International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11

ISSN: 2454-4248
542 – 544

Finding Frequent Text-Patterns on Big Data using Visualisation Techniques

Sangram Keshari Swain
Associate Professor, Department of Computer Science & Engineering, School of Engineering & Technology, Bhubaneswar Campus Centurion University of Technology and Management, Odisha, India
E-mail: sangrambapun@gmail.com

Dr. Srinivas Prasad
Professor, Department of Computer Science & Engineering, K L University, Andhra Pradesh, India.
E-mail: srinivas_prasad@hotmail.com

Dr. M. Vamsi Krishna
Associate Professor, Department of Computer Science & Engineering, School of Engineering & Technology, Parlakhemundi Campus Centurion University of Technology and Management, Odisha, India
E-mail: mvamsi@cutm.ac.in

Abstract: Big Data has got more importance in various industries over the last couple of years, which is applied on various huge and large data sets because it cannot be stored and process through traditional databases. Big data has huge potential to store and process such huge and large datasets in several ways because, during processing, we are analyzing the large datasets in required time. Text analysis is still somewhat in its infancy but is very promising. Because in most of the companies 80% of data is in unstructured form, while most types of analysis only work with structured data. In this paper, we are discussing some visualization techniques available for Big data. Finally, R package is used to analyze unstructured text because R is freely available and it comes with lots of free packages and powerful tools through which we can easily analyze the large data sets in a sufficient time.

Keywords: *Text Mining, Text Analysis, Data Visualization.*

*****

I. Introduction:

In the twentieth century ahead this World Wide Web has modified the means of expressing their views. Gift scenario is totally they're expressing their thoughts through on-line blogs, discussion forums and additionally some online applications like Facebook, Twitter, etc. Over the past 10 years, industries and organizations don't have the requirement to store and perform operations and analytics on the information for the shoppers. However, around since 2005, the requirement to rework everything into information is way amused to satisfy the wants of the individuals. Therefore huge information came into the image within the real-time business analysis of process information.

If we have a tendency to take Twitter as our example, nearly 1TB of text information is generated inside per week within the sort of tweets. Among these tweets will be classified by the hash worth tags that they're commenting and posting their tweets. So, currently, several corporations and additionally the survey corporations area unit mistreatment this for performing some analytics specified they will predict the success rate of their product or additionally they will show the various read from the information that they need to be collected for analysis. But, to calculate their views is extremely troublesome during a traditional means of taking this serious information that area unit about to generate day by day.

Text mining [1] has become an extremely great tool to analyze and understand massive data sets not done by the traditional analysis techniques. These tools are applied to a spread of data issues, like understanding themes in social media or facilitating info retrieval in unstructured information. Moreover, these tools also can assist in the improvement and structure text-based information for future analysis in the mental image or different graphical tools. Text mining is one amongst the foremost frequent nevertheless difficult exercise sweet-faced by beginners in information science/analytics consultants. The most important challenge is, one has to completely assess the underlying patterns in text, that too manually.

R is each a programming language and environment oriented towards applied mathematics, computing and graphics creation (R Core Team, 2016). R is formed obtainable underneath the GNU; as a result of the study, community involvement, there are various extensions, known as packages, developed over time, additionally as sturdy documentation. Because of this extensibility and flexibility, R has remained systematically standard for knowledge and text mining applications across several domains and includes powerful text mining tools.

II. Literature review:

Anjali Ganesh Jivani [2] mentioned that the aim of stemming is to cut back totally different grammatical forms or word varieties of a word like its noun, adjective, verb, adverb, etc. The goal of stemming is to cut back inflectional kinds and generally derivationally connected varieties of a word in a standard base form. This paper discusses totally different ways of stemming and their comparisons in terms of usage benefits similarly as limitations. The essential distinction between stemming and lemmatization is additionally mentioned.

Vishal Gupta et.al [3] has analyzed the stemmers performance and effectiveness in applications like spell checker varies across languages. A typical straightforward stemmer rule involves removing suffixes employing a list of frequent suffixes, whereas an additional complicated one

 ISSN: 2454-4248
542 – 544

would use the morphological information to derive a stem from the words [6]. The paper offers a close define of common stemming techniques and existing stemmers for Indian languages.

K.K. Agbele [4] mentioned the technique for developing pervasive computing applications that square measure versatile and elastic for users. During this context, however, data retrieval (IR) is commonly outlined in terms of location and delivery of documents to a user to satisfy their data would like. In most cases, morphological variants of words have similar linguistics interpretations and may be thought of as equivalent for the aim of IR applications. The rule Context-Aware Stemming (CAS) is projected, that may be a changed version of the extensively used Porter's stemmer. Considering solely generated purposeful stemming words because of the stemmer output, the results show that the changed rule considerably reduces the error rate of Porters rules from seventy six.7% to 6.7% while didn't compromise the effectiveness of Porters rules.

Hassan Saif [5] has investigated whether or not removing stop words helps or hampers the effectiveness of Twitter sentiment classification strategies. For this investigation he has applied, six totally {different / completely different} stop word identification strategies to Twitter knowledge from six different datasets and observe however removing stop words affects 2 well-known supervised sentiment classification strategies. The result shows that victimization pre-compiled lists of stopwords negatively impact the performance of Twitter sentiment classification approaches. On the opposite hand, the dynamic generation of stop word lists, by removing those occasional terms showing just once within the corpus seems to be the best methodology for maintaining a high classification performance whereas reducing the info meagreness and well shrinking the featured house.

There are various text classification and clustering methods. Most text categorization techniques reduce this large number of features by eliminating stop words or stemming. This is effective to a certain extent but the remaining number of features is still huge. It is important to use feature selection methods to handle the high dimensionality of data for effective text categorization. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. There are many feature selection methods. Mainly they are classified as filter and wrapper feature selection methods. One of the feature selection methods can choose to identify relevant information from documents based on their content. Selection of feature selection is an important and challenging step in text mining.

III. Problem definition:

Text mining [11] can help an organization to derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Data mining or Text mining plays an important role in decision making because through these mining techniques we can analyze the data and on the basis of the result, we can take a decision.

Nowadays social media sites like Twitter are widely used to share user opinions on various topics, Twitter gives a platform to the user to share their views and thoughts on the various field like political, industrial, education and there is a petabyte of data generated by Twitter in a day. So the mining techniques are used to analyze the social twitter data through we get a lot of data sets to analysis. So the analysis of twitter data provides a better way for making the decision.

IV. Different visualization techniques:

Text mining of Twitter data with R packages twitteR, tm, and wordcloud.

a. twitterR

Package *twitteR* provides access to Twitter data, first, we are creating a Twitter streaming API called Twitter app by which we can get out access token keys. through this package, we are retrieving data called tweets from the twitter server and provide storage.

b. tm

tm package [9] provides functions for text mining, The main structure for managing documents in tm is a so-called Corpus, representing a collection of text documents. A corpus is an abstract concept, and there can exist several implementations in parallel. The default implementation is the so-called VCorpus.

c. wordcloud

wordcloud visualizes the result with a word cloud. A word cloud (or tag cloud) can be a handy tool when you need to highlight the most commonly cited words in a text using a quick visualization. It is produced a word cloud based on the titles' word frequencies calculated using the powerful tm package for text mining [10].

V. Proposed methodology:

Our Steps or Algorithm Steps will follow:

First Step: First we create a Twitter API for retrieving text from Twitter for analysis.

Second Step: After retrieving we transformed the text, tweets are first converted to a data frame and then to a corpus [8]. After that, the corpus needs a couple of transformations, including changing letters to lower case, removing punctuations/numbers and removing stop words.

Third Step: In many cases, words need to be stemmed to retrieve their radicals. For instance, "example" and "examples" are both stemmed to "example". However, after that, one may want to complete the stems to their original forms, so that the words would look "normal".

Fourth Step: After transforming and stemming process [7] is done then we build a document-term matrix. Based on the matrix, many data mining tasks can be done, for example, clustering, classification and association analysis.

Fifth Step: With the help of matrix we can identify the frequent words and their association between words.

Six Step: After building a document-term matrix [12], we can show the importance of words with a word cloud (also known as a tag cloud).

 ISSN: 2454-4248
542 – 544

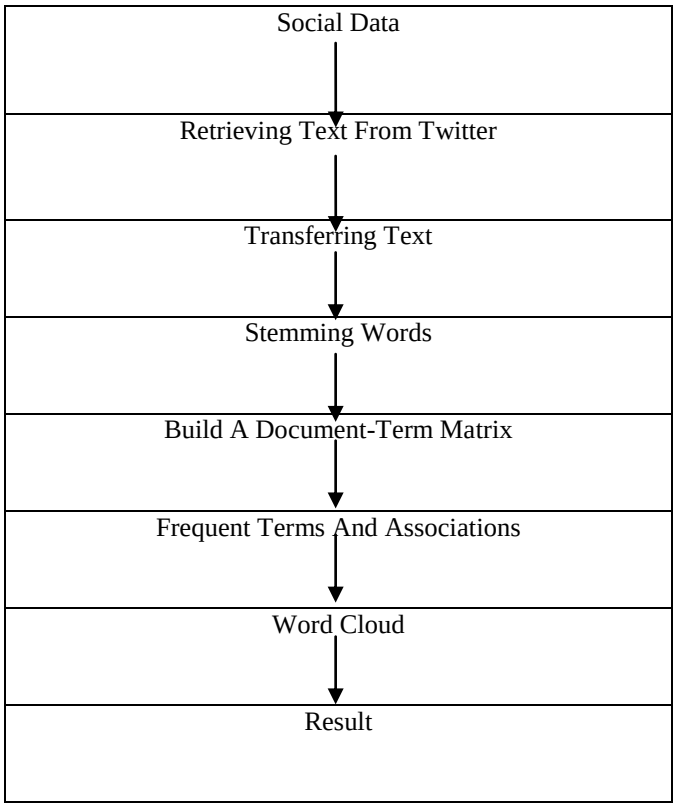

Fig – 1: ANALYSIS STEPS

VI. Conclusion:

Twitter data is very useful in decision making because it provides a variety of options on various topics. So the texting mining will perform on twitter data and we are using visualizing techniques called R which comes with a variety of packages. Through which we are using twitter package for retrieving a real-time data, tm package a very powerful tool for text mining and wordcloud package for visualization.

References

[1] Mehmet Ula_ Çak_r‖ Text Mining Analysis in Turkish Language Using BigData Tools‖ 2016 IEEE 40th Annual Computer Software and Applications.

[2] Anjali Ganesh Jivani, A Comparative Study of Stemming Algorithms, International Journal of Computer, Technology, and Application, Volume 2, ISSN:2229-6093.

[3] Vishal Gupta, Gurpreet Singh Lehal, A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages, Journal of Emerging Technologies in Web Intelligence, VOL. 5, NO. 2, MAY 2013.

[4] Agbele, A.O. Adesina, N.A. Azeez, & A.P. Abidoye, Context-Aware Stemming Algorithm for Semantically Related Root Words, African Journal of Computing & ICT.

[5] Hassan Saif, Miriam Fernandez, Yulan He, Harith Alani, On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter.

[6] Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009.

[7] Porter M.F, Snowball: A language for stemming algorithms. 2001.

[8] Mladenic Dunja, Automatic word lemmatization. Proceedings B of the 5th International Multi- Conference Information Society IS. 2002, 153-159.

[9] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, ―Effective Pattern Discovery for Text Mining‖, *IEEE Transactions on Knowledge And Data Engineering*, Vol. 24, No.1, January 2012.

[10] Yuefeng Li, Sheng-Tang Wu and Xiaohui Tao, ―Effective Pattern Taxonomy Mining in Text Documents‖, *ACM*, October 26–30, 2008.

[11] Pattan Kalesha, M. Babu Rao, and Ch. Kavitha, ―Efficient Preprocessing and Patterns Identification Approach for Text Mining‖, *International Journal of Computer Trends and Technology (IJCTT)*, Vol. 6, No. 2, December 2013.

[12] Zakaria Elberrichi, Abdelattif Rahmoun, and Mohamed Bentaalah, ―Using WordNet for Text Categorization, *International Arab Journal of Information Technology*, Vol. 5, No. 1, January 2008.