*, vol.30, no.10, pp.918-920, DOI: 10.1038/nbt.2377, 2012.

[27] I. Binanto, H. Leslie, H. Spits, N. F. Sianipar and W. Budiharto, Understanding LCMS data for identification of chemical compounds contained in Rodent Tuber: Timeseries or not, *Syst. Rev. Pharm.*, vol.12, no.1, pp.648-654, 2021.

[28] R. Arnab, Systematic sampling, in *Survey Sampling Theory and Applications*, R. Arnab (ed.), Academic Press, 2017.

[29] S. L. Lohr, *Sampling: Design and Analysis*, 2nd Edition, Brooks/Cole, Cengage Learning, Boston, 2010.

[30] *Requests: HTTP for Humans*$^{\text{TM}}$ *– Requests 2.24.0 Documentation*, https://requests.readthedocs.io/en/master/, Accessed on Nov. 05, 2020.

[31] L. Richardson, *Beautiful Soup: We Called Him Tortoise Because He Taught Us*, https://www.crummy.com/software/BeautifulSoup/, Accessed on Nov. 05, 2020.

[32] A. V. Saurkar, K. G. Pathare and S. A. Gode, An overview on web scraping techniques and tools, *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, vol.4, no.4, pp.363-367, 2018.