IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 10, October 2019

Automated E-Learning Platform using Data Mining Techniques

Prajeeth Kumar M.J[1], Gopalakrishnan.T[2]

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India[1,2]

Abstract: The increase in the amount of information available online has made it a difficult job to search for unique documents and relevant information. This in turn also creates a problem to find the correct required piece of information with a lot of time spent in browsing and searching. And we come up with a faster and more accurate way to make training easier in this process. Today, with the upcoming e-learning trends, the system not only aims to provide efficient searches but also provides the desired information to a consolidated material and provides various tools such as useful links and videos to make our system a reliable and comprehensive source for the user's needs. A process series, namely data cleaning, tokenization, and frequency calculation and log-likelihood function, is performed on all articles in this method after collecting the data of the article. In contrast to measuring the byte structure, it considers the capacity of Natural Language so that more accuracy can be obtained during the measure of similarity. During the estimation of text frequency, it performs pre-processing and inverse report frequency. Pre-processing is performed using a standard set of stop words that also provides almost accurate measurements. The following framework makes it possible to use the different data mining techniques to pre-process a report before uploading it and to identify the main topic covered by the post. Later on, searching we get the article whose match for feature vector of the search query is the highest; this is another step into the world of e-learning.

Keywords: Data Cleaning, e-Learning, Tokenization, Natural Language

I. INTRODUCTION

Since its inception, the Internet has undergone exponential growth, and this development has produced many problems. Consumers find it hard to handle content as well as test the validity of reports or find a single file to help them get the data they want. Today, with the e-learning trends, it is important that the information is available in a clean and condensed form and that more time is spent on understanding concepts instead of wasting time on various searches.

A. *Aim*

The main purpose of the system is to simplify the way learning is done and to help improve distance learning, etc. It is kept in mind that one puts more energy into researching or learning concepts than searching in today's advancing world through heaps of data available.

B. *Objectives*

The following targets were suggested to achieve the above-mentioned goal:

- A fully functional web app that provides students with easy learning.
- On the same page, the platform will take input from the pdf and provide correct links.
- A research material aggregation option should also be provided and similarity tests should be maintained.

II. OVERVIEW OF THE PROPOSED SYSTEM

The proposed system first by parsing the pdf collects the articles. After collecting the article data, a system sequence is performed on all posts, namely data cleaning, tokenization, and frequency calculation and log-likelihood. The RV Co-efficient is then applied to find articles with a similar profile. Eventually, the articles are grouped based on factors of closeness or patterns of the article's most important words. In comparison to measuring the byte structure, it considers the ability of Natural Language so that more precision can be obtained during the measure of similarity. During the calculation of text frequency, it performs pre-processing and inverse document frequency. Pre-processing is done using a standard set of stop words that also provides almost accurate measurements.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 10, October 2019

A. *Architecture*

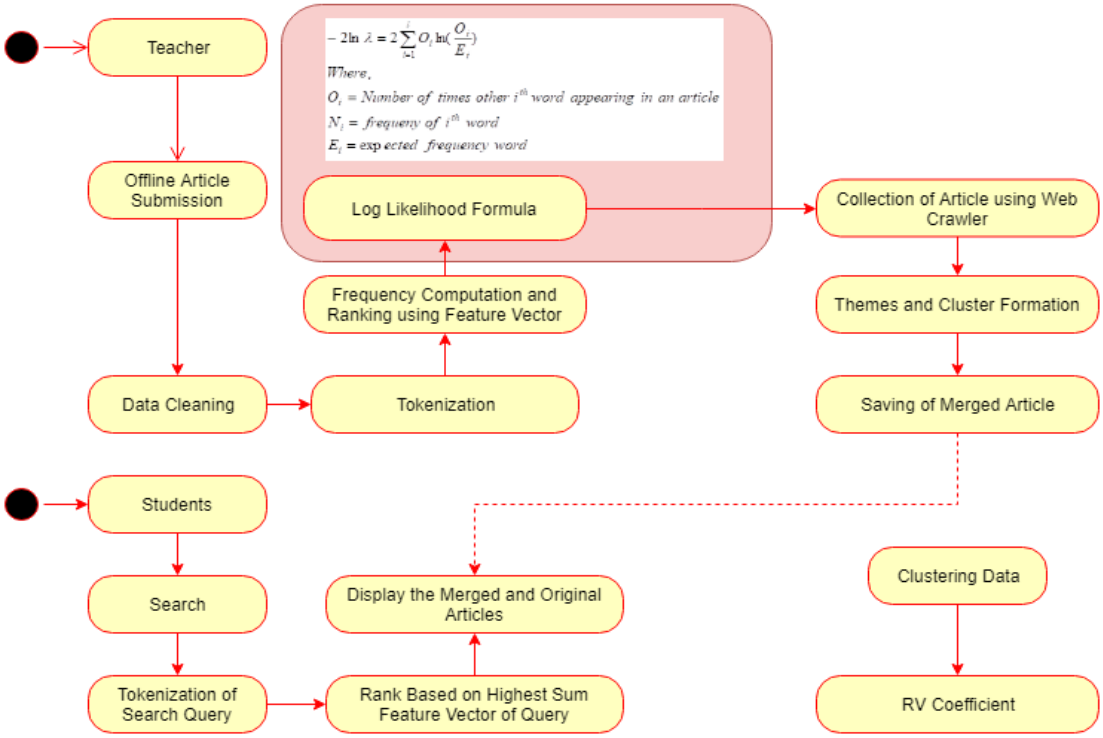

Fig. 1 Architecture

B. *Validation Modules*

- Registration: This module is responsible for allowing users to enroll in the system in order to identify the user as a teacher / student and to give the client the correct view. It also takes specifics of the student age and classification
- Login: This module is responsible for authenticating the end user and providing access to different functions depending on the type of user being a regular user or an admin user.
- Offline Article Submission: This is unique to teachers in order to enable them to submit articles on different topics with a short brief on their material and encourage the educator to provide personalized notes to students.

C. *Pre-Processing*

- Collection of Articles by Web Crawler: When a search is performed along with the already existing articles a query is made to the web crawler to provide the top searches that are also incorporated with our PDF and even among the top searches the following algorithm runs to increase the data quality and make it more accurate.

The Data Cleaning algorithm is responsible for removing stop words. These are the set of words that have no particular meaning. Keywords have been identified by the data mining forum. Stop words are words that are filtered out before or after natural language data (text) has been processed. There is not a definite list of stop words used by all tools and not always used such a filter.. The list of stop words used in the algorithm is as follows:

"a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your"

- Tokenization: Tokenization is a process by which the data got after data cleaning is converted into a set of words known as tokens. Each of the tokens can be represented as Token Id, Token Name and Article ID.

Token Id	Article Id	Token Name

- Frequency Computation: This is a method that performs the frequency measurement. The frequency is determined for each of the posts. Frequency is the number of times in the document that a token appears. The frequency matrix is computed in the following format:

Freq ID	Article ID	Token Name	Frequency

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 10, October 2019

- Feature Vector and Ranking: This module is responsible for calculating the number of articles found in the term, and then finally IDFT Feature Vector.

D. *Further Grouping & Similarity Check*

- Log-Likelihood Function Computation: The log-likelihood is determined for each of the tokens in the articles and is given by the equation. The expected Frequency is computed in the following way for all the words of articles.
- RV Coefficient Similarity: This module helps to check the article similarity thus helps in providing non-redundant information
- Clustering Process: This process helps to assign a group number to each article based on the results of the likelihood, we can set a threshold value so we can assign it to our requirements and form separate groups.
- Themes: These are terms that have the highest log likelihood; these are the main topics in the article of which we will also get the suggested connections.

III. PROPOSED SYSTEM ANALYSIS AND DESIGN

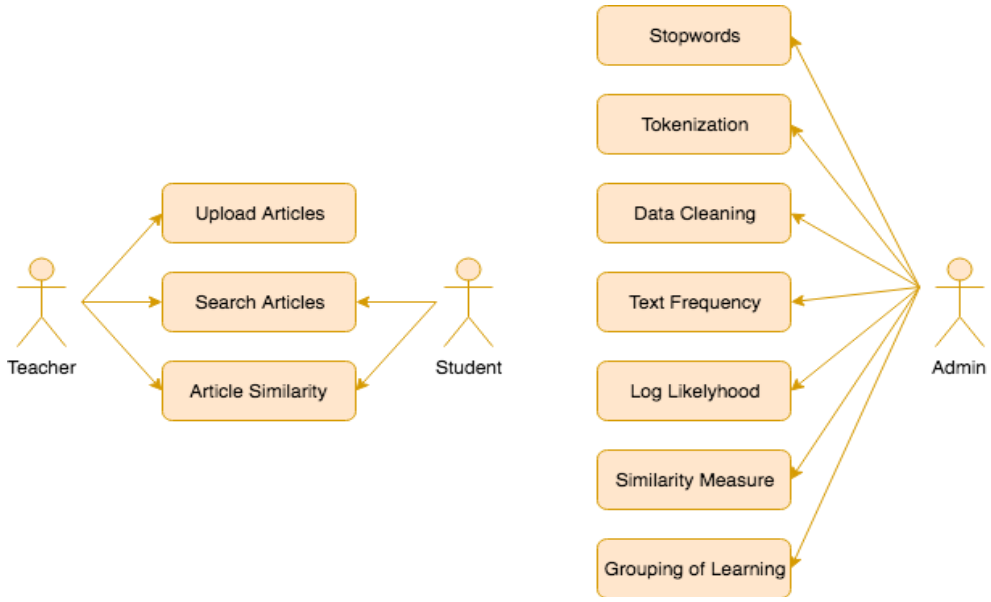

Fig. 2 Users and their actions

A. *Article Upload*

The article module is responsible for the storage of articles. Article name and article description acts as an input.

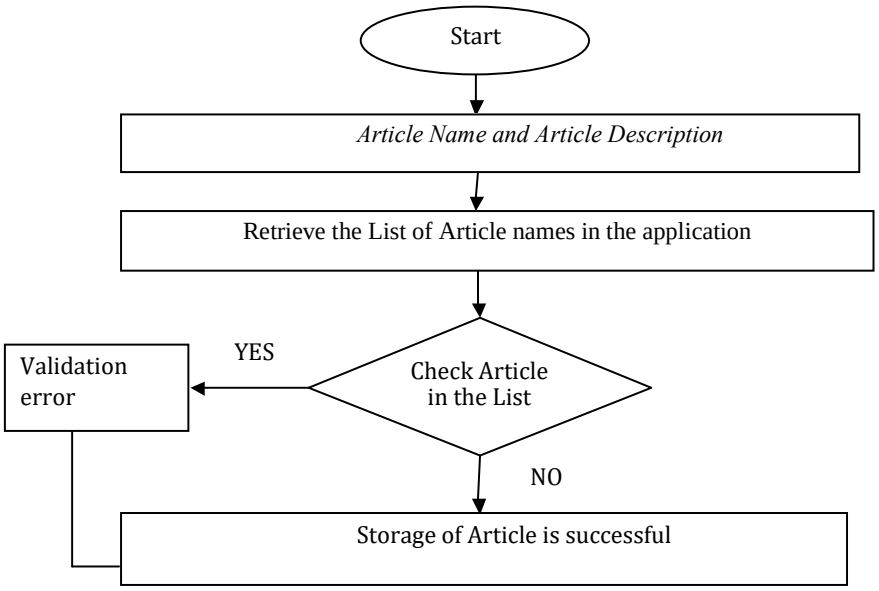

Fig. 3: Article upload

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 10, October 2019

B. *Data Cleaning Algorithm*

The process of removing the stop words from the articles is referred to as data cleaning.

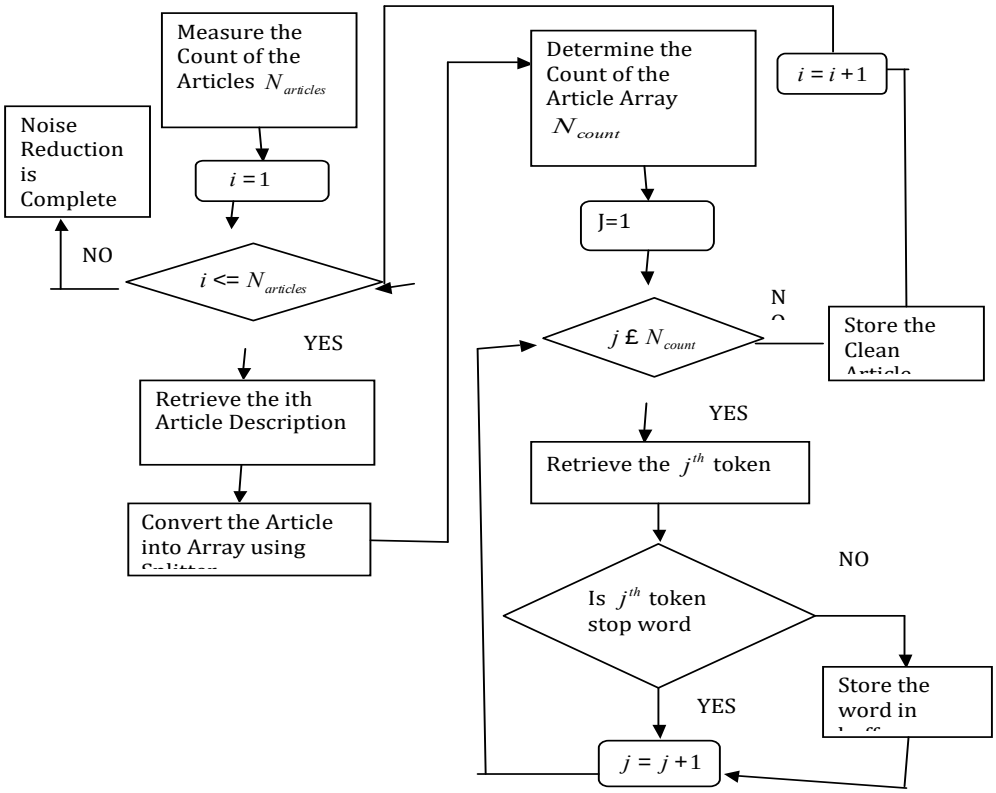

Fig. 4: Data cleaning

C. *Tokenization*

Tokenization is a way of transforming clean data into a series of words called tokens. You may represent each token as Token ID, Token Name and Article ID.

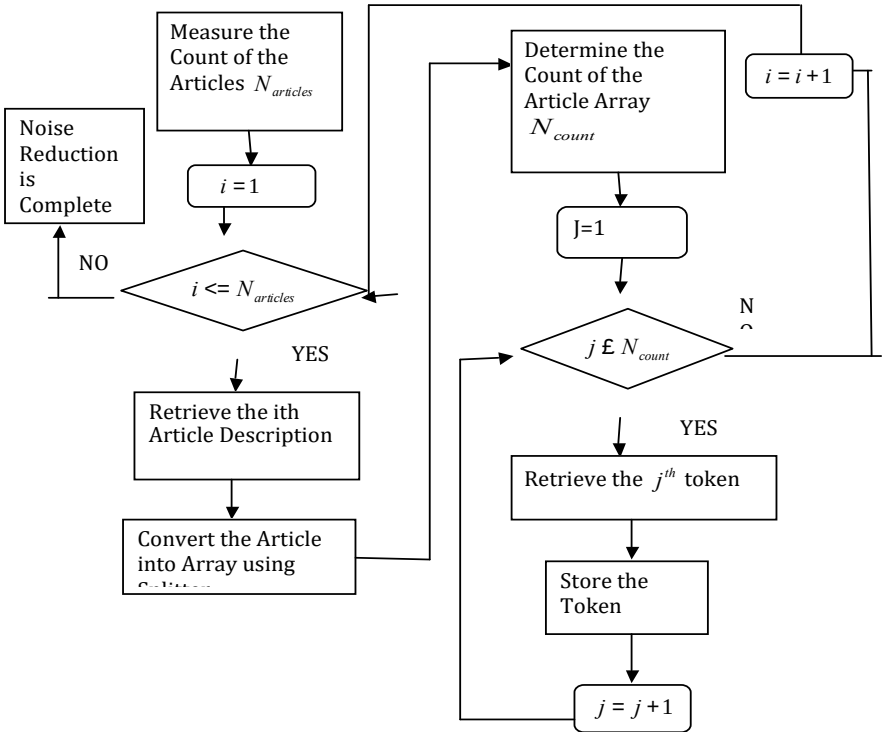

Fig. 5: Tokenization

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 10, October 2019

D. *Frequency Computation*

This is a process in which the frequency computation is performed. For each of the articles, the frequency is computed.

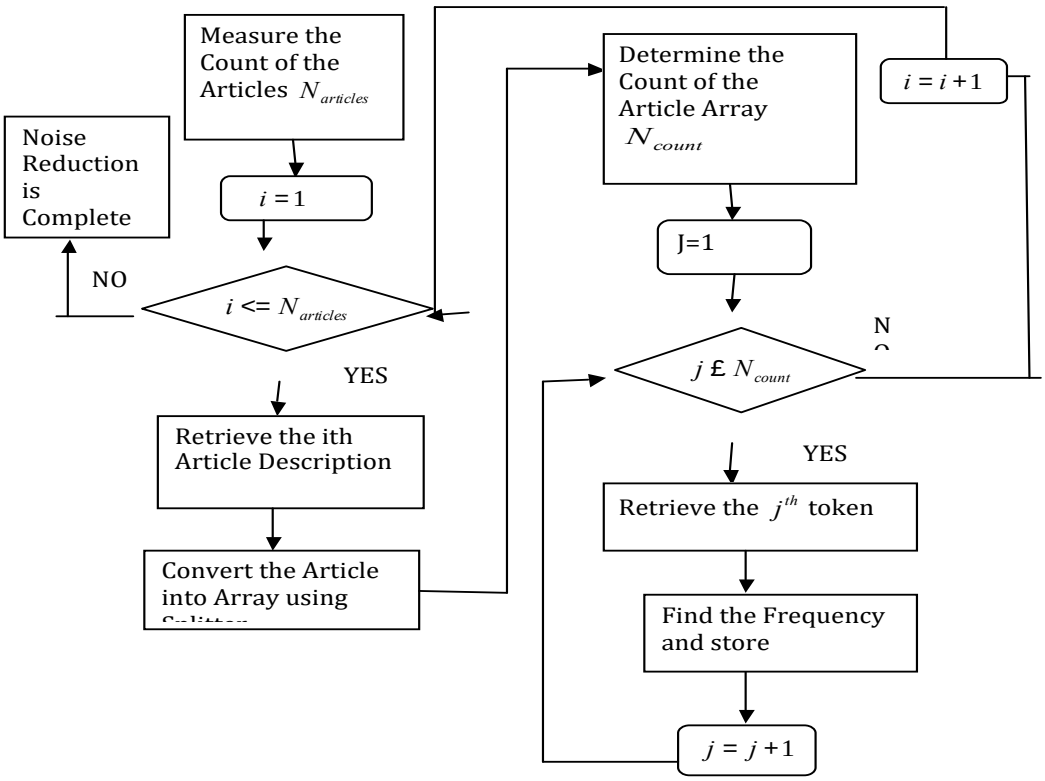

Fig.6: Frequency Computation

E. *Log-Likelihood Algorithm*

The log-likelihood is determined for each of the tokens in the articles and is given by the equation. The expected Frequency is computed in the following way for all the words of articles.

$$E_i = \frac{N_i \sum_{i=1}^{N_{wordsother}} O_i}{\sum_{i=1}^{N_{wordsin\,article}} N_i}$$

$$-2\ln\lambda = 2\sum_{i=1}^{l} O_i \ln(\frac{O_i}{E_i})$$

Where,

O_i = *Number of times other* i^{th} *word appearing in an article*

N_i = *frequeny of* i^{th} *word*

E_i = *expected frequency word*

F. *RV Coefficient Similarity*

- Select the two articles to be compared
- Find the List of unique words in Article A1
- Find the List of unique words in Article A2
- Find the intersection set between the two lists
- Compute the mean of the values of frequency for the words in article1
- Compute the mean of the values of frequency for the words in article 2
- Compute the standard deviation for the words in article1
- Compute the standard deviation for the words in article2

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 10, October 2019

- Compute the RV Coefficient value for the 2 articles using

$$\varphi = \frac{1}{N_{values} - 1} \frac{\sum_{i=1}^{N_{values}} (x_i - \bar{x})(y_i - \bar{y})}{\bar{\bar{x}}\,\bar{\bar{y}}}$$

where,

$x_i = word\ \ frequency\ of\ \ i^{th}\ word\ \ in\ article\ x$

$y_i = word\ \ frequency\ of\ \ i^{th}\ \ word\ \ in\ article\ y$

$\bar{x} = mean\ of\ \ words\ \ for\ article\ x$

$\bar{y} = mean\ of\ \ words\ for\ article\ y$

$\bar{\bar{x}} = standard\ deviation\ of\ words\ \ for\ article\ x$

$\bar{\bar{y}} = standard\ deviation\ of\ words\ \ for\ article\ y$

$N_{values} = Number\ of\ \ unique\ values$

G. *Clustering & Themes*

Clustering Process: Clustering is a process in which each of the papers are calculated & then clustered into one cluster if the value is greater than a certain threshold. The cycle will be repeated until all papers have been scanned & grouped.

Themes: These are set of words whose frequency is highest and are the main topics of various pages.

H. *StopWord Creation*

Stopword Creation Module is responsible for creating the stopword. If the stopword exists then validation error is shown otherwise stopword is created.

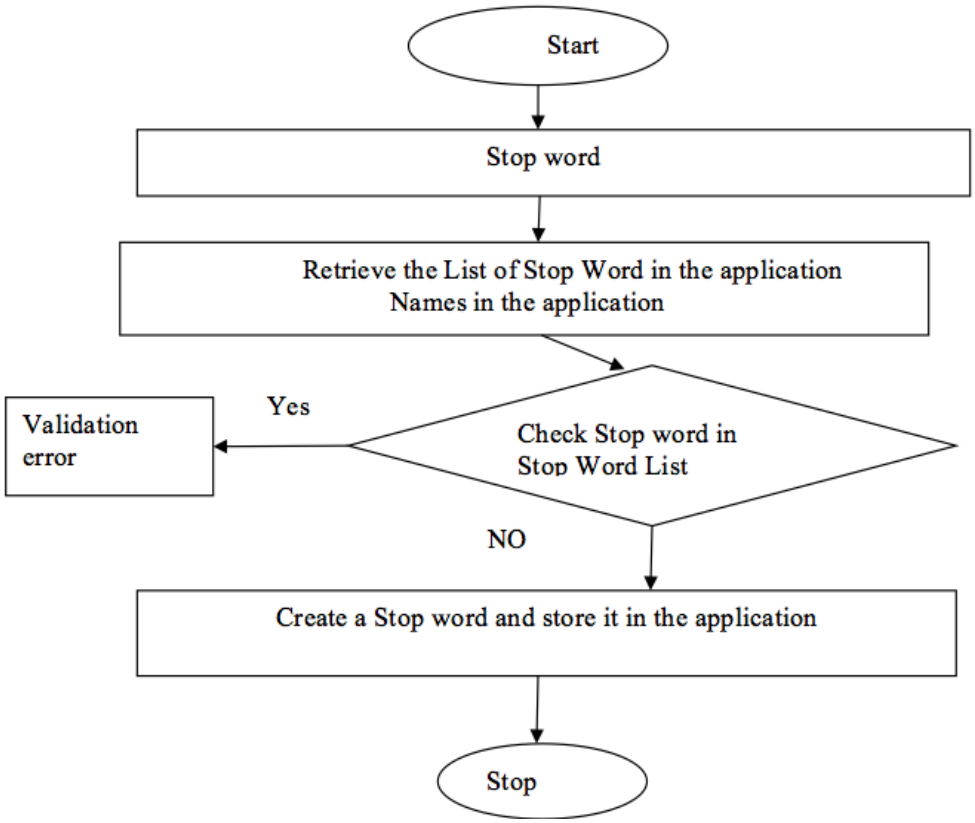

Fig. 7: Stopword creation

IV. RESULTS & DISCUSSIONS

The application was able to get the top articles for a search faster and faster with the article's main topics having more page-wise links for further studies, as well as a similarity check algorithm that was faster than the traditional convert to a binary algorithm. In this paper, a better solution for e-learning platforms is enabled and the summarization of information is achieved in a more apt way.

V. CONCLUSION

The current approach is a conventional model where we waste time looking for relevant information, this software will save time and provide the world of e-learning with more efficient and accurate methods. The software can be further built in the age of smartphones by creating a user-friendly interface. This can become a robotic teacher's e-learning bot and improve distance learning from tools and means to education for rural areas or cities. This application also uses a unique similarity control algorithm that can serve multiple purposes in itself. The software can therefore be an enormous benefit for the world of data mining, analysis and even the field of e-learning if it is explored further.

REFERENCES

[1]. I. Sommerville, "Software Engineering", 9th Ed., Pearson Education, 2010.
[2]. F. Brooks, "No Silver Bullet – Essence and Accidents of Software Engineering", Proceedings of the IFIP Tenth World Computing Conference, pp 1069-1076, 1986.
[3]. P. Jackson, I. Moulinier, "Natural Languge Processing for Online Applications, Text Retrieval, Extraction and Categorization", John Benjamins Publishing Company, 2002.
[4]. C. Aggarwal, C. Zhai, "A Survey of Text Clustering Algorithms", in Mining Text Data, Springer, pp77-129, 2012.
[5]. A. Huang, "Similarity Measures for Text Document Clustering", in New Zealand Computer Science Research Student Conference, pp 49-56,April 2008.
[6]. M. Thelwall, "Bibliometrics to webometrics", Journal of Information Science, 34(4), pp 605-621, 2008.
[7]. G. Keshaval, M. Gowda, "ACM transaction on information systems (1989-2006): A bibliometric study", Information Studies, 14(4), pp 223-234, 2008.
[8]. M. Lee, T. Chen, "Revealing research themes and trends in knowledge management from 1995 to 2010", Knowledge Based Systems 28, pp 47-58, 2012.
[9]. J. Ma, W. Xu, Y. Sun, E. Turban, S. Wang, O. Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans, Vol. 42, No. 3, 2012.
[10]. S. Mogil, K. Simmonds, J. Simmonds, "Pain Research from 1975 to 2007: A categorical and bibliometric meta-trend analysis of every research paper published in the journal Pain." Pain, 142, pp 48-58, 2009.
[11]. J. Hung, K. Zhang, "Examining mobile learning trends 2003-2008: a categorical meta-trend analysis using text mining techniques", Journal of Computing in Higher Education, Volume 24, Issue 1, pp 1-17, 2012.
[12]. P. Rayson, R. Garside, "Comparing Corpora using Frequency Profiling" in proceedings of the Workshop on Comparing Corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics, ACL 2000, Hong Kong.
[13]. A. Mundade, T. Pattewar, "Comparison Study of Optimized Test Suite Generation Using Genetic and Memetic Algorithm", International Conference on Pervasive Computing, ICPC, 2015.
[14]. K. Cosh, R. Burns, T. Daniel, "Content Clouds, Classifying Content in Web 2.0", Library Review, Vol. 57, Issue 9, 2008, pp 722-729.
[15]. P. Robert, Y. Escoufier, "A Unifying Tool for Linear Multivariate Statistical Methods: the RV-Coefficient", Journal of the Royal Statistical Society, Vol. 25, No. 3, 1976.
[16]. K. Cosh, "On Automatically Extracting Discoveries from User Generated Content" in proceedings of CISIS 2014, the 8th International Conference on Complex, Intelligent and Software Intensive Systems, 2014.
[17]. V. Yusifoğlu, Y. Ammannejad, A. Can, "Software test-code engineering: A systematic mapping", Information and Software Technology 50, 2015, pp. 123-147.129
[18]. Design and Application of Intelligent Dynamic Crawler for Web Data Mining"2017 May in 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC) Page(s):1098 - 1105