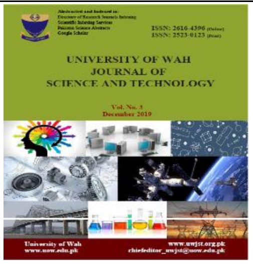
University of Wah
Journal of Science and Technology

www.uow.edu.pk



Abstracted and Indexed in:
Directory of Research Journals Indexing
Scientific Indexing Services
Pakistan Science Abstracts
Google Scholar

ISSN: 2616-4396 (Online)
ISSN: 2523-0123 (Print)

UNIVERSITY OF WAH
JOURNAL OF
SCIENCE AND TECHNOLOGY

Vol. No. 3
December 2019

University of Wah
www.uow.edu.pk

www.uwjst.org.pk
chiefeditor_uwjst@uow.edu.pk

Data Analysis of Educational Websites Using RapidMiner

Anum Akhtar, Saeeda Javaid, Sidra Ejaz and Rabail Shafique Satti

***Abstract*—With an exponential increase in the World Wide Web's size, collecting precise information which satisfies particular requirements is a challenge, especially when time is an entity of great importance leading to demand efficiency. In this paper, RapidMiner is used to collect information from educational websites. Collected data is analyzed and visualizations of data is created to evaluate it.**

***Index Terms*—Data mining, RapidMiner, Web mining, Web pages and Big data**

I. Introduction

THE internet contains almost 1.6 billion websites and 6.08 billion webpages available for use, and this number is continuously increasing [1], [2]. Retrieving relevant information from all this is a challenging task. Since the volume of data to be analyzed is in a significant amount, hence, it is termed as Big data [3]. The primary purpose of Big data is the extraction of useful information from the vast collection and use it for decisions based analysis [4].

For retrieval of this massive amount of data, a process known as data mining is done to summarize it and test the data from different perspectives. Web mining is a part of the large group data mining in which data is extracted from web links, summarized and then processed for required results. In the presented work, web mining techniques are utilized for the collection of data which is then processed for analysis. Section II of paper includes review of previous relevant literature on data mining, web mining and RapidMiner. Section III and IV discusses the architecture of the model and data collection methodology whereas implementation method is discussed in Section V. Section VI presents simulation and result based analysis. Conclusions are drawn in Section VII.

II. Literature Review

The worldwide web is an online network that comprises thousands of creditworthy information for the users. Big data is a phenomenon not restricted to a particular area but applies in multiple fields [5]. Previously used processing applications cannot be used to process vast amount of data with a complex structure [6]. Hence, handling significant amount of data efficiently for processing is a difficult task [7].

In the present age, data is generated by every person in the society and not just by large corporations and administrations [8]. A report by International Data Corporation (IDC) expects that the data volume from 2005 to 2020 to grow from 130 exabytes to a volume of 40,000 exabytes [9].

Data mining is known as a computerized exploration in which huge data is processed. Using data mining, "patterns" can be identified by summarizing and filtering the processed data. Since time is an entity of great importance, all the processing in data mining is done in a short time, which is a great advantage [10].

Oren Etzioni firstly used web mining as a term in [11], where an attempt was made to use "data mining techniques" using websites data for retrieval of useful material from web resources available. It was further pointed in [11] that web resource can be a web document or a web-service.

Web content mining is a classification of web mining in which knowledge is extracted from the actual content of the web, which may include web documents or web pages having information [12]. It can be also described as extending the work that is done by conventional search engines [13].

Text mining is a related field of data mining in which patterns are identified from the vast amount of data [14]. Different patterns may include text classification, text extraction and text clustering. data for text mining is mostly unstructured, which requires some text processing procedures, unlike data mining. Hence, both these fields are not similar.

RapidMiner is a software platform mostly used in data Science. It provides an integrated development

Manuscript received June 15, 2019; accepted October 11, 2019.

A. Akhtar (email: anumakhtar97@gmail.com) is student of Fatima Jinnah Women University, Pakistan.
S. Javaid (email: saeeda2905@gmail.com), S. Ejaz (email: cidraejaz@gmail.com) and R. S. Satti (email: emailrabail@gmail.com) are affiliated with Department of Software Engineering, Fatima Jinnah Women University, Pakistan.
Corresponding author email: anumakhtar97@gmail.com

environment (IDE) for data preparation, predictive analysis, data mining, machine learning, and many others [15].

RapidMiner operators can be used to improve summarization accuracy with sentiment analysis [15]. Text processing and Aylien Text Analysis. Extensions were used in [15] to build two different models for sentiment analysis. In the first's case, an accuracy of 90% was achieved, while for the second model, 96% accuracy was achieved.

In [16], RapidMiner was used to analyze big data. Social media was employed to extract financial news data. However, results produced were not much meaningful as home pages of most social media websites, which were mined to extract financial new data, did not focus on the topic. One factor which affected the results was the Hypertext Markup Language (HTML) code. It was not removed from the data in the model. The similarity of the webpages was low and most of frequently used words were not relevant to the topic.

III. Methodology

RapidMiner is a popular data analytics tool, which is a simulation IDE written in Java language. It is a “free” and "open-source." Formats in which data can be imported includes CVS, Excel, Access, Oracle and several others.

Microsoft Excel 365 is used in presented research for data preparation and transformation. There is an enormous amount of functionality within RapidMiner. RapidMiner's additional extensions are used for text mining web mining, and text processing functionalities. *Fig. 1* shows the conceptual architecture of the model simulated in RapidMiner.

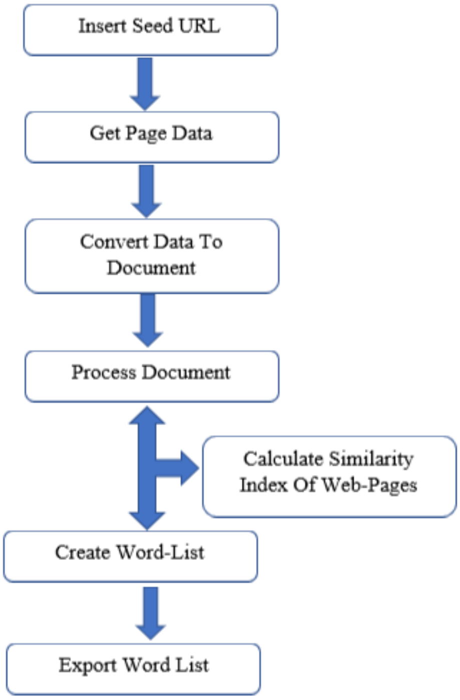


Fig. 1 Model Architecture.

IV. Data Collection

Data is the most crucial component of this research. The data used hereby in research is collected from the website of ten universities web links of Pakistan. The dataset is saved in an excel file and is retrieved at the time of process execution.

Two types of data are taken for this study. In the first case, links of the home page of ten (10) Pakistani universities websites are taken as shown in Table I, and in the second case, links of a sub page in which details of programs offered by the university were described as shown in Table II.

Table I
Home Page Of Universities Websites

Links
http://pu.edu.pk/
http://www.nust.edu.pk/Pages/Default.aspx
http://fjwu.edu.pk/
http://qau.edu.pk/
http://hyd.szabist.edu.pk/
http://nu.edu.pk/
http://www.uop.edu.pk/
http://ue.edu.pk/
http://preston.edu.pk/
http://gcwuf.edu.pk/

Table II
Programs Offered Sub Page Of Universities

Links
http://pu.edu.pk/program/index/Graduate
http://www.nust.edu.pk/Academics/Pages/Undergraduate.aspx
http://fjwu.edu.pk/admission-criteria/graduate-and-under-graduate-program/
http://qau.edu.pk/faculty-of-natural-sciences-2/
http://hyd.szabist.edu.pk/academics.html#undergradute
http://nu.edu.pk/Degree-Programs
http://www.uop.edu.pk/admissions/?q=Undergraduate&r=Undergraduate-Departments
http://ue.edu.pk/beta/programDetail/programlist.aspx?pageID=11
http://preston.edu.pk/courses.php
http://gcwuf.edu.pk/admissions/fall-2018/bs-chart-of-seats/

V. Implementation

The model created in the study includes fundamental RapidMiner Studio operators and extension operators of Web mining and text processing. The procedure that is followed is described in forthcoming subsections.

A. *Step 1: Inserting read excel operator*

Read excel is a standard RapidMiner operator. This operator is used to load data from an excel file. Input given to this operator is the file path of the excel file and the output is data table containing data. The format of the table is such that the rows contain the example and the columns show the characteristics. The name of the attributes are specified by a parameter in the first row of the excel sheet. The data table can contain the rows and columns which are empty and the instructions related to formatting. The table can be positioned at any place in the excel sheet. The input is taken in the form of a file containing data, and the operator converts the file through the port and represent it in the form of a table containing data. In the parameter panel, the import configuration wizard can quickly import the file.

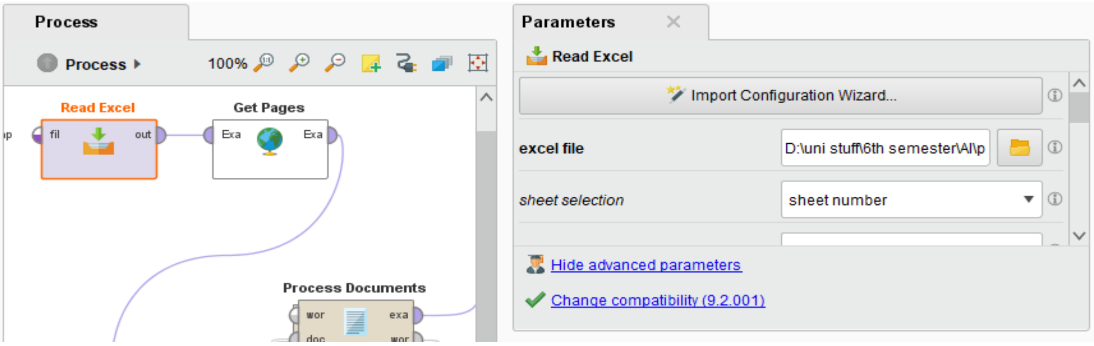
Fig. 2 shows the read excel operator connected with other operators in the simulation on the left side while on the right side, it shows the parameters of the read excel operator.

B. Step 2: Inserting get pages operator

Get pages is an operator used in web mining. In which, those pages are retrieved from the data set whose URLs are encompassed in the data set. The input of the operator is the data table obtained from the read excel operator. A request is send to get the page. The page attribute parameter later on specifies a new attribute to store the requested page. The output of the operator is an example set. The parameter is set in the parameter panel as a link because it has a URL of the pages. The accept cookies checkbox is checked, so it should accept all the cookies. The cookie scope parameter is set as global and the connection and read timeout are set 10,000 in the parameter panel. *Fig. 3* shows the Get pages operator in the IDE on the left side while the parameters of the operator on the right side.

C. Step 3: Inserting data to documents operator

Data to documents is a text processing operator. It creates the documents by using the values present in the data set. These documents are then grouped to generate a document that contains every example of the data. The input to the operator is an example set and the output is the document. *Fig. 4* shows the data to documents operator in the IDE.



Read Excel

Import Configuration Wizard...

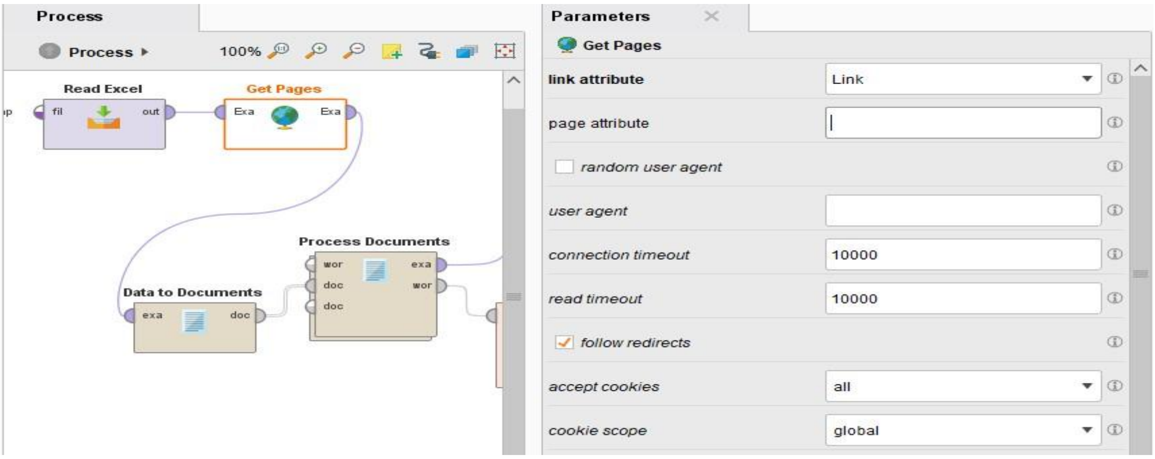
excel file D:\uni stuff\6th semester\AI\p

sheet selection sheet number

Hide advanced parameters

✓ Change compatibility (9.2.001)

Fig. 2. Read excel operator.



Get Pages

link attribute	Link
page attribute	
□ random user agent	
user agent	
connection timeout	10000
read timeout	10000
☑ follow redirects	
accept cookies	all
cookie scope	global

Fig. 3. Get pages operator.

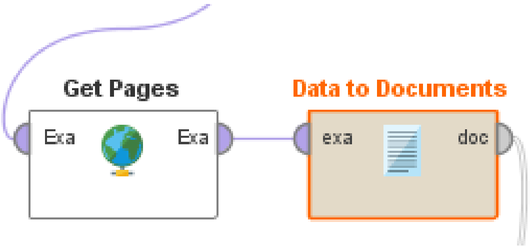


Fig. 4. Data to documents operator.

D. Step 4: Inserting process documents operator

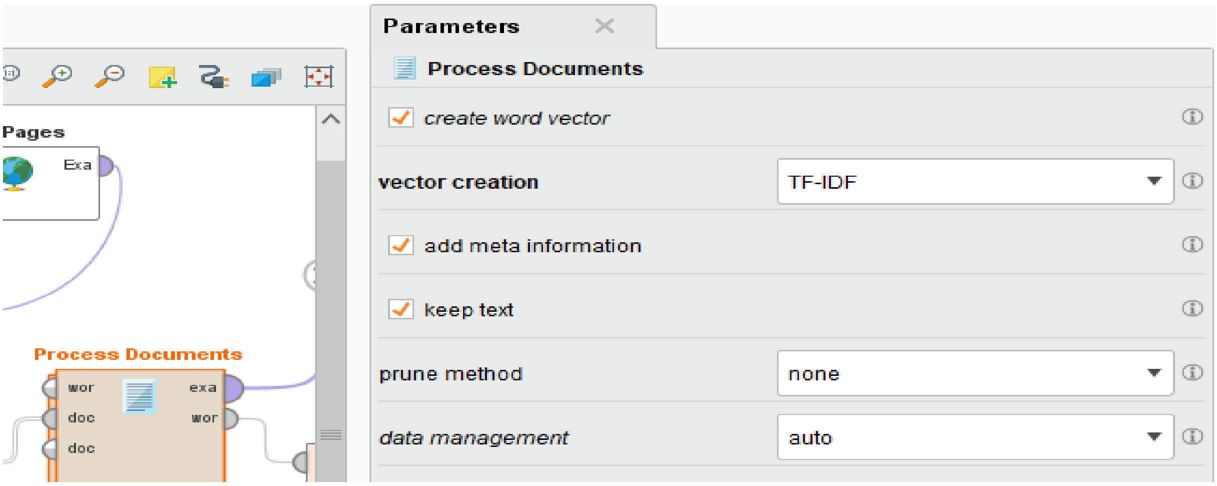
It is a text processing operator. In which, a word vector is generated from a text object. The term vector is generated by using a single text object. As a result, the example set contains a single example and it becomes easier to apply a model on a single text. The input to the operator is word list and documents. The word list is provided through word list port and the documents or collection of documents are provided through the document port. The output of the operator is a wordlist, and an example set. In the parameter panel, the created word vector is checked to produce a vector to represents the document by using the tokens. The vector creation parameter is set to term frequency-inverse document frequency (TF-IDF) for the scheme of word vector creation. The data management is set to auto option to represent the data internally. This is a nested operator; hence, document processing operators are inserted inside it. Process is completed from step 5 to step 7. *Fig. 5* shows the process documents operator.

E. Step 5: Inserting extract content operator

Extract content is an operator used in web mining, using which content from the HTML document is extracted. The text blocks are taken out of the document to avoid single word. They contain many words. The input to the operator is a document and output from the operator is also a document. They are received and delivered through the respective document ports. The extract content parameter in the parameter's panel is checked to extract the content. The minimum block length is set 4, i.e., a block can comprise four characters. All other parameters are checked for this operator. The extract content operator is shown in *Fig. 6.*

F. Step 6: Inserting tokenize operator

This operator is used in text processing. Primary purpose of this operator is to tokenize a document. The text is fragmented into a set of tokens. The points for separation of tokens can be specified by using parameters. The input is a document that is sent through the document port and output is a document that is delivered through the document port. The mode parameter in the parameter's panel is set non-letters. This means that the operator tokenizes according to the non-letters. *Fig. 7* shows the tokenize operator in the IDE.



Parameters

Process Documents

☑ create word vector

vector creation: TF-IDF

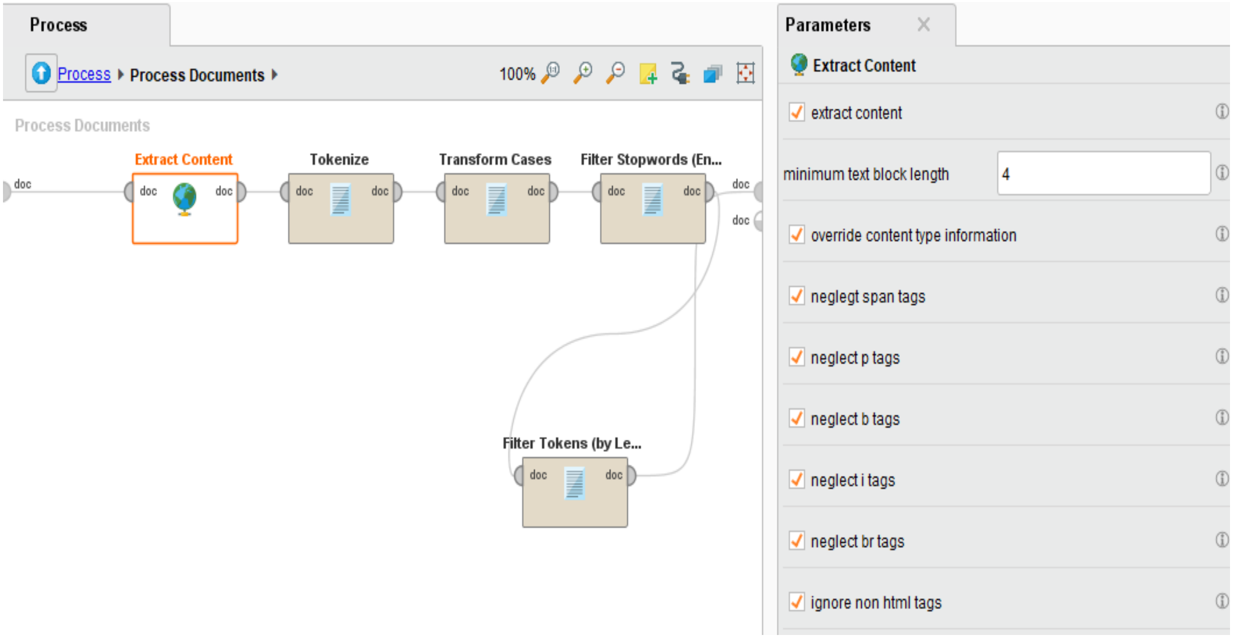
☑ add meta information

☑ keep text

prune method: none

data management: auto

Fig. 5. Process documents operator.



Parameters

Extract Content

☑ extract content

minimum text block length: 4

☑ override content type information

☑ neglegt span tags

☑ neglect p tags

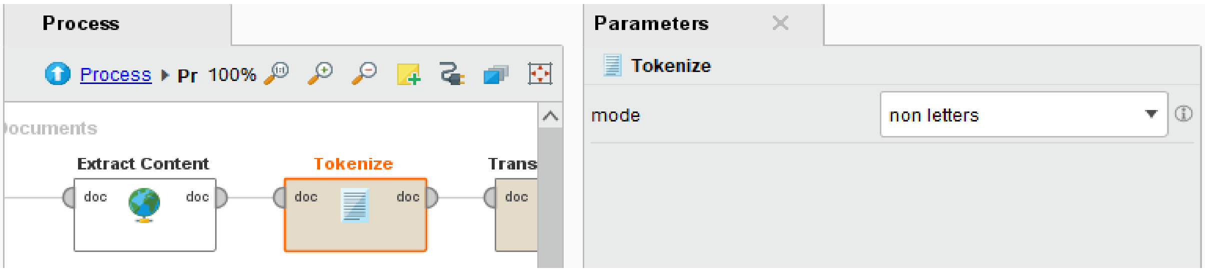
☑ neglect b tags

☑ neglect i tags

☑ neglect br tags

☑ ignore non html tags

Fig. 6. Extract content operator.



Tokenize

mode: non letters

Fig. 7. Tokenize operator.

G. Step 7: Inserting a transform cases operator

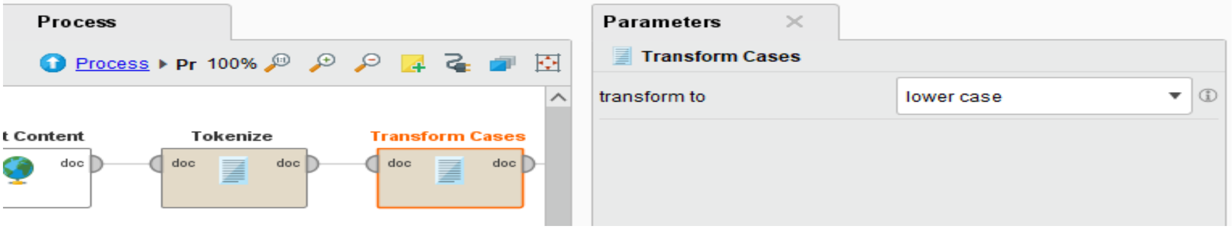
It is an operator used in text processing. It transforms the cases of characters in a document. The characters can be transformed to either lower or uppercase characters as required. The input of the operator is a document and output got from the operator is a document. In the parameters panel, the "transform to" parameter is set to lowercase. *Fig. 8* shows the transform cases operator.

H. Step 8: Inserting filter tokens (by length) operator

This operator is used in text processing and comes after tokenize operator. The document containing the tokens delivered from the output document port of the tokenize operator is received at the input port of this operator. The primary function of this operator is to filter out the tokens based on their length. The input of the operator is a document and the output is also a document. *Fig. 9* shows the filter tokens (by length) operator in the IDE.

I. Step 9: Inserting multiply operator

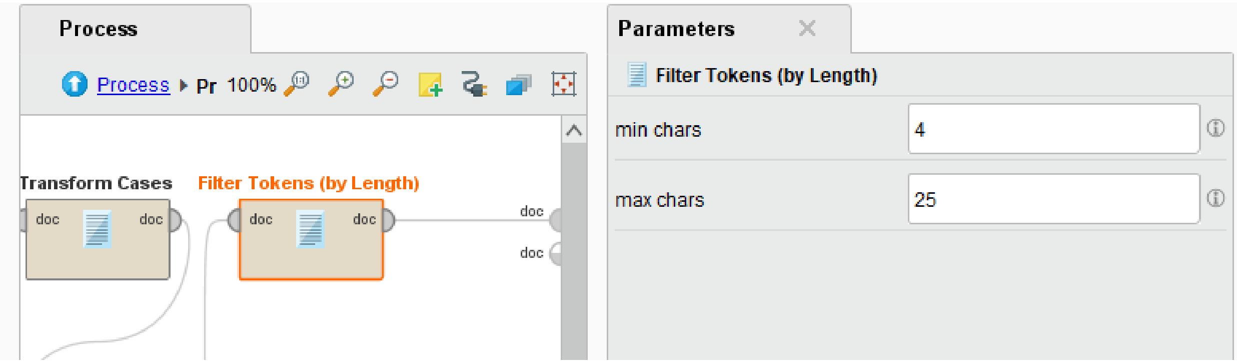
It is an operator used in RapidMiner. The input to the operator is a RapidMiner object which is transferred to it through an input port. It makes several copies of the object and sends it to the output. The alteration in one copy does not affect the other copies because every port generates its copy. It helps in better processing. *Fig. 10* shows the multiply operator in the IDE.



Transform Cases

transform to: lower case

Fig. 8. Transform cases operator.



Filter Tokens (by Length)

min chars: 4

max chars: 25

Fig. 9. Filter tokens (by length) operator.

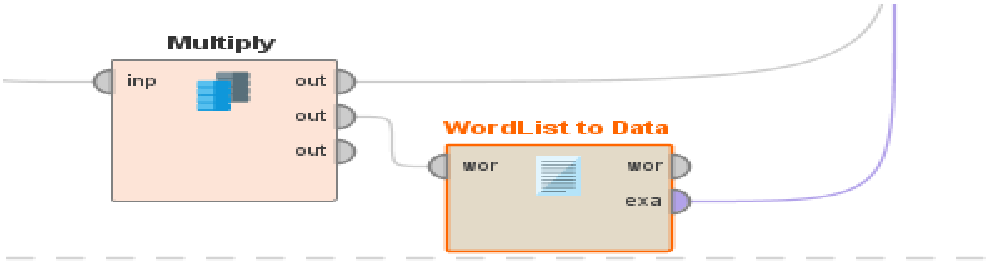


Fig. 10. Multiply operator.

J. Step 10: Inserting wordlist to data operator

This operator is used in text processing for transforming the Wordlist into data. The data set has separate rows and columns. There is a separate row for each word and its characteristics. The data set also contains information about how many times a word occurs in a document and classes. This operator is helpful in a way that it filters out a wordlist before the next processing. The input of the operator is the wordlist and the output is an example set and the wordlist through their respective input and output ports. *Fig. 11* shows the Wordlist to data operator in the IDE.

K. Step 11: Inserting write excel operator

It is an operator of RapidMiner. As name shows, it is used to write an example set into an excel sheet. It generates the excel files. The example set can contain missing data represented by blank cells. The attributes of example set are represented in the first row. The read excel operator is used to load the files in the RapidMiner, which are written by the "write excel" operator. The input of the operator is an example set. The output of the operator is a file. The file port sends the file object for the operators, which might require file objects. In the parameter panel, the file format is set to xlsx. *Fig. 12* shows the write excel operator.

L. Step 12: Inserting data to similarity operator

Data to similarity is an operator of RapidMiner. In this operator, the similarity of the examples of the data table is cross-checked with the other examples of the same data table. It is used to check the similarity between examples. The input to the operator is the data table and the output of the operator is similarity and example set. The similarity port sends the object which include similarity between examples of the data table. The example set output port sends the data table with no modification. In parameter panel, the measure type is set to numerical measures so that the computed results for similarity is represented in numeric. The numerical measure parameter is set to cosine similarity. *Fig. 13* shows the "data to similarity" operator.

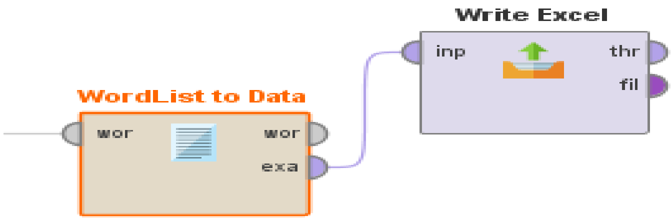
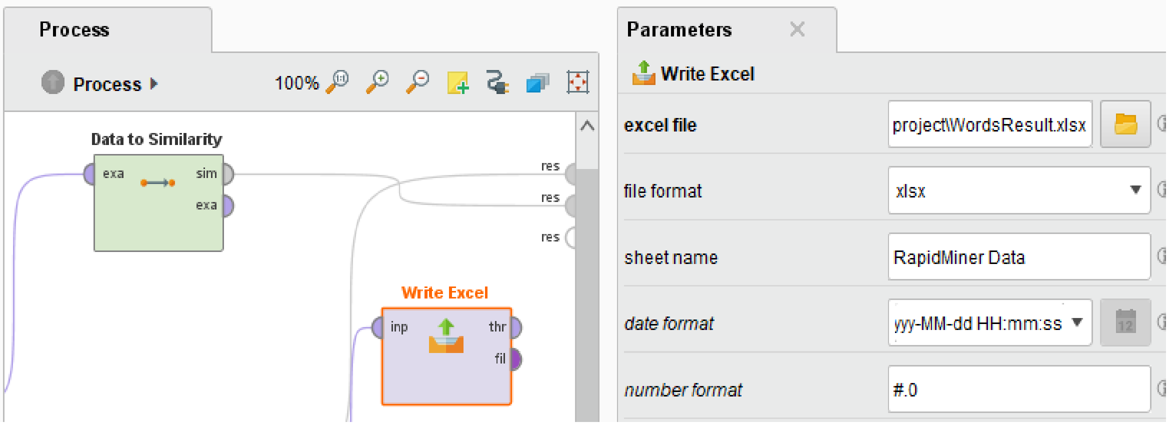


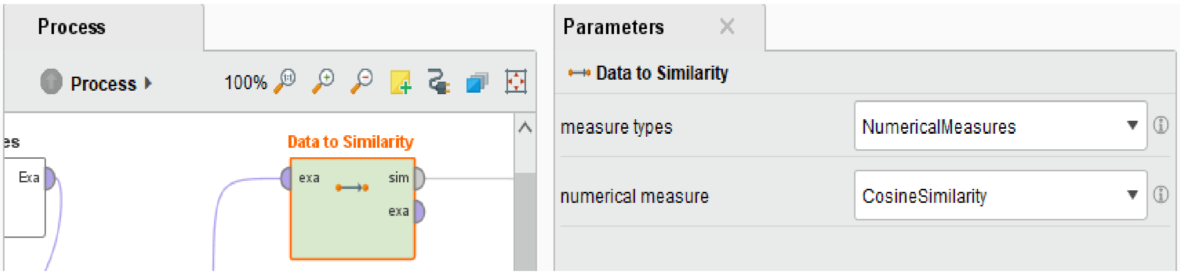
Fig. 11. Wordlist to data operator.



Write Excel

excel file	project\WordsResult.xlsx
file format	xlsx
sheet name	RapidMiner Data
date format	yyy-MM-dd HH:mm:ss
number format	#.0

Fig. 12. Write excel operator.



Data to Similarity

measure types	NumericalMeasures
numerical measure	CosineSimilarity

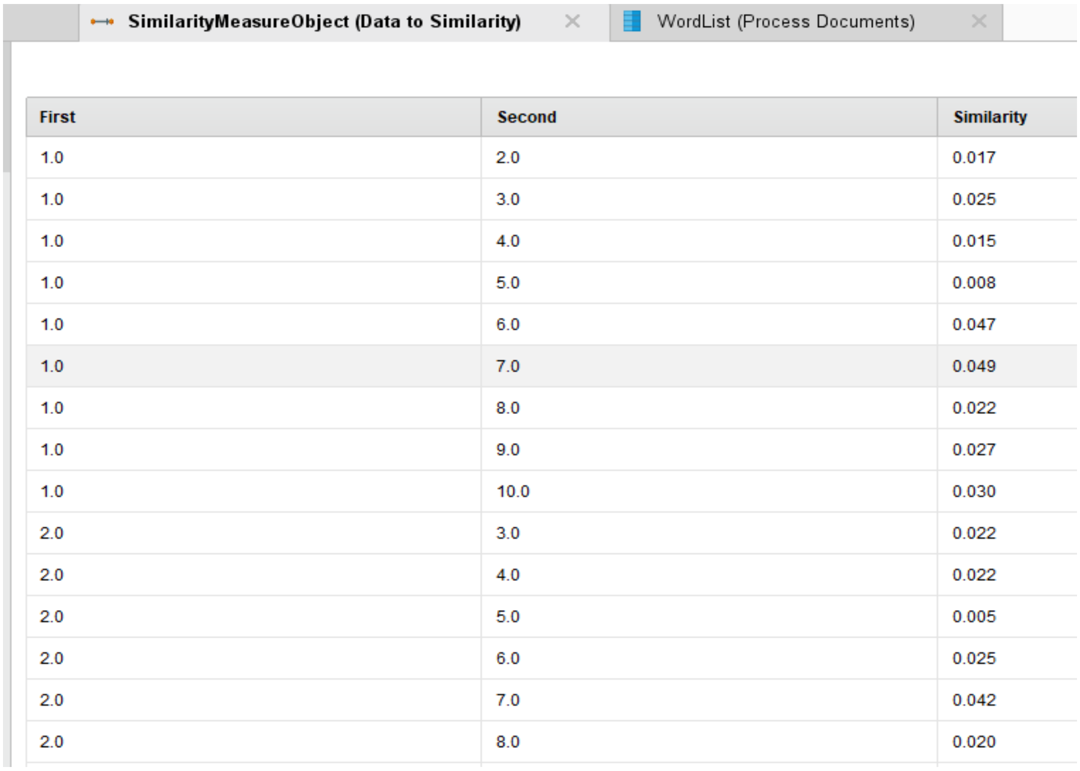
Fig. 13. Data to similarity operator.

VI. Results And Discussion

Two types of results are obtained from the study; similarity index of websites and the wordlist visualization and summarization of sample of websites.

A. *Case 1*

The similarity index of a sample set is shown in the *Figs. 14-21*. In all the websites of universities, there must be some similarity. The similarity of every webpage is calculated compared to others as in all cases. For visualization purposes, a graph in the circular form depicts all webpages as nodes and their edges show the similarity between them. *Fig. 14* shows the similarity index between the selected websites in case 1, while *Fig. 15* shows a graph of similarity index where websites are marked at nodes and weights of edges are their similarity index.


SimilarityMeasureObject (Data to Similarity)
WordList (Process Documents)

First	Second	Similarity
1.0	2.0	0.017
1.0	3.0	0.025
1.0	4.0	0.015
1.0	5.0	0.008
1.0	6.0	0.047
1.0	7.0	0.049
1.0	8.0	0.022
1.0	9.0	0.027
1.0	10.0	0.030
2.0	3.0	0.022
2.0	4.0	0.022
2.0	5.0	0.005
2.0	6.0	0.025
2.0	7.0	0.042
2.0	8.0	0.020

Fig. 14. Similarity Index Case 1.

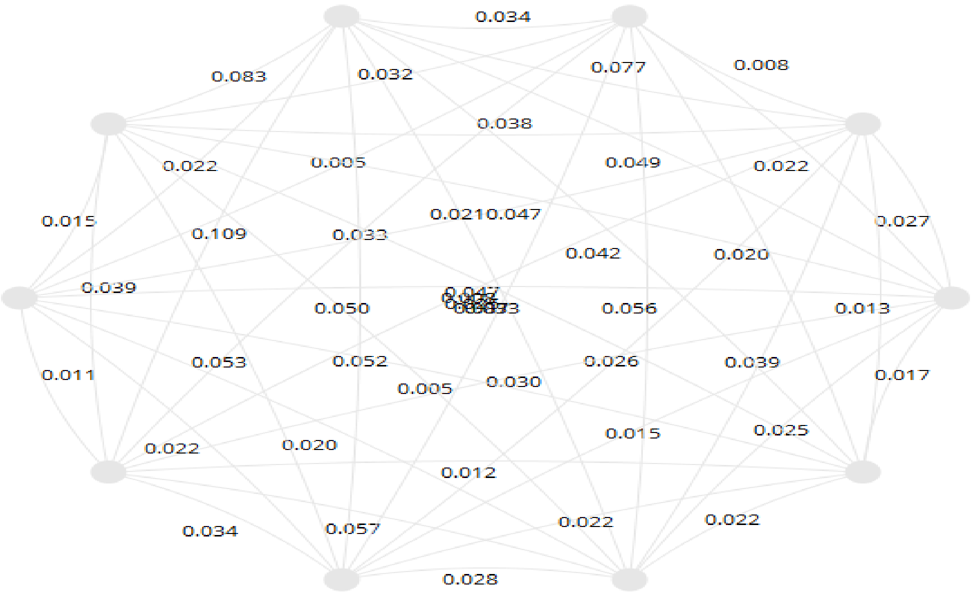
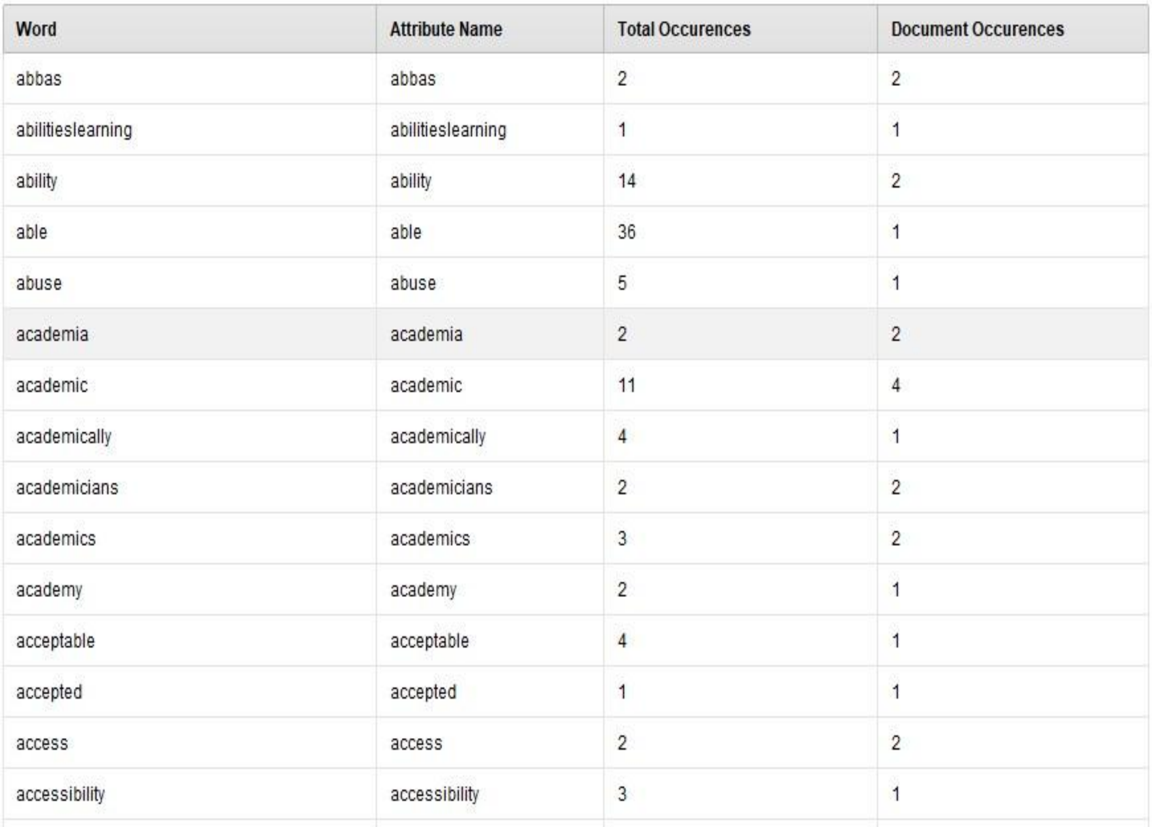


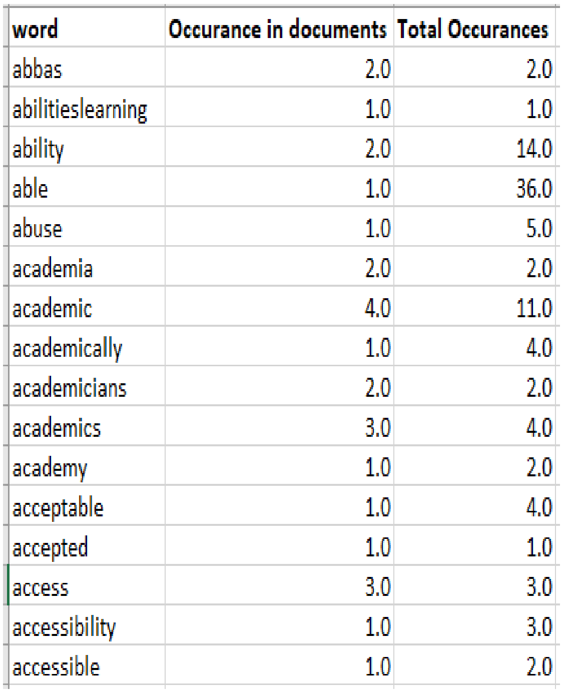
Fig. 15. Similarity index Graph.

Fig. 18 is a summarized pie chart of the wordlist. It shows that topic related words like education, university, and science are present in abundance in the content. *Fig. 19* is a Radar chart visualization of the result.



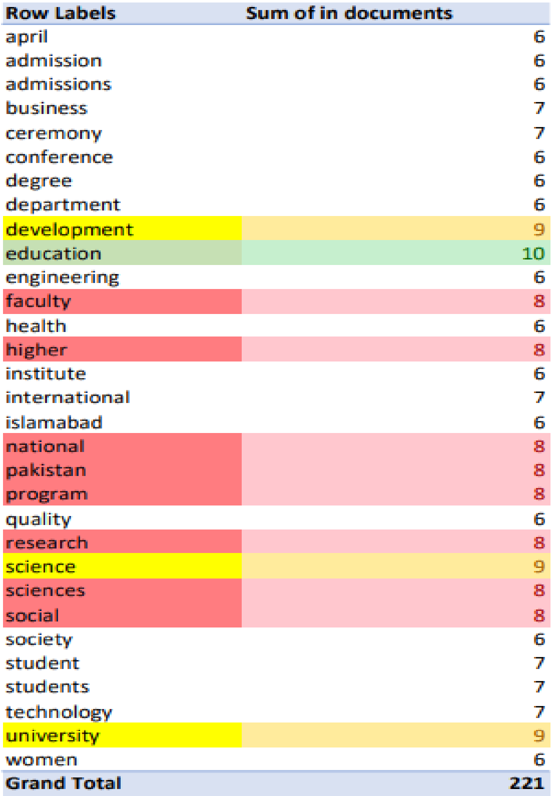
Word	Attribute Name	Total Occurences	Document Occurences
abbas	abbas	2	2
abilitieslearning	abilitieslearning	1	1
ability	ability	14	2
able	able	36	1
abuse	abuse	5	1
academia	academia	2	2
academic	academic	11	4
academically	academically	4	1
academicians	academicians	2	2
academics	academics	3	2
academy	academy	2	1
acceptable	acceptable	4	1
accepted	accepted	1	1
access	access	2	2
accessibility	accessibility	3	1

Fig. 16. Wordlist.



word	Occurance in documents	Total Occurances
abbas	2.0	2.0
abilitieslearning	1.0	1.0
ability	2.0	14.0
able	1.0	36.0
abuse	1.0	5.0
academia	2.0	2.0
academic	4.0	11.0
academically	1.0	4.0
academicians	2.0	2.0
academics	3.0	4.0
academy	1.0	2.0
acceptable	1.0	4.0
accepted	1.0	1.0
access	3.0	3.0
accessibility	1.0	3.0
accessible	1.0	2.0

Fig. 17. List in excel.



Row Labels	Sum of in documents
april	6
admission	6
admissions	6
business	7
ceremony	7
conference	6
degree	6
department	6
development	9
education	10
engineering	6
faculty	8
health	6
higher	8
institute	6
international	7
islamabad	6
national	8
pakistan	8
program	8
quality	6
research	8
science	9
sciences	8
social	8
society	6
student	7
students	7
technology	7
university	9
women	6
Grand Total	**221**

Fig. 18. Pivot table.

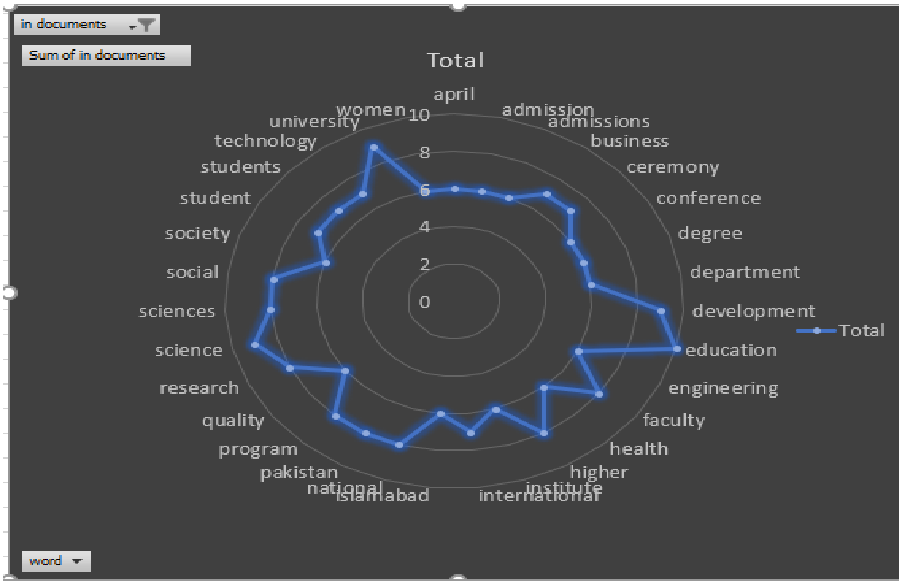
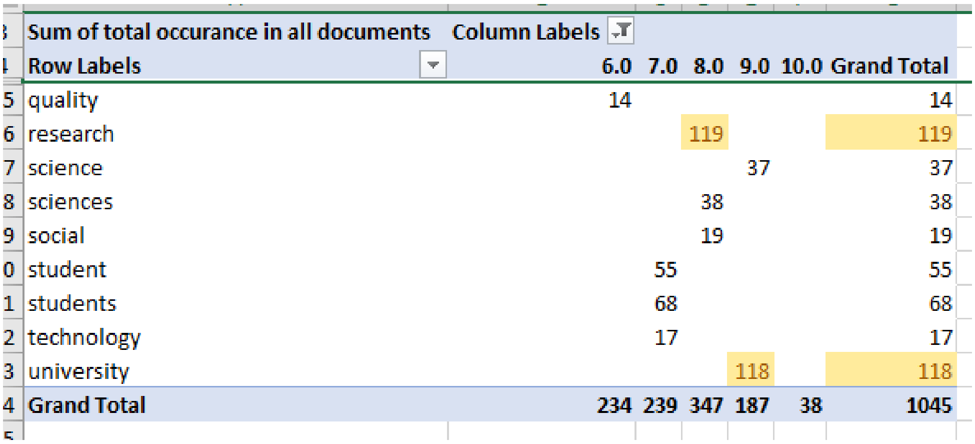


Fig. 19. Radar chart of pivot table.

Fig. 20 below shows a pivot chart that summarizes the result of wordlist. It depicts that words like research and university appeared in 8 to 9 websites, respectively. It shows the relevancy of content to the topic. *Fig. 21* is the bar graph visualization of results.



Sum of total occurance in all documents	Column Labels					
Row Labels	6.0	7.0	8.0	9.0	10.0	Grand Total
quality	14					14
research			119			119
science				37		37
sciences			38			38
social			19			19
student		55				55
students		68				68
technology		17				17
university				118		118
Grand Total	234	239	347	187	38	1045

Fig. 20. Pivot table.

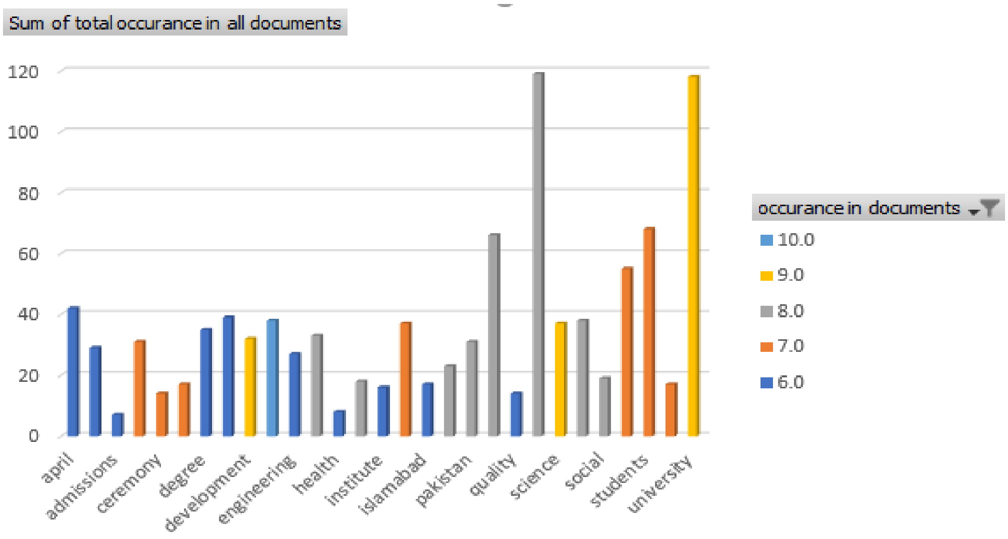
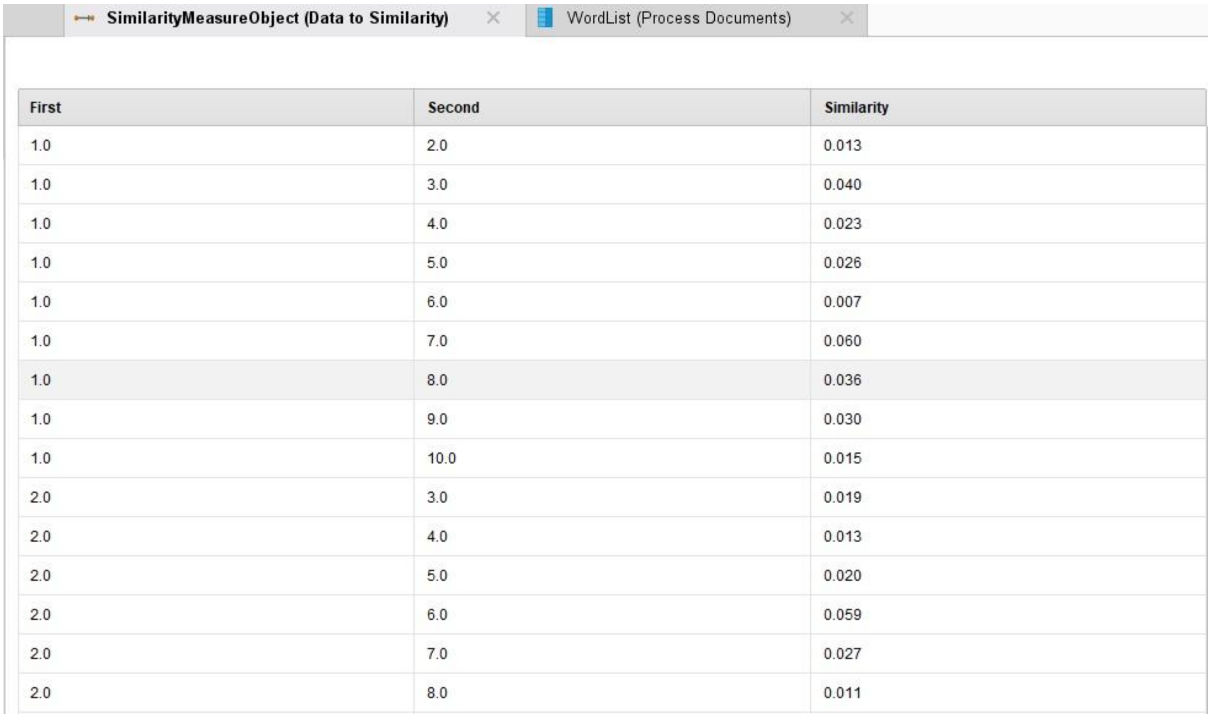


Fig. 21. Bar graph of pivot table.

B. Case 2

The similarity index of a sample set is shown in the *Figs. 22-29*. As all the websites are of universities, there must be some similarity between the websites. The similarity of every webpage is calculated compared to others for all of them. For visualization purposes, a graph in the circular form depicts all webpages as nodes, and their edges show the similarity between them. *Fig. 22* below shows the similarity index between the selected websites in case 2, while *Fig. 23* shows a graph of similarity index where websites are the nodes and weights of edges are the similarity index.


SimilarityMeasureObject (Data to Similarity)
WordList (Process Documents)

First	Second	Similarity
1.0	2.0	0.013
1.0	3.0	0.040
1.0	4.0	0.023
1.0	5.0	0.026
1.0	6.0	0.007
1.0	7.0	0.060
1.0	8.0	0.036
1.0	9.0	0.030
1.0	10.0	0.015
2.0	3.0	0.019
2.0	4.0	0.013
2.0	5.0	0.020
2.0	6.0	0.059
2.0	7.0	0.027
2.0	8.0	0.011

Fig. 22. Similarity Index of case 2.

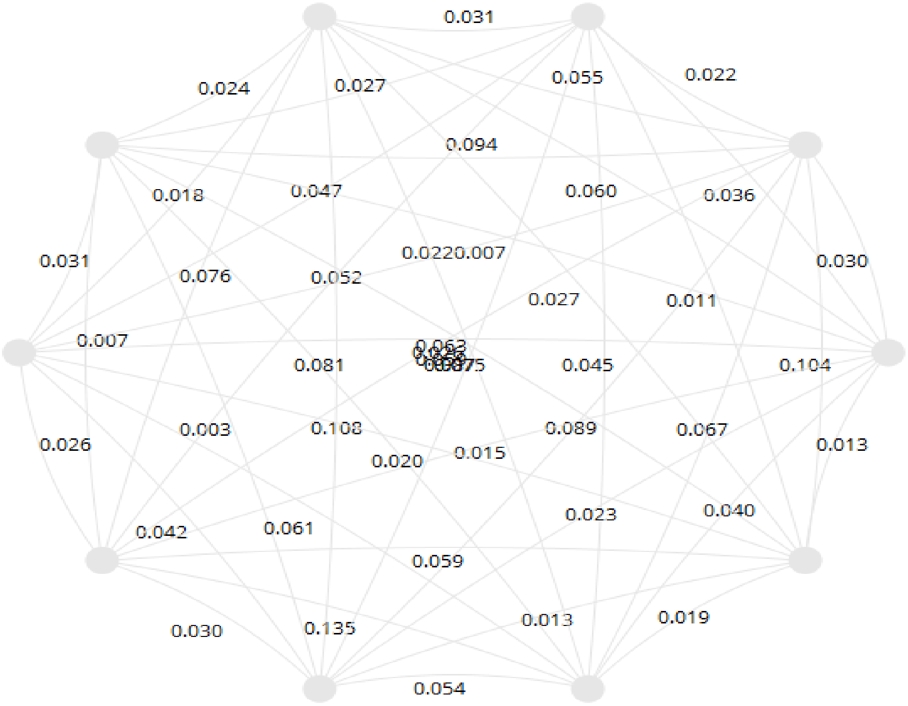
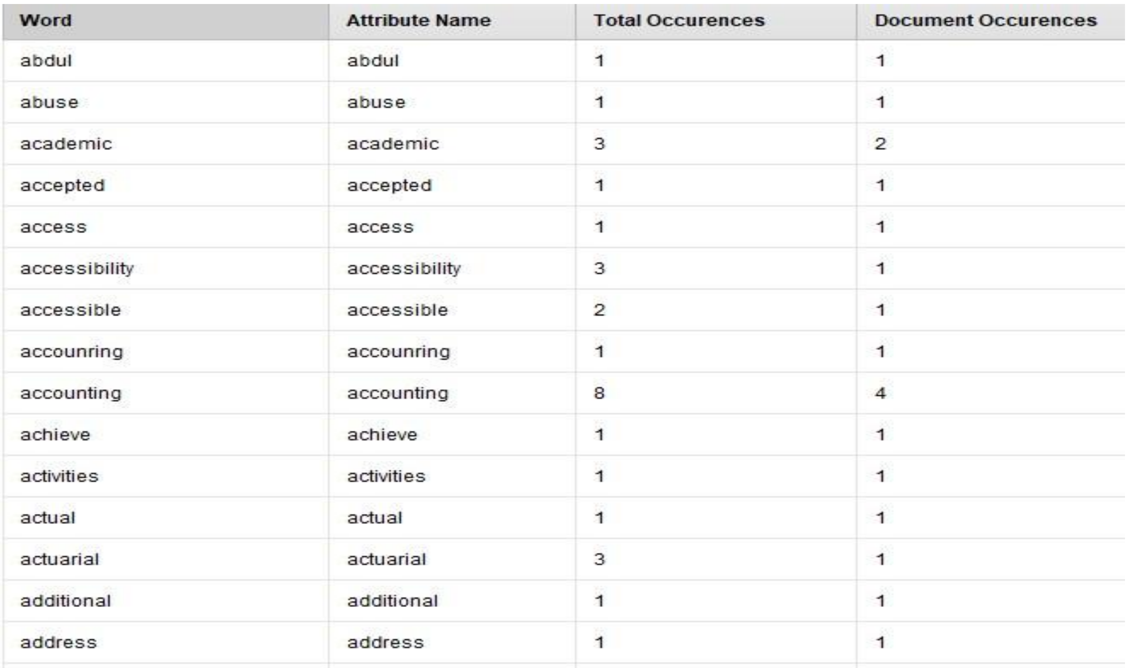


Fig. 23. Graph of similarity index.

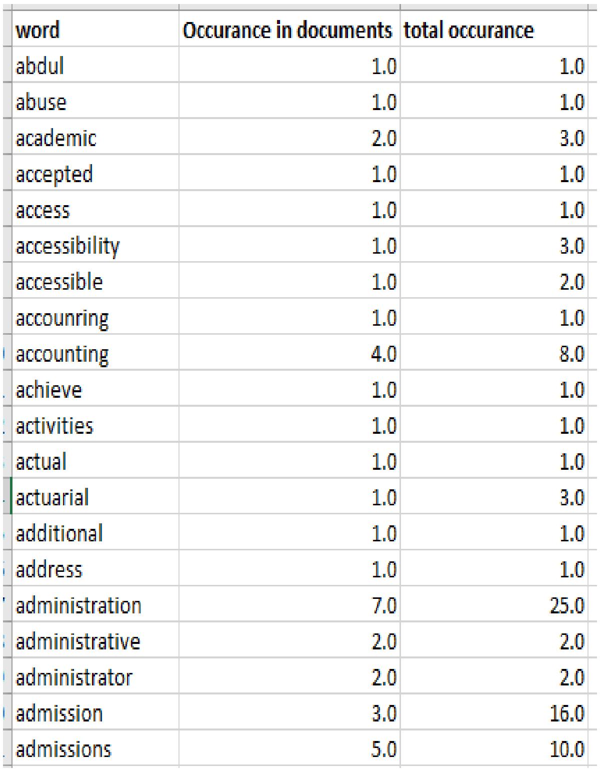
A sample of the generated Wordlist is shown below in *Fig. 24*. It contains the keyword, attributes, total number of occurrences in all the web data and occurrence in websites.

The wordlist is exported to an excel spreadsheet, and the pivot table of the data is extracted, which summarizes the data. *Fig. 25* shows a portion of the exported wordlist in excel. *Fig 26* shows a summarized pivot chart of the wordlist. It shows that topic related words like the computer, management, university, and science are present in abundance in the content. *Fig. 27* below is a Radar chart visualization of the result. *Fig. 28* shows a pivot chart that summarizes result word list. It depicts that words like science, engineering, and the program appeared in 9,7 and 4 websites, respectively. Result shows the relevancy of content to the topic. *Fig. 29* shows the bar graph visualization of presented results.



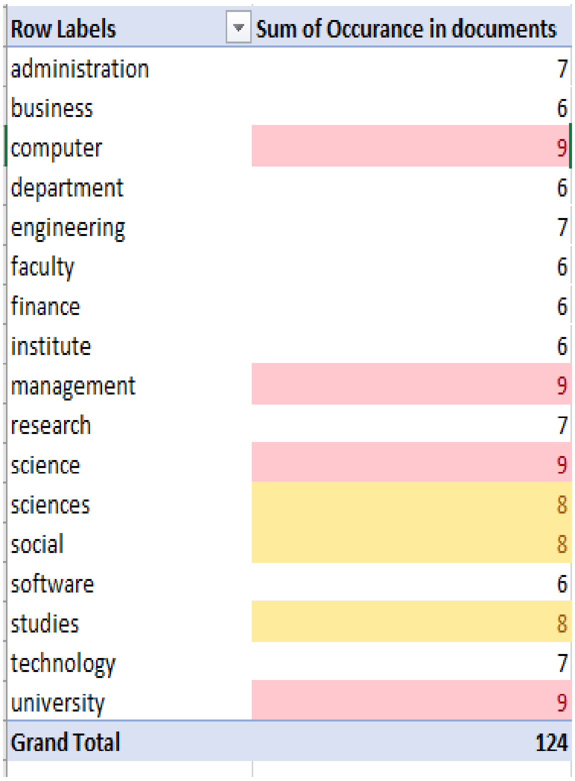
Word	Attribute Name	Total Occurences	Document Occurences
abdul	abdul	1	1
abuse	abuse	1	1
academic	academic	3	2
accepted	accepted	1	1
access	access	1	1
accessibility	accessibility	3	1
accessible	accessible	2	1
accounring	accounring	1	1
accounting	accounting	8	4
achieve	achieve	1	1
activities	activities	1	1
actual	actual	1	1
actuarial	actuarial	3	1
additional	additional	1	1
address	address	1	1

Fig. 24.Wordlist.



word	Occurance in documents	total occurance
abdul	1.0	1.0
abuse	1.0	1.0
academic	2.0	3.0
accepted	1.0	1.0
access	1.0	1.0
accessibility	1.0	3.0
accessible	1.0	2.0
accounring	1.0	1.0
accounting	4.0	8.0
achieve	1.0	1.0
activities	1.0	1.0
actual	1.0	1.0
actuarial	1.0	3.0
additional	1.0	1.0
address	1.0	1.0
administration	7.0	25.0
administrative	2.0	2.0
administrator	2.0	2.0
admission	3.0	16.0
admissions	5.0	10.0

Fig. 25. Wordlist in excel.



Row Labels	Sum of Occurance in documents
administration	7
business	6
computer	9
department	6
engineering	7
faculty	6
finance	6
institute	6
management	9
research	7
science	9
sciences	8
social	8
software	6
studies	8
technology	7
university	9
Grand Total	**124**

Fig. 26. Pivot table.

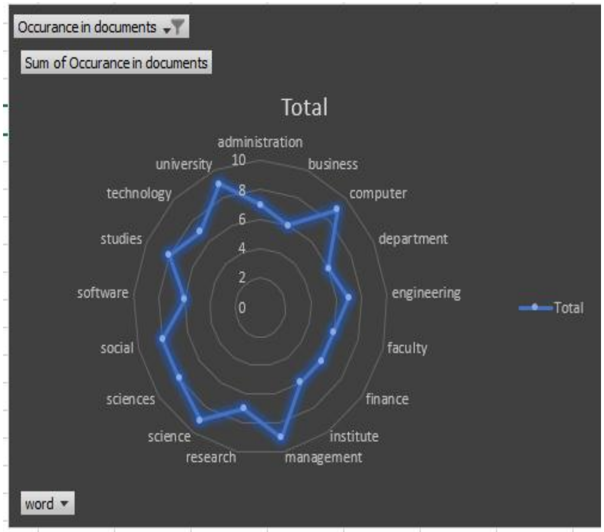
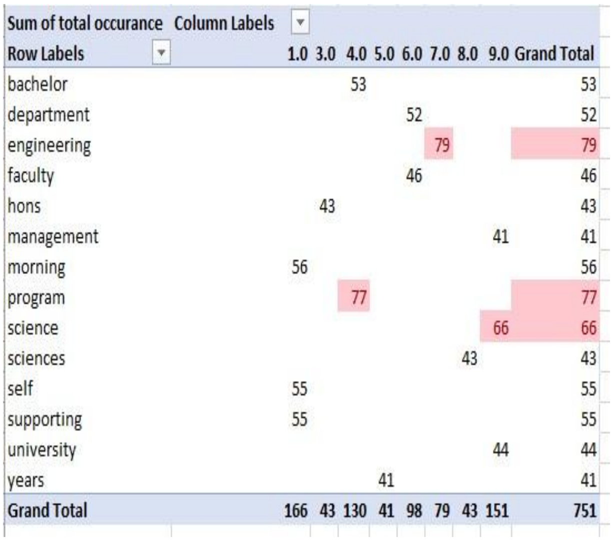


Fig. 27. Radar chart of pivot table.



Sum of total occurance	Column Labels								
Row Labels	1.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	Grand Total
bachelor			53						53
department					52				52
engineering						79			79
faculty					46				46
hons		43							43
management								41	41
morning	56								56
program			77						77
science								66	66
sciences							43		43
self	55								55
supporting	55								55
university								44	44
years				41					41
Grand Total	166	43	130	41	98	79	43	151	751

Fig. 28. Pivot table.

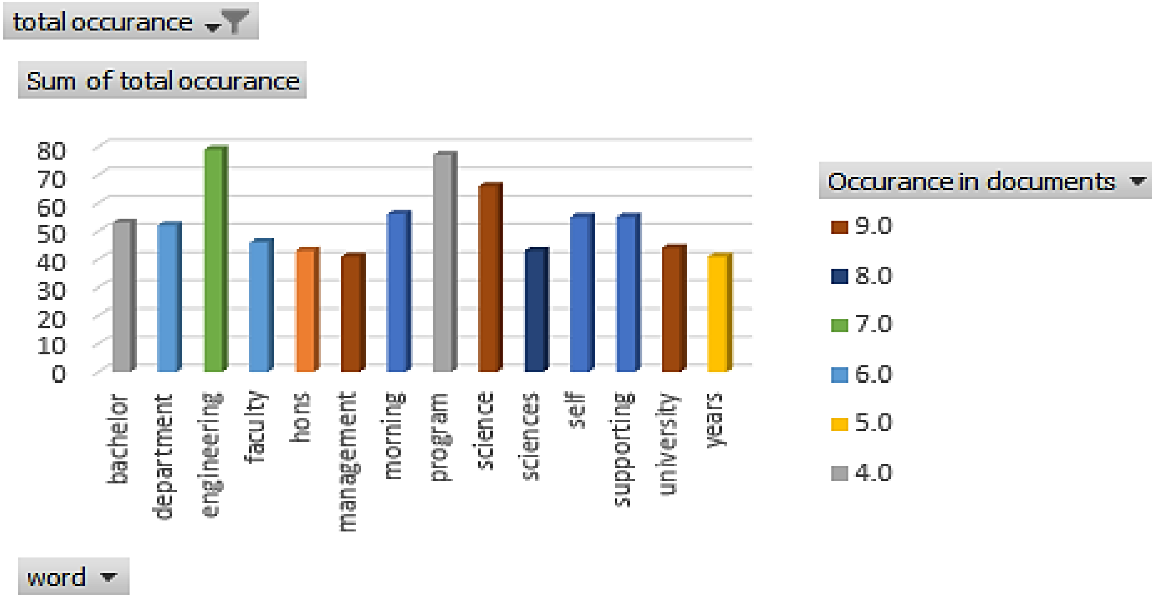


Fig. 29 Bar Graph of pivot table.

VII. Conclusion

In the current world, Big data has been created which is beyond the capacity of "database software tools" to process, analyse, and create visualizations. Hence techniques of data mining are beneficial in this era. The constructed model shows satisfactory results with a high percentage of keywords related to topic of university websites. Visualizations and data summary extracted is very much relevant to the presented theme. The similarity index is low, but there is some level of similarity between the webpages. Similarity index results are satisfactory as the contents of websites cannot be very much similar. While comparing presented cases using RapidMiner, it is concluded that utilization of a specific sub-page of a website is a better option instead of focusing on the home page. Result shows that the websites homepage does not concentrate on the theme as much as the sub-pages.

References

[1] M. Kunde. The Size of the Worldwide Web. Available online at https://www.worldwidewebsize.com/ on June 15, 2019.

[2] G. Deyan. How Many Websites are there in 2019 - Active and Inactive. Available online at https://techjury.net/blog/how-many-websites-are-there/ on June 15, 2019.

[3] R. Gupta, S. Gupta and A. Singhal. Big data: overview, International Journal of Computer Trends and Technology, Vol. 9 (5), pp. 266-268, 2014.

[4] National Research Council. Frontiers in Massive Data Analysis, The National Academies Press, Washington DC, 2013.

[5] H. U. Buhl, M. Roglinger, D. K. F. Moser and J. Heidemann. Big data, Business and information Systems Engineering, Vol. 5(2), pp. 65–69, 2013.

[6] A. A. Cardenas, P. K. Manadhata and S. P. Rajan. Big data analytics for Security, IEEE Security and Privacy, Vol. 6, pp. 74–76, 2013.

[7] C. P. Chen and C. Y. Zhang. Data intensive applications, challenges, techniques and technologies, a survey on big data, Information Sciences, Vol. 275, pp. 314–347, 2014.

[8] A. Elragal and M. Haddara. Big Data Analytics, Text Mining Based Literature Analysis, NOKOBIT, 2014.

[9] J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows and biggest growth in the far east, IDC iView: IDC Analyze the future, Vol. 2007, pp. 1-16, 2012.
[10] R. Gupta. Journey from data mining to web mining to big data, International Journal of Computer Trends and Technology, Vol. 10(1), pp. 18-20, 2014.
[11] O. Etzioni. The world wide web: quagmire or gold mine, Communications of the ACM, Vol. 39(11), pp 65-68, 1996.
[12] S. B. Joshi and S. D. Mundhe. Web mining and qualities of a website design to be evaluated for customer browsing behavior a review, International Journal of Computer Applications Technology and Research, Vol. 6(6), pp. 269-272, 2017.
[13] H. Dunham, Margaret and Sridhar. Data Mining Introductory and Advanced Topics, Pearson Education India, 2006.
[14] S. A Salloum, M. Alemran, A. A. Monem and K. Shaalan, Intelligent Natural Language Processing: Trends and Applications, Springer, 2018.
[15] A. F. Alsaqer and S. Sasi. Movie review summarization and sentiment analysis using rapidminer, International Conference on Networks and Advances in Computational Technologies, 2017, pp. 329-335.
[16] J. Lane and H. J. Kim. Big data: Web crawling and analysing financial news using rapid miner, International Journal of Business Information Systems, Vol. 19(1), pp. 41–57, 2015.

Anum Akhtar is a student in Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan. She worked as an intern in Data Warehouse, BI and Analytics Department at Ufone, Islamabad. Her areas of interest are data science, big data and web development.

Saeeda Javaid is a student in Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan. Her areas of interest are artificial intelligence and web development.

Sidra Ejaz is working as Assistant Professor in the Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan. She has done her MS and Bachelors in software engineering in 2007 and 2010 respectively. She is currently pursuing her Ph.D. from EME College, NUST. Her area of research is healthcare data analytics.

Rabail Shafique Satti is working as Assistant Professor in the Department of Computer Science, Fatima Jinnah Women University, Rawalpindi, Pakistan. She has done her MS in software engineering in 2014. She is currently pursuing her Ph.D. from Capital University of Science and Technology, Islamabad. Her areas of interest are Computer Networks, Artificial Intelligence, and Information Security.