2nd International Conference on Current Research Trends in Engineering and Technology
© 2018 IJSRSET | Volume 4 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Text Mining and Natural Language Processing in Web Data Mining

Bhoomika Joshi, Nancy Macwan, Taxak Mistry, Digvijaysinh Mahida

Information Technology, Sigma Institute of Engineering, Vadodara, Gujarat, India

ABSTRACT

Now days, most research works are on sentiment classification, here we discuss about text mining and the technique used in text mining, web mining, video mining sentiment classifications, sentiment analysis. And further we are discussing about Natural language processing which support multiple languages. This is the technique which is most generally used in analyzing the unstructured data sets. Various methodologies are being used but the NLP is the easiest and commonly used among them.

Keywords: Web Mining, Text Mining, Natural Language Processing, Artificial Intelligence, Sentiment classification, NLP

I. INTRODUCTION

Here, we are discussing on the web data mining and its techniques like text mining and natural language processing. We have examined various papers on web mining, where the focus of discussion is about, what is text mining and how the techniques are being used. And our main content for this review paper is all about the natural language processing, its techniques and all about it.

Text mining is also referred to as text analytics. It is a process of analysing or structuring the large collection of data or written resources & converting them to the relatable or relevant information. Text mining is being performed through various analytics techniques like Natural Language Processing. Natural Language Processing is nothing but the area how the computer and human interacts with each other. It is an area or a study of computer science and artificial intelligence. Natural Language Processing works as method to mine the web data. Text mining is sometime referred as web search processes but it is actually different. In web search the specific keyword is used and we get the information which is already known by us. Instead of that in text mining, we get the information, which is relevant for what we are seeking to get. The goal of text mining is to transform the relevant information to the data or data sets.

II. Natural Language Processing

Natural Language Processing is a field of computer science, which includes manipulation of human language and understanding the computer system [1]. We can say that it the way for computer system to understand and analyse the human language and their meanings and transform and achieve some useful information relevant to that. For deciphering the ambiguities in the human language, NLP uses variety of methodologies. It includes various methods like disambiguation, Part-of-speech tagging, automated summarization, relations extraction, as well as disambiguation and natural language understanding and recognition [2].

Any NLP software needs to have a consistent knowledge based such as a detailed thesaurus, data sets for linguistic rules and grammatical rules, an otology, and up-to-date entities. A developer can organize and structure knowledge to perform tasks such as speech recognition, automatic translation and summarization, relationship extraction and topic segmentation, just by utilizing NLP [3].

In general we are using NLP in various ways which we are not aware of. Like NLP is being used to analyse the behaviour of human on the bases of its social media activities. Personal assistance in our phone that translates the information relevant to our data [4]. There are many more ways we are using NLP in our day-to-day life.

III. Algorithms

A. Supervised Learning Techniques:

This technique is most commonly used by any machines to map the input to the output, Where firstly it assumes input as X and output as Y and the mapping function can be given by

$$Y = f(X)$$

B. Unsupervised Technique:

In this technique we have input data and there is no corresponding output available. It can be further grouped into clustering and association problem [6].The best examples of it are K-means for clustering problem and apriori algorithm for association rule learning problem.

C. Semi-supervised Technique:

In that we have large amount of data and some data is labeled. It can use to discover and learn the structure in the input variables [7]. The good example is a photo archive where some of images are labelled, and the majority are unlabelled.

IV. How Algorithms are being used

It typically based on various machine learning algorithms. Instead of using various large data rules, it relies on automatic learning these rules for analysing the examples.

D. To Generate keyword tags Automatically

It is a technique, which uses Auto tag in generation of topics contained with the body of text. It majorly find outs the topic contained in the text or a large collection of text.[8] The generation of keyword tags help in identifying the main topic and the focus then shifted to that main topic.

E. To summarize block of text

Summarizer is used to discard all the irrelevant information and only extract the useful and summarize the data sets. The unrequired information is discarded.[9] In English language, there are some words which re being used repeatedly. Like A, An, The, Who, When, Where, What, How, Something etc.

F. Identify the entity extracted

The process, to identify the type of entity being extracted. Like name, place, or person or organization by using the Named Identity Recognition. Examining the entity type which is being extracted for a reason.

G. Chat Bot

Create a chat bot using parsey McParseFace by Google, which is a language parsing deep learning tool that uses point-to-point speech learning tool [10]. Describe the point to point meaning of the text and helps the system generating the relevant information.

H. Sentiment analysis

To identify the string of the text and its meaning, from very positive to very negative. The hidden meanings used in the text. How it describes the sentiments of the text. What is the hidden means to this text.

I. Reduce words

Reduce words to their roots so that they can only have the meaningful and important information. Reduction of un-wanted words. Which are just increasing the size of the text and their absence create no fuss. Use PorterStemmer to reduce words to their roots. Also break it in a tokens using tokenizer. Tokenizing also helps in a good way as it decreases the size of text and contains the useful and most importantly the relevant and required information from the large collection of data sets.

V. CONCLUSION

In this paper we have discussed and studied about the text mining and the techniques used in text mining, web mining, video mining sentiment classifications and sentiment analysis and also studied on the natural language processing in different languages. We learned different algorithms on NLP which are supervised technique, semi-supervised technique and unsupervised technique. We also learned how these algorithm works, how it generates keyword texts automatically, text block summarization works, identifies entity to extract, how chat-bot works, sentiment analysis is performed and to reduce words to give accurate data. Overall we studied the whole natural language process.

VI. REFERENCES

[1] Jamie Murdoch, Peter J. Liu, Bin Yu "Beyond word importance: using contextual decompositions to extract interactions from LSTMs." ICLR 2018

[2] Navdeep Jaitly, Richard Sproat "An RNN Model of Text Normalization" Interspeech 2017 (2017)

[3] Kedar Dhamdhere, Kevin McCurley, Mukund Sundararajan, Qiqi Yan, Ralfi Nahmias " Analyza: Exploring Data with Conversation" 2017

[4] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, Kamps,"Avoiding Your Teacher's Mistakes: Training Neural Networks with Controlled Weak Supervision",2017

[5] Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan Mcdonald, Slav Petrov" Natural Language Processing with Small Feed-Forward Networks" ,2017

[6] Peter J. Liu, Mohammad Ahmad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, Noam Shazeer,"Generating Wikipedia by Summarizing Long Sequences",2018

[7] Fadi Biadsy, Michael Alexander Nirschl, Min Ma, Shankar Kumar ,"Approaches for Neural-Network Language Model Adaptation",2017

[8] Adams Wei Yu, Hongrae Lee, Quoc V. Le," Learning to Skim Text",2017

[9] Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan Mcdonald, Slav Petrov "Natural Language Processing with Small Feed-Forward Networks",2017

[10] Avneet Pannu, "Artificial Intelligence and its Application in Different Areas" ,2015