International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 4 | ISSN : 2456-3307

Advanced Text Mining & Natural Language Processing

Kamal M[1], Dr. R. Chinnaiyan[2]
[1]PG Scholar , Department of MCA, New Horizon College Of Engineering, Bangalore, Karnataka, India
[2]Professor, Department of MCA, New Horizon College of Engineering, Bangalore, Karnataka, India

ABSTRACT

In the data driven world it is important for the organization to make important business decisions. Analytics helps in achieving the goals, Business analysts and data analyst work together for providing overall insights with historical data that has been generated by the organization. Tackling with the quantitative data has N numbers of methodologies. When it comes with qualitative data business analysts & data analysts face difficulties tackling huge amount of textual data and there are no proper techniques. To overcome these problems there is a techniques called text analytics. Text analytics always deals with textual data, refers to the representation, processing and modeling of the textual data to derive insights. An important component of text analysis is text mining, the process of discovering relationship and interesting patterns in the large text collections and extraction of meaningful information from the unstructured data can either enrich the customer like and dislike and new product sentiment for all of these tasks text analytics plays a vital role in business.

Keywords : Natural Language Processing, Text Mining, NLP

I. INTRODUCTION

The world is over growing with data which might give rise to Big data. When big data is combined with analytics the following tasks can be performed.

- ✓ Determination of faults and failure's in real time and finding the root cause of it.
- ✓ Fraudulent transactions can be detected.
- ✓ Filtering unwanted e-mails and spam detection

At this point text mining helps in better understanding of components and ideas present on the online. The web crawlers scans the texts that are present through the text that might be considered to be important of the words and their specific situation cannot be interpreted.

Text analytics can overcome this such problem for a greater extent. It distinguishes the patterns and connections among the words and text analytics makes it simple in identifying the hidden patterns in big data using statistics for significant insights.

II. STEPS IN TEXT ANALYTICS

A text analysis problem usually consists of three important steps: Parsing, search and retrieval and text mining. Text analysis may also contain of other subtasks such as discourse and segmentation.

1. PARSING

The data that resides on the web such as the xml, html, or word document are in unstructured format which needs to be analyzed. Parsing process helps in achieving it. The process that takes unstructured text and imposes a structure for the further analysis, parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps

2. SEARCH AND RETRIEVAL

Identifying the documents in a corpus is a large collection of texts that might contain specific words, phrases, topics or entities like people, organizations. These search items generally contains items called key terms.Search and retrieval originated from the field of library science and now most commonly used in web search engines

3. TEXT MINING

The term mining are one of the most important aspects of text data. It uses the terms and indexes produced by two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques such as clustering, classification algorithms can be adapted to text mining, clustering can be used the text documents into groups, where each group represents the collection of documents with similar topic. The distance of a document to a centroid represents how closely the document talks about a topic. Classification tasks such as sentiment analysis and spam filtering are being widely used. Text mining may utilize methods and techniques from various fields such as statistical analysis, information retrieval, data mining and also natural language processing. There are few drawbacks While performing the text analysis project we do not have follow the above steps. If the goal is to construct a corpus we would be using the parsing along with one or more text preprocessing techniques such as part-of-speech tagging, named entity recognition, lemmatization, or stemming. The following three steps do not have any sequential. Because one could use parsing to build the data store and choose either for search and retrieval the documents used in the text mining on the entire data to store to gain more insights.

Following are steps in Text Mining

- ✓ Collection of unstructured data
- ✓ Transforming of unstructured
- ✓ Finding hidden pattern
- ✓ Pattern recognition
- ✓ Storage

III. APPLICATIONS OF TEXT MINING

Text mining are being widely used in document analysis and provides analyzation of text documents in the business world.

- ✓ Digital Libraries:huge amount of patterns can be derived from other documents & journals that resides on digital libraries. Text mining helps in accessing trillions of documents that are being present on libraries. Documents extraction can be used in form of audio, image format along with text.
- ✓ Business Intelligence: most of the business firms use text data to derivate patterns and to analyze text data so that it's important for making decisions about competitors. Better the business decisions better will be the performance.
- ✓ Filtering of unwanted and spams in emails: text mining helps in removing and filtering unwanted emails and spams. It also helps in routing the messages and emails for appropriate destinations.
- ✓ Health care and life science: The medical domains contains huge tons of patients information disease, treatment etc. extraction of these information is a hectic and challenging task. Text mining helps in evaluations and effectiveness of medical treatments that shows effective comparison between different disease and symptoms and their course of treatment.

IV. TECHNIQUES IN TEXT MINING

Text mining techniques depends on the requirement and the applications that is being used. The following are the steps

1. EXTRACTION OF INFORMATION

Extraction of vital information from a large container of database, which is needed for decision

making.Information that is related to attributes and entities from different document. These documents can be further be used in analysis includes name of person, organization in which he works and location

2. RETRIEVING OF INFORMATION

Collection of information from various sources. A best example is the search engines that retrieve's the information from world wide web.

3. CATEGORIZING

It is a way used for sorting of documents from predefined steps into different categories. This technique's is being used in several applications such as spam filtering, survey coding, patent filling etc. the main aim is to train the classifier the unknown terms.

4. CLUSTERING

The technique helps to group the documents which are similar. This leads to formation of clusters that contains the one or more documents.

5. SUMMARIZATION

It gives us the overall view of the documents. That describes the entire documents and also contains the points and meanings and information of the complete documents.

6.VISUALIZATION

The text data have been arranged into visual hierarchy. User are provided with data in the form of tables and graphs. So that it can be easily understood.

V. TEXT MINING CHALLENGES

The major challenges that text analysis is the text are not structured format. This might include quasi-structured,semi-structuredand unstructured data. Another important challenges are the text that have the similar meaning or the multiple word can have same meaning. Classification and categorization being used to overcome structure document.

VI. NATURALLANGUAGE PROCESSING

Natural language processing provides an automated way to analyze the text data and to gain the information from the human languages. this approach uses the machine algorithms to extract meaningful information. There are various methods which is being used in the process of NLP such as Tokenization, stemming, Tf-idf, semantic analytics, Disambiguation, Topic models, Word boundaries. Tokenization refers to splitting the words, phrases and idioms, Stemming is done to validate root word, the term Tf-idf Represents term frequency and inverse document frequency, widely being used in information retrieval and text analysis semantic analytics refers to methods of comparing words, phrases, and idioms in a set of documents to extract meaning, Disambiguation Determines meaning and sense of words (context vs. intent), Topic models Discover topics in a collection of documents and word boundaries Determines where one word ends and the other begins these are the term which are being used in the process of NLP. Some of the common approaches for analyzing textual data are Conduct basic text processing by using sample data, Categorize the words in a specific format and tag them all together, classify the text that are relevant, Extract information from the groups of words that is being classified, Analyze sentence structure, Build feature based structure, Analyze the meaning to gain an insights. Nowadays these kind of processing are widely being used in Machine Translation, Speech recognition, and Sentiment Analytics.

VII. APPLICATIONS OF NATURAL LANGUAGE PROCESSING

NLP is termed as the one of the most widely used and future it might become one of the most important technology which help in building the human communication and digital data. some of the important applications are listed below.

- ✓ Machine translation**:** The information which we access is from online. The task of data accessing becomes more important. The challenge is to make the information available for everyone. Machine translation is used to translate one language into another. Google Translate is an example. It uses NLP to translate the input data from one language to another. The NLP uses machine learning algorithm for language translation.
- ✓ Speech recognition**:** the Applications of speech recognition is widely being used in daily aspects it can be any mobile devices or any other system applications such as Cortana. Speech recognition applications take human speech as input data and analyzes it and provides the necessary output**.** It is useful for applications like Siri, Google Now, and Microsoft Cortana.
- ✓ Sentiment Analysis**:** the data that has been generated from all over the social media such as comments, tweets, feedback are in the form of text that might contain all most tons of textual data and these data needs to be analyzed for determining the behavior or attitude of humans the sentiments analysis uses the data mining tasks algorithms.Sentiment analysis is achieved by processing tons of data received from different interfaces and sources. For example, NLP uses all social media activities to find out the popular topic of discussion.

VIII. CONCLUSION

Text analytics and NLP techniques help the users to find meaningful information and provides an insights in large amount of data that is being present on the web and databases. The main aim of using text analytics and NLP extracting information from unstructured data which uses Advanced text mining with algorithms That makes it efficient to use. There are many techniques as discussed above and these the developers or researchers must select an appropriate techniques who use it.

IX. REFERENCES

[1]. Amreen kausar, Rekha B.S "Application of Text mining in effective document analysis vol.6 Issue 04, April 2017.

[2]. M.Hu and B.liu, "mining and summarization, International conference on knowledge Discovery and data mining,pp.168-177,2004