Library Waves
Volume 6, No. 2 (2020)
ISSN 2455-2291 (Online)

An Overview of Text Mining: Application and Free Software Tools

*** Khusbu Thakur[#]** **** Dr. Vinit Kumar**

***** Research Scholar, Department of Library and Information Science, Babasaheb Bhimrao Ambedkar University, Lucknow (UP) INDIA; Email: thakurkhusbu02@gmail.com

****** Assistant Professor, Department of Library and Information Science, Babasaheb Bhimrao Ambedkar University, Lucknow (UP) INDIA; Email: mailvinitkumar@gmail.com
# Corresponding author.

Received: 02 December 2020; Accepted: 28 December 2020; Published: 31 December 2020

Abstract

As we keep everything systematic in our daily life, the same rule we apply in the professionals' work. Nowadays, we have tons of textual unstructured information and it needs to be stored in a proper sequence into the database system. For this, there are needs for extraction tools and its knowledge of how to extract and generate interesting unknown patterns from unstructured data. Having lots of confusion and various questions generated in mind such as how to do, where to begin, what are open-source software tools, its features, and what are the possible applications in libraries management systems? So, it is being described in this study, which is basic information of an overview of text mining, its application in the library management system and in the last section we have discussed the most trending free & open-source software of text mining, which is widely used in text mining, machine learning and data science. So, the goal of this paper will be helpful to beginners of text mining research who are new in the field of text mining research for decision making over research trend & library management services within minimum efforts.

Keywords: Text mining tools, open source tools of text mining, application of text mining.

1. Introduction

Text mining is one of exciting techniques, which always receives the best attention to manage the information that occupies the vast documents. Information retrieval is easy to search documents and find large collections of data in the form of unstructured and fulfill the user's demands. How does this happen? Scientists have developed AI i.e. Artificial Intelligence. It is a powerful technique that works like a human mind. It has more capabilities to control, to analyze, to manage and to demonstrate all the resources. However, Text mining is a trending technique which is used in marketing research and business analytics. But in the last few years, text mining has started in the social science disciplines. Text Mining is a technical process to extract hidden knowledge. Nowadays, more than 80 % of unstructured data is growing in various forms such as images, texts, emojis, and so on. Therefore, all these unstructured information needs to purify them in a particular form for better analysis and to find out good results. We can define text mining as the trend solution to purify the unstructured data and discover the unknown pattern. Text mining has various synonyms such

as Intelligent Text Analysis, Text engineering, Text Data Mining or Knowledge Discovery in Text (KDT), refers generally to the problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. It has several applications such as to classify the text documents into different categories, Search large text, translation, information extraction, understanding speech (Gohil, 2015). NLP techniques involve part of speech, stemming, lemmatization by text mining. TM is a powerful technique which develops two models one is to predict in future and from data gaining deep knowledge, discovers new pieces of information from textual data which is earlier unidentified information by extracting it using different techniques. Here in this paper, we will discuss the framework and text pre-processing steps of text mining with the help of figure 1 and figure 2. In the third phase of paper, table 1 will discuss the basic characteristics of free software tools of text mining, which are mostly trending software in text mining scholarly articles. In the last section, we will talk about the applications of text mining in the library management system. So, this study aims to discuss the important features of the popular text mining tools used by researchers in their scholarly articles and compare the script coder software and graphical visualization tools. This study will be helpful for beginners of text mining research to overcome the challenges while selecting text mining tools and to do different tasks with their research interest and to take decisions over the research trends along with the library management system to develop library services within minimum efforts.

Basic steps of text mining framework

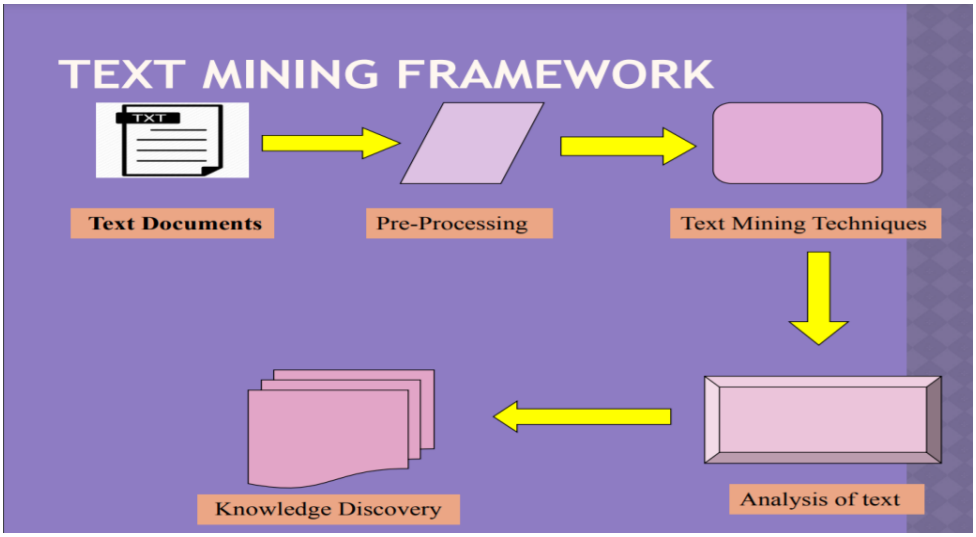

Figure - 1: Process of text mining

Figure 1, represents the five major steps involved in text mining. Text mining is used to extract interesting information, knowledge and patterns from the unstructured documents that are from different sources. Then apply pre-processing (compare to NLT – Natural Language Text) of text and then Text Mining techniques. Then, analyse the text document and finally discover the knowledge from text documents (Sheela and Bharathi, 2018).

2. Steps of text pre-processing in text mining

This section has discussed the general outline of text mining pre-processing. It is a technique of reducing unwanted data from collection of text documents. In other words, it is defined as a task which converts the raw data into well-defined knowledge. There are some essential operations of text mining pre-processing steps as follows:

a) Text Pre-processing

- **Tokenization Process:** It is the process of breaking a stream of textual content up-to words, terms, symbols or some other meaningful elements called tokens. Filtering (Stop words) removes unnecessary information which includes preposition, articles, conjunction and many more.
- **Lemmatization Process:** It is also a process which is like stemming but it is identifies the dictionary forms of words. This technique is used to reduce the length of text words. In text mining techniques, pre-processing steps play a major role in transforming the words roots into proper roots for appropriate text analysis.

b) Text Transformation Process

- Text transformation uses a bag of words or a vector space model. It performs a feature selection task. Under this it reduces the dimensionality by removing redundant and irrelevant features (Sheela and Bharathi, 2018). Even removes the sequence of words which occur frequently which has no sense or unimportant content in the collection of text documents.

c) Text Mining Methods

There are various types of text mining methods which includes Text Clustering, Text Summarization, Text Categorization as given below:

- **Text Clustering Process:** Text clustering is a technique that can be used in measuring similarities and grouping the similarities of texts. For the best result, text documents depend on the quality of clustering and find out good results between the lower and higher similarities from multiple text documents.
- **Text Summarization Process:** Text summarization is a technique which shortens the documents text and mentions the only essential point which is very much important in the text documents. TS is also known as text review which means to reduce the length of documents. The key advantage of text summarization is “save the time” which follows the five laws of library science, which was given by Dr. S. R. Ranganathan.
- **Text Categorization Process:** Text Categorization is a process to category the multiple documents automatically into various categories, categorization is a supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content (Krishnamoorthy and Mani, 2014). The main purpose of text categorization is to train classifiers based on supervised and unsupervised categories automatically. Statistical classification techniques like Naïve Bayesian classifier, Nearest Neighbour classifier, Decision Tree, and Support Vector Machines can be used to categorize text (Gaikwad, Chaugule and Patil, 2014).

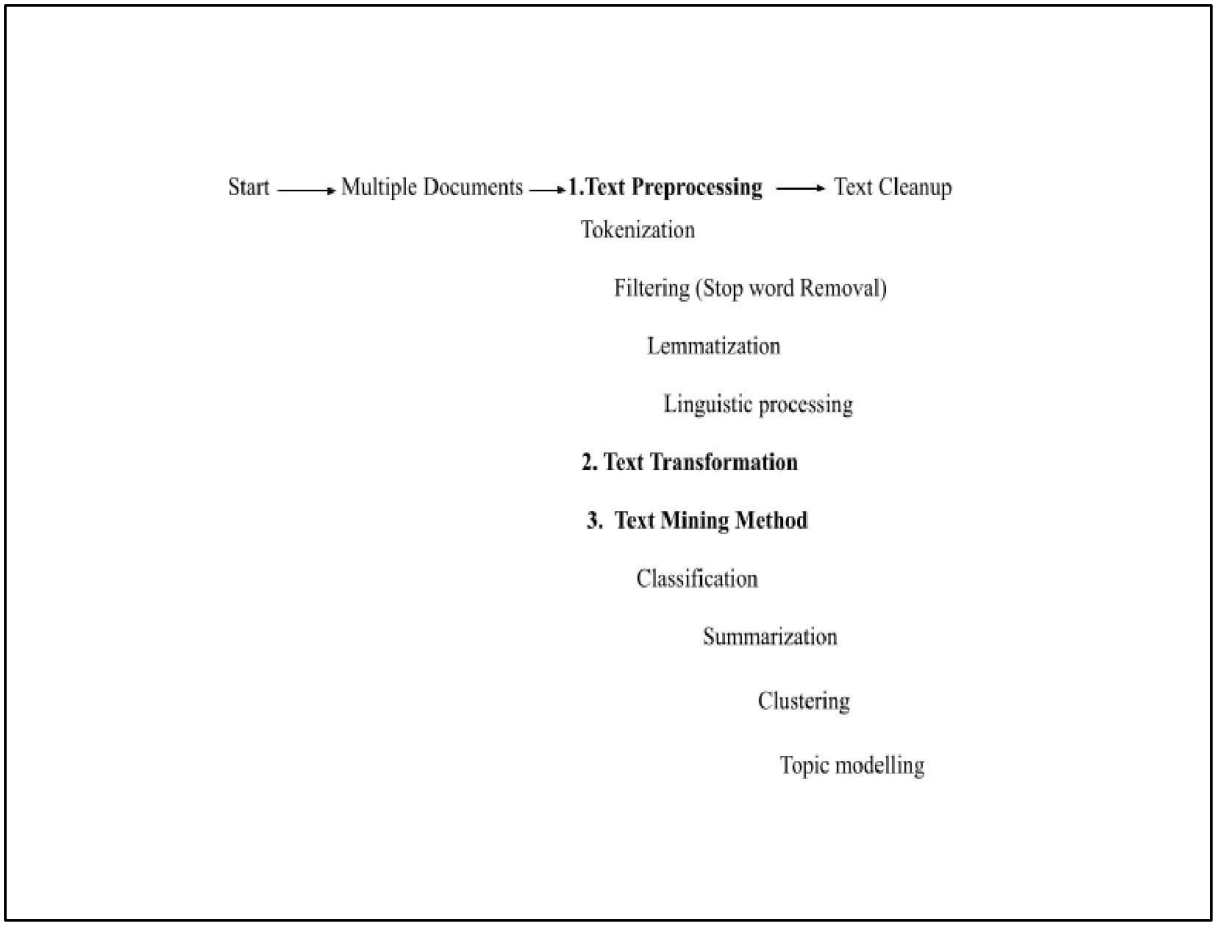

Figure - 2: Steps of text pre-processing for text mining

Above figure 2, shows the text cleaning process. It is a specific task to remove unwanted words, signs, symbols, tables, figures and many more. It is the process for reducing modulated words to their words stem, root or base which is filtered out before or after the processing of natural language text (Sathya and Rajendran 2015).

3. Objectives of the study

Q1. To discuss an overview of text mining and its application in the library management system?

Q2. To evaluate the most important features of trending tools of text mining techniques?

4. Methodology

The data for this study has been collected from 85 scholarly articles where the text mining researchers mostly utilized text mining tools in their scholarly articles, i.e. RStudio, Python are open source software and non-coders are Orange and RapidMiner. Which were further considered in the preparation of checklist to evaluate script code and graphical Visualization tools.

5. Trending of text mining tools an overview

In this section, discuss the general characteristics of most utilized text mining tools, i.e. RStudio, Python, RapidMiner and Orange. All four tools implement on Windows, Linux and Mac OS Operating System as well as it supports the various applications of text mining techniques. Nowadays, R and Python are the most trending tools for text mining visualization especially for script code but RapidMiner and orange are graphical visualization tools especially for non-code.

Table - 1: General characteristics of free software tools of text mining

Basic characteristics	RStudio	Python	RapidMiner	Orange
Year of published	1992	1991	2006	1996
Developer	Worldwide development	web development	RapidMiner, Germany	University of Ljublijana
Programming Language	C, Fortran, R	C, C++, java etc.	Java	Python, C, C++, java
Available Version	3.4.3	3.9.1	9.6	3.4.5
Licence	free-open-source GNU	Free and open	RapidMiner Studio Free is limited	GPLv3
Main purpose	data analytics, statistical analysis, as well as machine learning	For data science	general data mining	Text mining , Machine Learning, Data Visualization, Data Analysis
Community support	very large (~ 2 M users)	N/A	large (~200 000 users)	N/A
platform	Cross- platform	Cross- platform	cross-platform	Cross-Platform

Above table 1, shows the general characteristics of RStudio, Python, Orange and RapidMiner. RStudio, Python, and Orange are open and free tools and RapidMiner is free software. These tools are very much flexible to implementation for various tasks of text

mining. All these four tools are compared to general characteristics such as license, running platform, languages, support system for text mining, data science, machine learning. RapidMiner and Orange software are user friendly, users simply drag and drop and connect the wire together. Overall, it is very much flexible and sophisticated for business statistical, prediction analysis, machine learning and for data visualization.

6. How is text mining beneficial for libraries management systems?

In the library management system, text mining is highly beneficial to explore the quality of text, meaningful pattern of knowledge, as well as insights are extracted from multiple documents. We can say that nowadays, every platform can benefit from text mining.

a) Using text mining techniques, we can automatically identify the academic libraries user identity such as name of person, place, organization, customer demands, interest, location and so on even create users related information in libraries.
b) Predicting whether a packet of network data can pose a cyber-security threat.
c) It helps to investigate the issues, trends as well as solutions of the particular research topic.
d) It also helped in promoting and responding to the market or promotion offers.
e) In the library's centre, it helps to identify the users' details and then take appropriate actions.
f) It identifies which users are most likely to subscribe to which magazine/ journals.
g) It helps in identifying people likely to subscribe to what kind of content that they are publishing through their magazine/Journals.
h) It also works in improving the research process and its quality of research study.

7. Conclusion

From the above discussion of this paper, text mining is a technique that is used for text mining. This study has represented the basic ideas of an overview of the text mining framework that has been introduced in the above. This paper also discussed the basic characteristics of free software that are easily available & trending tools nowadays for text mining, i.e RStudio, Python, Orange and RapidMiner. With respect to this providing an overview of text mining to the beginners who will explore and use any particular tool in searching, exploring the knowledge in their research interest. Eventually, using ™ tools techniques will predict the trends and themes of the research, the new concepts of research and check duplicate text documents from multiple documents such as articles, news, blog and many more. Even librarians will also improve their services like reference services, CAS, SDI more effectively in their library management system with minimum efforts. Overall, this paper will improve the understanding level of text mining researchers.

References:

1. Agrawal, R., & Batra, M. (2013). A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering, 2*(6), 118-121.
2. Dang, S. & Ahmad, P. H. (2015). A review of text mining techniques associated with various application areas. International Journal of Science and Research (IJSR), 4(2), 2461-2466.
3. Dang, S. & Ahmad, P. H. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations, 1*(4), 866-2348.

4. Gohil, L. (2015). Text mining: process and techniques. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 3*(3), 2347-5552.
5. Gupta, V. & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.
6. Krishnamoorthy, M. & Mani, M. (2014). A Brief Survey on Text Mining and its Applications. *Int. J. Computer Technology & Applications, 5*(5), 1637-1640.
7. Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, M. (2014). Pre-processing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.
8. Pynam, V., Spanadna, R. R., & Srikanth, K. (2018). An Extensive Study of Data Analysis Tools (Rapid Miner, Weka, R Tool, Knime, Orange). *SSRG International Journal of Computer Science and Engineering (SSRG–IJCSE), 5*(9), 4-11.
9. Prakash, K., Chand, Prem & Gohel, Umesh (2004). Application of data mining in library and information services. 2nd Convention PLANNER-2004, Manipur University, Imphal (pp.168-177). INFLIBNET Centre, Ahemdabad.
10. Sathya, S. & Rajendran, N. (2015). A review on text mining techniques. *Int. J. Computer Sci. Trends Technology*, *3*(5), 274-284.
11. Sheela, S. & Bharathi, T. (2018). Analysing Different Approaches of Text Mining Techniques and Applications. *International Journal of Computer Science Trends and Technology*, *6*(4), 23-29.

■■■