Webscraping Data Labeling System

by Iwan Binanto

Submission date: 09-May-2021 09:00AM (UTC+0700)
Submission ID: 1367616810
File name: ICICELB-2103-021-rev.docx (152.27K)
Word count: 3473
Character count: 19750

ICIC Express Letters
Part B: Applications
Volume **11**, Number **9**, September **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 1-8

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], Widodo Budiharto[1]

[1]Computer Science Department, BINUS Graduate Program, Doctor Computer Science Bina Nusantara University, Jakarta, Indonesia

[2]Informatics Department, Sanata Dharma University, Yogyakarta, Indonesia
*Corresponding Author: iwan@usd.ac.id

[3]Food Technology Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia

[4]Research Interest Group Biotechnology, Bina Nusantara University, Jakarta, Indonesia

Received March 2020; accepted April 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], 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 provide 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 which are LC-MS data of Rodent Tuber from the studies of Sianipar et. al [9]–[13]. One of the most important thing to

do with this data is labeling by giving all names of the chemical compounds. Currently, it is do 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 x 700,000 seconds. By doing a simple calculation, it will take 3,888.89 hours or more than 5 months if do this 24 hours per day to complete the labeling. If it takes more than 20 seconds to get a single compound name, thus it needs longer time to complete this job. This is the main problem. The answer to this problem is automation. But 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. But 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 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 ini 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 using three classifiers [20]. Kamminga et.al. [21] synchronized sensors and cameras for data labeling and comparing 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. combining LC-MS and machine learning to training and validating plant species identification algorithms [2] as well as Nazarenko et. al. [23]. Roux-Dalvai et. al. also

combine LC-MS and machine learning to identify bacterial species in urine specimens [24]. Planinc et. al. combining 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 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 applied to one of ten of the LC-MS dataset of Rodent Tuber which will be used in this paper. In this case, it 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 x 18,271 seconds, it will take about 101.5 hours or about 4 days if do this 24 hours per day.

Algoritma 1 LinearSystematicSampling

```
1:  listDF = read(Dataset)
2:  Create list arrRT by pivot technique to get "Retention Time" without duplication.
3:  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 is get a complete URL with its parameters. It is done by fill in provided forms which is the "Exact Mass". One of the m/z data copied from .xlsx and pasted in "Exact Mass" filling form's massbank.jp and click the Search button. After the Search button clicked, the complete URL with its options emerged 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 and bold section, the value of this section will be changed according to the data, as input for webscraping.

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-ITTOF&inst=ESI-QIT&inst=ESI-QTOF&inst=ESI-TOF&inst=LC-ESI-IT&inst=LC-ESI-ITFT&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

Frequently, there is more than one compound names 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 shown in Figure 2.

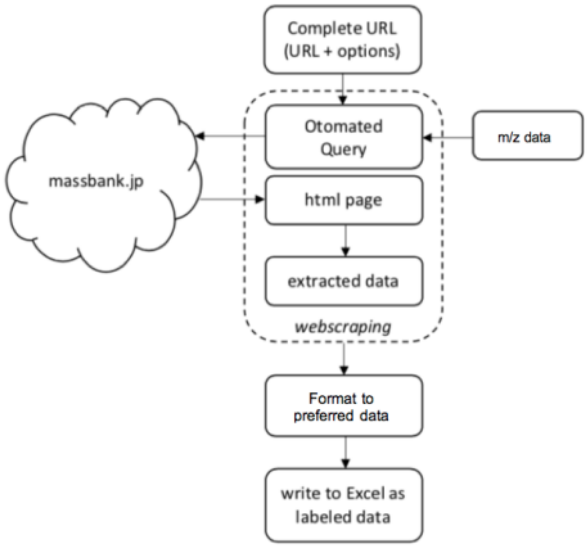

FIGURE 2. Developed Model

The next step is "Webscaping" stage by developing Algorithm 2. It is implementing webscraping technique for gathering compound names from masbank.jp. It is implemented

in Python programming by utilizing the library namely Requests [30] and BeautifulSoup [31].

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
```

There are more than one compound names 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 names in one record/row. Therefore, it is necessary to separate data to get one

compound name in one row. We developed an algorithm to processes this, the impact has duplication of the m/z.

As 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 the m/z. It is necessary 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".

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")
```

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 utilizing 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. But 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, every 3000 - 4000 records/rows, it always disconnects from an online database, it seems because the server refusing 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.

But 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 is work 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, 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, 2) complete 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," *Cogn. Technol.*, no. 9783642151835, pp. 213–241, 2011, doi: 10.1007/978-3-642-15184-2_13.

[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, 2018, doi: 10.1038/s41598-018-35399-z.

[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, 2011, doi: 10.1093/bioinformatics/btr079.

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

[5] C. Guijas *et al.*, "METLIN: A Technology Platform for Identifying Knowns and Unknowns," *Anal. Chem.*, vol. 90, no. 5, pp. 3156–3164, 2018, doi: 10.1021/acs.analchem.7b04424.

[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, 2012, doi: 10.1039/C1MB05350G.

[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, 2005, doi: 10.1074/mcp.R500005-MCP200.

[8] F. Fernández-Albert, "Machine Learning Methods for the Analysis of Liquid Chromatography-Mass Spectrometry datasets in Metabolomics," UNIVERSITAT POLIT` ECNICA 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)," in *The 3rd International Conference on Biological Science 2013 (The 3rd ICBS-2013)*, 2015, vol. 2, pp. 139–145.

[10] N. F. Sianipar, R. Purnamaningsih, and Rosaria, "Bioactive compounds of fourth generation gamma-irradiated Typhoniumflagelliforme Lodd . mutants based on gas chromatography-mass spectrometry," 2016, doi: 10.1088/1755-1315/41/1/012025.

[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, 2019, doi: 10.22159/ajpcr.2019.v12i3.29651.

[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, 2010, doi: 10.1002/jms.1777.
[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, 2019, doi: 10.5702/massspectrometry.A0076.
[16] R. McAlister, "Webscraping as an investigation tool to identify potential human trafficking operations in Romania," *Proc. 2015 ACM Web Sci. Conf.*, no. 2013, 2015, doi: 10.1145/2786451.2786510.
[17] M. Herrmann and L. Hoyden, "Applied Webscraping in Market Research," in *First International Conference on Advanced Research Methods and Analytics*, 2016, p. 2016, doi: 10.4995/carma2016.2016.3131.
[18] M. Shreesha, S. B. Srikara, and R. Manjesh, "A Novel Approach for News Extraction Using Webscraping Technique," in *The 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA 2018)*, 2018, no. NCICCNDA 2018, pp. 359–362, doi: 10.21467/proceedings.1.56.
[19] J. Yang, J. Fan, Z. Wei, G. Li, T. Liu, and X. Du, "A game-based framework for crowdsourced data labeling," *VLDB J.*, 2020, doi: 10.1007/s00778-020-00613-w.
[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, ICCCI 2020*, pp. 82–87, 2020, doi: 10.1109/ICCCI49374.2020.9145968.
[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," *DATA 2019 - Proc. 2nd ACM Work. Data Acquis. To Anal. Part SenSys 2019*, no. November, pp. 37–39, 2019, doi: 10.1145/3359427.3361920.
[22] S. Albert, M. Teletin, and G. Czibula, "Analyzing protein data using unsupervised learning techniques," *Int. J. Innov. Comput. Inf. 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, 2016, doi: 10.1016/j.chemolab.2016.06.003.
[24] Florence 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, 2017, doi: 10.1007/s00216-016-9683-9.
[26] M. C. Chambers *et al.*, "A cross-platform toolkit for mass spectrometry and proteomics," *Nat. Biotechnol.*, vol. 30, no. 10, pp. 918–920, 2012, doi: 10.1038/nbt.2377.
[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, "Chapter 4 - Systematic Sampling," in *Survey Sampling Theory and Applications*, R. Arnab, Ed. Academic Press, 2017, pp. 89–115.
[29] S. L. Lohr, *Sampling: Design and Analysis*, Second Edi. Boston: Brooks/Cole, Cengage Learning, 2010.
[30] "Requests: HTTP for Humans™ — Requests 2.24.0 documentation." https://requests.readthedocs.io/en/master/ (accessed Nov. 05, 2020).
[31] L. Richardson, "Beautiful Soup: We called him Tortoise because he taught us.," *www.crummy.com*. https://www.crummy.com/software/BeautifulSoup/ (accessed 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, [Online]. Available: http://www.ijfrcsce.org.

Webscraping Data Labeling System

ORIGINALITY REPORT

13%	7%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Strathclyde Student Paper	3%
2	eprints.binus.ac.id Internet Source	2%
3	silo.pub Internet Source	1%
4	Submitted to The British College Student Paper	1%
5	Submitted to President University Student Paper	1%
6	onlinelibrary.wiley.com Internet Source	1%
7	sersc.org Internet Source	<1%
8	link.springer.com Internet Source	<1%
9	journals.plos.org Internet Source	<1%

10	Raghunath Arnab. "Systematic Sampling", Elsevier BV, 2017 Publication	<1%
11	arxiv.org Internet Source	<1%
12	Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei et al. "MassBank: a public repository for sharing mass spectral data for life sciences", Journal of Mass Spectrometry, 2010 Publication	<1%
13	Sanjaya, Christine, May liana, and Agus Widodo. "Revenue Prediction Using Artificial Neural Network", 2010 Second International Conference on Advances in Computing Control and Telecommunication Technologies, 2010. Publication	<1%
14	Iwan Binanto, Harco Leslie Hendric Spits Warnars, Nesti Fronika Sianipar, Bahtiar Saleh Abbas. "LC-MS Analysis: Mini Review Frequently Used Open Source Softwares", 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), 2019 Publication	<1%
15	drops.dagstuhl.de Internet Source	

		<1%
16	innovareacademics.in Internet Source	<1%
17	mafiadoc.com Internet Source	<1%
18	www.mcponline.org Internet Source	<1%
19	www.science.gov Internet Source	<1%
20	Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, Gary Siuzdak. "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification", Analytical Chemistry, 2006 Publication	<1%
21	Pavel Kharyuk, Dmitry Nazarenko, Ivan Oseledets, Igor Rodin, Oleg Shpigun, Andrey Tsitsilin, Mikhail Lavrentyev. "Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task", Scientific Reports, 2018 Publication	<1%
22	hal.archives-ouvertes.fr Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		