International Journal of Current Research and Modern Education (IJCRME)
Impact Factor: 6.725, ISSN (Online): 2455 - 5428
(www.rdmodernresearch.com) Volume 2, Issue 2, 2017

A SURVEY ON TEXT PATTERN MINING TOOLS

K. Sundaravadivelu
Department of Computer Science, Madurai Kamaraj University, Madurai, Tamilnadu

Cite This Article: K. Sundaravadivelu, "A Survey on Text Pattern Mining Tools", International Journal of Current Research and Modern Education, Volume 2, Issue 2, Page Number 415-418, 2017.

Copy Right: © IJCRME, 2017 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract*:*

Today's world scenario is something homogeneous to people who are drastically flourishing in the cyber world as it grows much more expeditious. There arises a quandary of categorising, analyzing, summarizing and discovering of data that are given by users in gregarious media like Facebook, Twitter, E-mails etc., in the form of text. Text mining is essential if constructive cognizance needs to dig out from heaps of text. However, where to commence, what are the favourite implements, which techniques are utilized and what metrics can be used along with them are discussed. This paper endeavours to explain the available implements for text mining with its performance metrics which gives a smooth kick-start to the incipient researchers and practitioners in the field of text mining.

Key Words*:* Categorizing, Analyzing, Summarizing, Discovering, Metrics, Text Mining

1. Introduction:

Due to the rapid development of hardware technology, there is a desideratum for storing large data as its capacity increases immensely. As exponentially the data grows every day, the users need variants of data in which they utilize textual data hugely rather than the numerical one. Hence there comes the role of developing techniques that are applied to the textual data which are different from the numerical data. The mining of these kinds of textual data is called Text mining.

Text mining is flourishing nowadays. There is an elevating demand to build better and efficient solutions that are capable of dealing with some sticky situations. There is no bridge between the utilizing needs and the implements available in the industry. Text mining is an integrated ground concerning the retrieval of erudition, understanding of the text, extraction of information, clustering and categorization, visualisation of data, machine learning in data science and mining of data.

Text mining is a confronting task since it deals with unstructured and fuzzy data. Utilizing text mining the possibility of analyzing and structuring large colossal sets of documents that apply statistical and computational linguistics technologies is essential nowadays. The pattern and hidden information of the text can be elucidated utilizing Text Mining. The subsistence of data is ubiquitous in the form of text where we are unable to store all the available data in the database. From time to time the analyzing of data is done that is not backed up in synergistic databases preferably it can be extracted from the firm's website, emails or reports. Indeed, the technique utilized for structured data backed up in the database cannot be implemented in an unstructured data. Due to the complexity of an unstructured data, to extract information from them, more robust methods are required. Consequently, to analyze an unstructured data, i.e. Text, we obligate text mining implements with better performance.

A standard text mining task includes a categorization of text, clustering of text, extraction of text, engendered of granular taxonomies, analysis of text, summarization of documents, and entity cognition modelling. Simple text mining implements are inadequate to handle text, so we require some well-equipped implements with a high-performance rate that can analyze, categorize, summarize and discover text. Text mining can be utilized on any ground for business perspicacity, software process analysis, gregarious media analysis, biomedical analysis, sentiment analysis and even for security analysis. Thus text mining has extensiveutilization in different grounds, and one needs to ken about the implements that are subsisting and the metrics utilized by the implements to assist so that gaining of knowledge becomes more facile and more expeditious.

The overall goal and purpose are to convert text into data, for analyzing or classifying using categorization, information retrieval, lexical analysis, pattern recognition, tagging/annotation, information extraction using text analytics. Discovering of textual data built on the approach of probability and statistics can be recognized by the statistical functional words in the sample data, i.e. Text and whether or not the text is suspicious is identified utilizing mining tool. The term summarization refers to the automated summarizing of a sizably voluminous document into a crisp one, as opposed to the manipulation done manually. The utilization of text mining implements can explore the quandaries faced during each process of text mining.

For the text mining system development, comparison of metric evaluation is essential, but selecting congruous measures and metrics are not a simple task. The measures and metrics are perpetually developing in many domains, and it is consequential to consider the characteristics of the domain into account, then the metrics are decidable.

2. Motivation:

Text mining is the course of collaborating the information from a massive database by analyzing its relations, rules, and patterns from the text. Thecritical point is that rearranging the mined data together to produce a novel hypothesis that can be further developed by means of experimentations. The utilizer's search for a relevant document for their requirements, but simultaneously irrelevant documents is also fetched along with them. To resolve the above problem, the texts are mined so that the utilizer gets what they want by excluding the other kinds of stuff that are irrelevant to them. Text mining is classified as text summarization, text categorization, text analytics and knowledge discovery.

.Due to increment in the availability of information in digital form, there occurs a quandary of intricacy in accessing them in an efficient way. So text categorization acquires a wide ocular perceiver on it because the paramountcy of categorizing the textual data makes the utilizer to access their required information facilely. The application of text categorization varies from the indexing of the document according to the lexical and the extraction of a document. Text categorization is the intersection of machine learning and information retrieval which assimilates cognizance from texts as well as from text mining. The limber development of textual information in online, there is a desideratum for categorizing of texts according to the utilizer's requisites, we require a technique to relegate incipient facts and discover intriguing information available online

The concept of "Text Mining" is commonly used for characterizing a large amount of textual data, pattern recognition of the text and extraction of knowledge from it. The terms "text analytics" and "text mining" are basically transposable. They both have methods, tools, and applications in common. Their difference stalks from the people who is using what. Data miners use "Text Mining", and the organizations and individuals use "Text Analytics" in various domains to pave the way for business development methods that make a lot of difference in its growth. If structured data is immensely colossal, then unstructured data is sizably voluminous. It is generally accepted that analyzing textual data cannot be done 100% as structured data as 70% of them are unstructured and fuzzy. The gain of an organization is insufficient because of the huge amount of structured data. In order to overcome the barrier between structured and unstructured data potentially we need well-equipped tools and techniques for finding the solution. As a solution Text Analytics - extracts information from the conversation of language/ framework using 5-W's ("Who," "Where," "When", "What" and "Why") conversation and 1-H ("how" people feel,and the conversation comes to light). This is the reason why man is going to be replaced by the machine

Machines are needed more than a human if the size of the data is astronomically immense. The time consumed by a human to read, understand and analyze the data is far more than a machine. In order to engender an outline, we have to identify the most dominant pieces of information from the document, omitting impolite information and minimizing particulars, and pull together into a compact coherent report, we have the need of an automated tool that has paved the way for extensive research in automatic text summarization. Automatic text summarization can produce a precise and fluent summary of the preserving content that may be structured or unstructured. Man fails over machines as they yield far better results than human works.

Knowledge discovery is the process of extracting useful information that is spoken openly inside the text mining document, which is an efficacious approach for mining of text on the basis of erudition to be discovered. Information Retrieval is a technique that can accommodate new mining process for text. On the other hand, if an unstructured data resides in a document instead of conceptual knowledge, Information Retrieval system is just a supplementary technique for extracting that information to make over those unstructured data into a structured data. Then the extracted data are identified by the use of traditional cognizance mining tools to obtain the conceptual patterns.

3. Justification:

The vital role of text mining is to dig out information from textual resources and enable them to utilizers whereas text mining tool deals with the functions of categorizing/relegation, retrieving, Natural Language Processing (NLP). In order to relegate and discover patterns from various documents, some of machine learning tools can collaborate with text mining tools.

A. Text Categorization:

For handling and organizing an enormous amount of amorphous textual data that are yielded from online, text categorization plays a vital role in this work. This technique is habituated to relegate new stories to obtain exciting news that is available on WWW to match the utilizer's needs. Text Categorization is the process of transferring a document into a definite number of categories. The initial step in the categorizing of text is the transformation of documents which contains strings of characters into an illustration for the cognition algorithm and the relegation task. The precision of the modern text relegation systems rivals with the trained human professionals and thanks to an amalgamation of information retrieval technology and machine learning technology.

B. Text Analytics:

Text Analytics is the course of translating amorphous textual data into essential data for analysis, to quantify views and responsibility of the customer to offer search facilities, sentimental analysis, and text

modelling to fortify fact-predicated conclusions. It uses many linguistic, statistical and machine learning techniques for the retrieval of information from an amorphous text data and the techniques for obtaining patterns, evaluating and interpreting the yield. The analysis takes into account the keywords, meanings, and tags from extensive data that are obtained from a firm in the form of different file formats. Text analytics go in hand in hand with text mining.

C. Text Summarization:

A concise and fluent summary of text by preserving its crucial information and overall construct is called Text Summarization. A wide range of text summarization techniques has been developed in recent years in several domains. As a human, we are unable to summarize an astronomically large piece of text by reading it entirely which is facilely done by a machine with the availability of the implement highlighting the critical content. A machine lacks language capability which can be facilely handled utilizing an automatic summarizing tool that yields result far better than the human works.

D. Knowledge Discovery:

The extraction of information from a vast database that is implicitly accessible, interiorly indefinite and latently subsidiary for the utilizers is kenned as knowledge revelation. Text mining is the revelation of fascinating erudition in text documents. It is a challenging issue to find precise erudition (or features) in text documents to avail users to find what they operate. Knowledge revelation can be efficaciously used,and update discovered patterns and apply it to the field of text mining.

4. Metric Analysis:

Metrics are used for measuring the manners, activities of its supporting organization and needs of the utilizers. The metrics of text mining tools are examined broadly and are tabulated. In Comparative study of metrics among text mining tools is visualized. Indeed, new metrics are placing their importance in the market than the existing metrics.

A. Inter-Rater Reliability:

Inter-rater Reliability, inter-rater agreement, or concordance, is the degree of agreement among raters. It gives a score of how much homogeneity, or consensus, there is in the ratings given by judges. Different statistics are appropriate for different types of measurement. Some options are joint-probability, inter-rater correlation, concordance correlation coefficient.

B. Similarity Measures:

Similarity measure or Similarity function is an absolute-valued function that quantifies the similarity between the two data. Frey and Dueck suggest defining a similarity measure

$$s(x, y) = -\|x - y\|_2^2$$

Where $\|x - y\|_2^2$ quantifies the similarity between the data

C. Precision (P):

Precision in text mining context, is the fraction of recovered textual file that are related to the query:

$$\text{Precision (P)} = \frac{|\{\text{related textual file }\} \ \Lambda \ \{\text{ recovered textual file }\}|}{|\{\text{recovered textual file }\}|}$$

D. Recall (R):

Recall in text mining context, is the fraction of the related textual file that is successfully recovered.

$$\text{Recall(R)} = \frac{|\{\text{related textual file }\} \ \Lambda \ \{\text{recovered textual file }\}|}{|\{\text{ related textual file}\}|}$$

E. F-Measure:

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \ x\frac{P*R}{P+R}$$

5. Conclusion:

There are many text mining implements that are available in the market to enhance the conventional research on text mining. Some of them are unable to handle an enormous amount of unstructured data since most of the data is in the form of text. Virtually every utilizer works with text, there is a desideratum for text mining implements which has better performance rate, available in the form of open source, online and proprietary. This paper has given different ideas from different papers and websites for useful text mining with the help of tools that are available and the measures and metrics supported by the tool to achieve high-performance rate. The insight of metrics and measures discussed in this paper immensely guide the new researchers and practitioners to select the tool according to their requirements.

6. References:

1. Dekhtyar, Alexander, Jane Huffman Hayes, and Tim Menzies. & quot; Text is software too. & quot; MSR 2004: International Workshop on Mining Software Repositories at ICSE'04: Edinburgh, Scotland. 2004.

2. https://www.igi-global.com
3. Berry Michael W., Automatic Discovery of Similar Words, in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 2004, pp.24-43.
4. Vishal gupta and Gurpreet S. Lehal , "A survey of text mining techniques and applications", journal of emerging technologies in web intelligence, 2009,pp.60-76
5. Durga Bhavani Dasari & Dr. Venu Gopala Rao. K, "Global Journal of Computer Science and Technology Software & Data Engineering",Volume 12, 2012 , ISSN: 0975-4350
6. https://www.linguamatics.com
7. http://www.predictiveanalyticsworld.com
8. Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E Williams. 2007. Fast generation of result snippets in web search. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 127–134.
9. Raymond J. Mooney and Razvan Bunescu, "Mining Knowledge from Text Using Information Extraction" ,SIGKDD Explorations, Vol 7, pp.3-10
10. Sateli, B., Angius, E., Rajivelu, S. S., & Witte, R. (2012). Can text mining assistants help to improve requirements specifications. Mining Unstructured Data (MUD 2012), Canada.
11. Malhotra, Ruchika, et al. & quot; Severity Assessment of Software Defect Reports using Text Classification & quot; International Journal of Computer Applications83.11 (2013).
12. Sharma, Gitika, Sumit Sharma, and Shruti Gujral. & quot A Novel Way of Assessing Software Bug Severit Using Dictionary of Critical Terms & quot; Procedia Computer Science 70 (2015): 632-639.
13. Jurek, Anna, Maurice D. Mulvenna, and Yaxin Bi. & quot; Improved lexicon-based sentiment analysis fo social media analytics & Security Informatics 4.1 (2015): 1-13.
14. Niharika, V. Sneha Latha and D. R. Lavanya, "A Survey On Text Categorization", International Journal of Computer Trends and Technology , vol 3, pp. 39-45, 2012
15. Feldman, Ronen, and James Sanger. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.