iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45408

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Text Mining Using Natural Language Processing

Karan Nellimella[1], Pranav Thorat[2], Suraj Uikey[3], Rohan Godage[4], Shubhra Mathur[5]
[1, 2, 3, 4, 5]*Student Department of Computer Engineering Shree Ramchandra Collage Of Engineering*

I. INTRODUCTION

The amount of data is increasing exponentially every day. Almost all types of institutions, organizations, and industries store their data electronically. A huge amount of texts flow over the Internet in the form of digital libraries, repositories, and other textual information such as blogs, social media networks, and email. Determining the right patterns and trends to extract valuable knowledge from this large amount of data can be a daunting task. Traditional data mining tools cannot process textual data because it takes hours and effort to extract the information. Text mining is the process of extracting interesting and important patterns for exploring knowledge from text data sources. Text mining is an interdisciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics. Figure 1 shows the interaction of text mining Venn diagram with other fields. You can apply several text mining techniques such as summarization, classification, and clustering to extract knowledge. Text mining processes natural language text stored in semi-structured and unstructured formats. Text mining technology is continuously applied in industry, science, web applications, the Internet, and other areas. Text mining for opinion mining, feature extraction, sentiment, prediction, and trend analysis in application areas such as search engines, customer relationship management systems, filter e-mail, product proposal analysis, fraud detection, and social media analysis.

Text mining is based on a variety of advanced techniques derived from statistics, machine learning, and linguistics. Text mining uses technology to find patterns and trends in "unstructured data", more commonly due to, but not limited to, textual information. The goal of text mining is to be able to process large text data and extract "high quality" information. This helps to provide insight into the specific scenario to which text mining applies. Text mining has numerous applications such as extraction of concepts, sentiment analysis, and summarization.

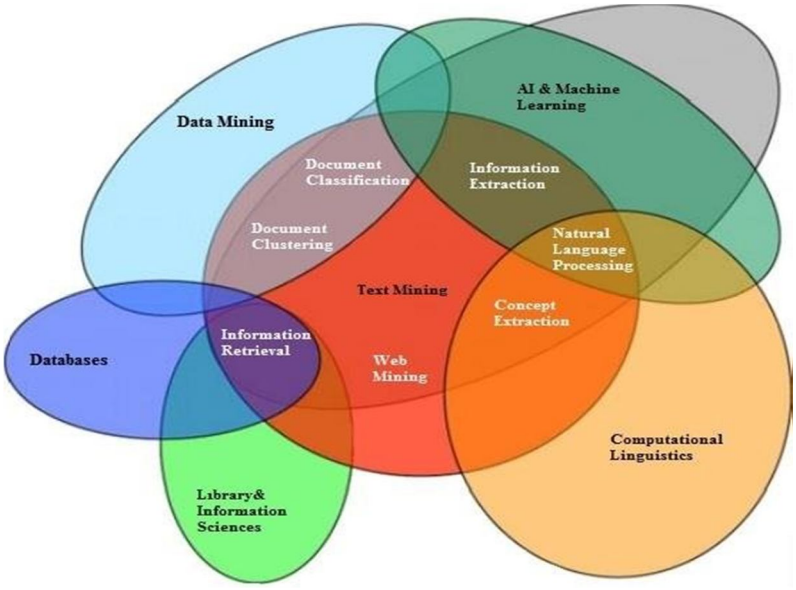


Fig. Venn Diagram Of Text Mining

II. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a field of computer science and artificial intelligence that deals with natural language interactions between computers and humans. The ultimate goal of NLP is to enable computers to understand languages as we do. This is the power behind virtual assistants, speech recognition, sentiment analysis, automated text summarization, machine translation, and more. In this post, I'll explain the basics of natural language processing and elaborate on some of its techniques. You will also learn how NLP benefits from recent advances in deep learning.

Natural Language Processing (NLP) is an interface between computer science, linguistics, and machine learning. This area focuses on the communication between computers and humans in natural languages, and NLP aims to help computers understand and generate human languages. NLP technology applications include voice assistants such as Amazon's Alexa and Apple's Siri, as well as machine translation and text filtering. Human language is special for several reasons. It is specially built to convey the meaning of the speaker. This is a complex system, but toddlers can learn fairly quickly.

Another notable point of human language is that it's all about symbols. According to Chris Manning, a professor of machine learning at Stanford University, it is a discrete and symbolic category of signalling systems. This means that the same meaning can be conveyed in different ways (for example, through language, gestures, symbols, etc.).

III. PROBLEM DEFINITION

In the field of text mining, we try to extract useful information from unstructured text data by identifying and searching for paths of interest.

Text mining, commonly used in knowledge-based organizations is the process of exploring a large collection of documents to discover new informationor answer specific research questions. Text mining identifies facts, relationships, and claims that are buried in most of the big data in text. When this information is converted to a structured format and can be analyzed or displayed directly in grouped HTML tables, mind maps, diagrams, etc., text mining uses a variety of text processing methods, it is one of the most important of them. Natural Language Processing (NLP).

IV. TEXT MINING INFRASTRUCTURE IN R

Text mining covers a wide range of theoretical approaches and methods that have one thing in common: text as input information. This allows for more sophisticated formulas such as "Using a large online collection of texts to discover new facts and trends about the world itself" (Hearst 1999), from classic data mining extensions to text.

Various definitions are possible up to the conversion. In general, text mining is an interdisciplinary field of activity between data mining, linguistics, computational statistics, and computer science.

Standard techniques are text classification, text clustering, ontology and classification construction, document summarization, and potential corpus analysis. In addition, techniques from related disciplines such as Information Retrieval are often used.

The classic application of text mining (Weiss et al. 2004) comes from the data mining community such as document clustering (Zhao and Karypis 2005b, a; Boley 1998; Boley et al. 1999) and document classification (Sebastiani 2002).

To do. In both cases, the idea is to convert the text into a structured format based on the frequency of terms before applying standard data mining techniques. A typical application for document clustering is grouping documents from news articles or information services (Steinbachetal. 2000), Text Classification Method Example: B. Automatic Labeling of Documents in Email Filters and Business Library (Miller 2005). Certain distance measures such as cosine (Zhao and Karypis 2004; Strehl et al. 2000) play an important role, especially in relation to clustering.

RStudio primary purpose is to create free and open-source software for data science, scientific research, and technical communication. This allows anyone with access to a computer to participate freely in a global economy that rewards data literacy; enhances the production and consumption of knowledge; and facilitates collaboration and reproducible research in science, education, and industry. We spend over 50% of our engineering resources on open-source software development and provide extensive support to the open-source data science community.

Code-Friendly: Real-world data science problems are complex, and not easily solved with black box solutions. Code is vital for communication and reuse.

Mitigate Vendor Dependency: Unlike the exclusively proprietary software sold by traditional scientific and technical computing companies, RStudio provides core productivity tools, packages, protocols, and file formats as open-source software so that customers aren’t overly dependent on a single vendor.

Modular Platform for impactful data science: Unlike monolithic data science platforms that require a massive investment upfront in licensing costs and process changes before any real value is delivered, we provide a modular platform for open-source data science.

Scalable, enterprise-friendly, and production-ready:

Unlike fragmented open-source tooling that can be difficult to scale and maintain, our modular platform can scale to large numbers of users and large amounts of data, and integrate into existing enterprise systems, platforms, standards, and processes.

V. LIMITATIONS

Text mining is looking for patterns in text, as data mining can be broadly described as looking for patterns in data. However, the two superficial similarities hide the actual difference. Data mining can be more fully characterized as, which extracts implicit, previously unknown, and potentially useful information from the data. Information is included in the input data. The information is hidden and unknown and can hardly be extracted without the use of automatic data mining techniques. In text mining, on the other hand, the information to be extracted is clearly stated in the text. It's not hidden at all, and most authors put a lot of effort into expressing themselves clearly and clearly. From a human point of view, it was "previously unknown" that the restrictions of staff would make it impossible for humans to read the text themselves. The problem, of course, is that the information is not formulated in a way that is suitable for automatic processing. Text mining strives to output Text in a format suitable for direct computer consumption, without the need for human intervention.

VI. METHEDOLOGIES

Traditionally, numerous techniques have been developed to solve the text mining problem. This is nothing more than retrieving relevant information according to the user's request. According to information gathering, four main methods are used

A. Term Based Method (TBM)

The terms in the document are words that have a semantic meaning. The term-based method has the advantages of analysing documents based on terms, efficient computing power, and a well-designed theory for term weighting. These technologies have emerged from the information retrieval and machine learning communities over the last few decades. The concept-based method has the problem of ambiguity and synonyms. Ambiguity means that a word has multiple meanings, and synonym means that multiple words have the same meaning. The semantic meaning of many of the terms found is uncertain as to what the user wants. Information retrieval has provided many concept-based methods to solve this challenge.

B. Phrase-Based Based Method

The phrase carries more semantics like information and is less ambiguous. In the phrase-based method, the document is analyzed on a phrase basis as phrases are less ambiguous and more discriminative than individual terms. The likely reasons for the daunting performance include:

1) Phrases have inferior statistical properties to terms,
2) They have a low frequency of occurrence, and
3) Large numbers of redundant and noisy phrases are present among them

C. Concept-Based Based Method

Conceptually, concept-based terms are analyzed at the sentence and document level. Text mining technology is mainly based on the statistical analysis of words and phrases. Statistical analysis of term frequency captures the meaning of words without documents. Two terms can have the same frequency in the same document, but that means that one term contributes better than the other. New concept-based mining is being introduced because of the need for more emphasis on terms that capture text semantics. This model consisted of three components. The first component analyses the semantic structure of the sentence. The second component builds a conceptual 4 ontology graph (COG) for describing semantic structures, and the last component extracts higher-level concepts based on the first two components, building a feature vector using a standard vector space model. Concept-based models can effectively distinguish between non-essential terms and meaningful terms that explain the meaning of a sentence. Concept-based models typically rely on natural language processing technology. Feature selection is applied to the concept of queries to optimize the representation and remove noise and ambiguity.

D. Pattern Taxonomy Method

Patterns analyze classification documents on a pattern basis. The pattern can be structured into a classification method using the is-a relationship. Pattern mining has been extensively studied in the data mining community for many years. Patterns can be discovered through data mining techniques such as association rule mining, common item set mining, sequential pattern mining, and closed pattern mining. It is difficult and ineffective to use the knowledge (patterns) found in the field of text mining. This is due to the lack of support for some of the highly specific useful long patterns (i.e., infrequent problems). Not all common short patterns are useful, so they are known as pattern misunderstandings and reduce performance.

The research paper proposes an effective pattern recognition technique to overcome the problems of low frequency and misunderstanding of text mining. The technique based on pattern uses two processes: pattern expansion and pattern evolution. This technique improves the patterns found in text documents. Experimental results show that pattern-based models perform better than concept-based models as well as other pure data mining-based methods and concept-based models.

VII. TEXT CLUSTERING

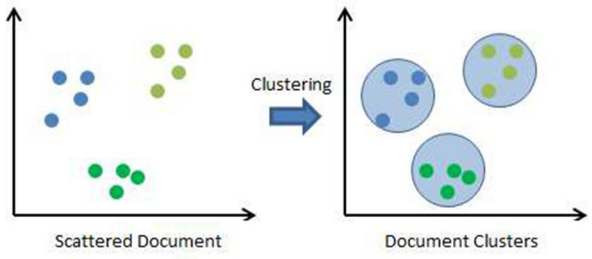


Fig. Text Clustering

Text classification or text classification is a matter of library science, information science, and computer science. The task is to assign text to one or more classes or categories. This can be done "manually" (or "intelligently") or with an algorithm. Intellectual classification of texts is primarily in the field of library science, and algorithmic classification of texts is primarily in information science and computer science. However, because of the overlapping issues, there are interdisciplinary studies on text classification. The classified texts are texts, images, music, and so on. Each type of text has its own classification problem. Unless otherwise stated, text classification is included. Text can be categorized by subject or other attributes (for example, text type, author, year of publication, etc.). For the rest of this article, only the subject classification will be considered. The text subject classification has two main philosophies: a content-based approach and a query-based approach.

Text clustering is the task of grouping a set of unlabelled text so that texts in the same cluster are more similar to each other than texts in other clusters. The text clustering algorithm processes the text and determines if a natural cluster (group) is present in the data.

VIII. ALGORITHOMS

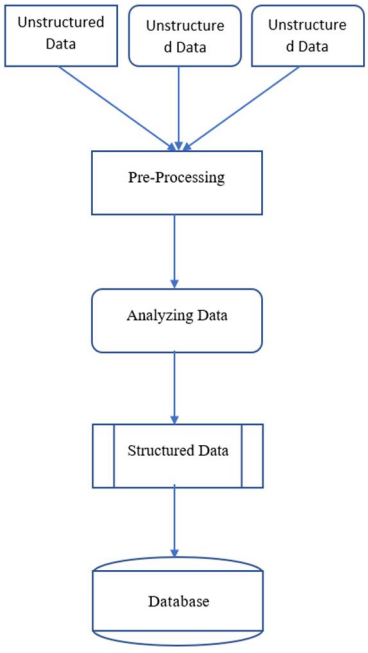


Fig. Algorithm

Unstructured data is information that cannot be stored in a traditional relational database or RDBMS because it is not organized according to a particular data model or schema. Text and multimedia are two common types of unstructured content. Like email messages, videos, photos, web pages, and audio files, many business documents are unstructured. Unstructured data stores contain a wealth of information that you can use as the basis for your business decisions. However, unstructured data has historically been very difficult to analyse.

Data processing occurs when data is collected and transformed into available information. It is usually performed by a data scientist or team of data scientists, but it is important that the data is processed correctly so that it does not adversely affect the final product or data output. Data processing begins with raw data, transforms it into a more readable format (graphics, documents, etc.), interprets it on a computer, and provides the format and context that users need to use.

The benefits of data processing are by no means limited to large enterprises. In fact, small businesses can have great benefits for themselves.

Data analysis is the process of investigating, cleaning, transforming, and modelling data with the goal of discovering useful information, drawing conclusions, and assisting decision making. Data analysis has multiple aspects and approaches, including different methods under different names and used in different areas of business, science, and social sciences. In today's business world, data analytics play a role in making more scientific decisions and enabling organizations to operate more effectively.

Structured data have a well-defined structure that helps in easy storage and access of data. Data can be indexed based on text string as well as attributes. This makes search operation hassle-free. Business Intelligence operations such as Data warehousing can be easily undertaken.

IX. CONCLUSION

We introduced a new framework for text mining applications in R via the tm package. It offers functionality for managing text documents, abstracts the process of document manipulation, and eases the usage of heterogeneous text formats in R. The package has integrated database backend support to minimize memory demands. Advanced metadata management is implemented for collections of text documents to alleviate the usage of large and metadata-enriched document sets. With the package ships native support for handling the Reuters-21578 data set, the Reuters Corpus Volume 1 data set, Gmane RSS feeds, e-mails, and several classic file formats (e.g., plain text, CSV text, or PDFs).

The data structures and algorithms can be extended to fit custom demands since the package is designed in a modular way to enable easy integration of new file formats, readers, transformations, and filter operations. tm provides easy access to pre-processing and manipulation mechanisms such as whitespace removal, stemming, or conversion between file formats (e.g., Reuters to plain text). Further, a generic filter architecture is available to filter documents for certain criteria or perform a full-text search. The package supports the export from document collections to term-document matrices which are frequently used in the text mining literature.

This allows the straightforward integration of existing methods for classification and clustering. tm already supports and covers a broad range of text mining methods, by using available technology in R but also by interfacing with other open-source tool kits like Weka or open NLP offering further methods for tokenization, stemming, and sentence detection, and part of speech tagging. Nevertheless, there are still many areas open for further improvement, e.g., with methods rather common in linguistics, like the latent semantic analysis. We are thinking of better integration of tm with the lsa package. Another key technique to be dealt with in the future will be the efficient handling of very large term-document matrices.

In particular, we are working on memory-efficient clustering techniques in R to handle high dimensional sparse matrices as found in larger text mining case studies. With the ongoing research efforts in analyzing large data sets and using sparse data structures, tm will be among the first to take advantage of new technology. Finally, we will keep adding reader functions and source classes for popular data formats.

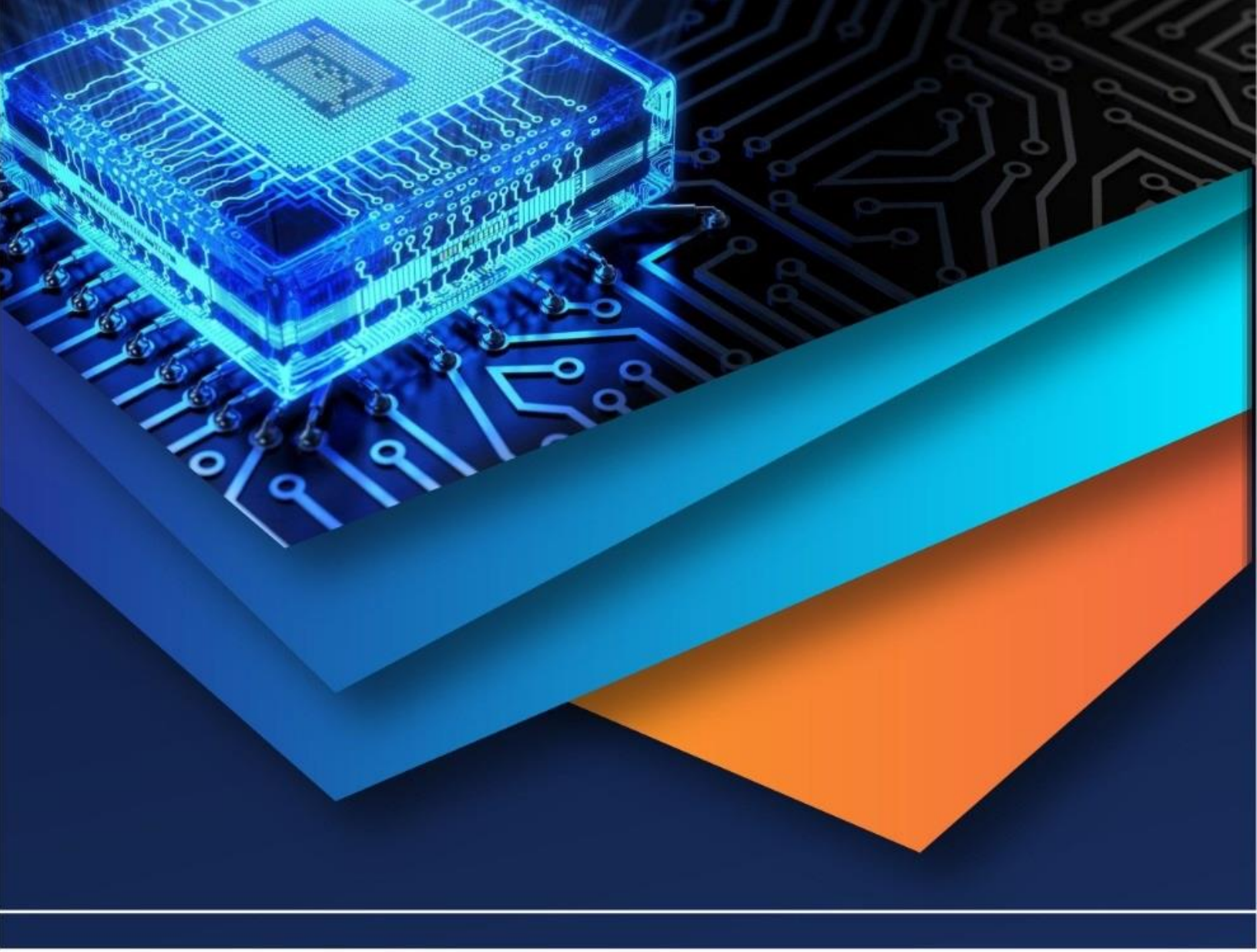
X. ACKNOWLEGEMENT

We are sincerely thankful to Dr. A. D. DESAI Principal of Shree Ramchandra College of Engineering Lonikand Pune, for providing us the opportunity to write the research paper in the form of desertion on this topic "Text Mining Using NLP". We thankful to Prof. Shubhra Mathur for guiding us into every stage of this research paper.

Without her support it would have been difficult for me to prepare the paper so meaningful and interesting. We are also thankful for the teaching and non-teaching staff. This project has helped me acknowledge the business and market aspects in the future endurance.

REFERENCES

[1] Asuncion A, Newman D (2007). "UCI Machine Learning Repository." URL http://www. ics.uci.edu/~mlearn/MLRepository.html.
[2] Bates D, Maechler M (2007). Matrix: A Matrix Package for R. R package version 0.999375-2, URL http://CRAN.R-project.org/package=Matrix.
[3] Berry M (ed.) (2003). Survey of Text Mining: Clustering, Classification, and Retrieval. Springer-Verlag. ISBN 0387955631.
[4] Bierner G, Baldridge J, Morton T (2007). "OpenNLP: A Collection of Natural Language Processing Tools." URL http://opennlp.sourceforge.net/.
[5] Boley D, Gini M, Gross R, Han EH, Karypis G, Kumar V, Mobasher B, Moore J, Hastings K (1999). "Partitioning-based Clustering for Web Document Categorization." Decision Support Systems, 27(3), 329–341. ISSN 0167-9236. doi:10.1016/S0167-9236(99
[6] Feinerer I (2007a). openNLP: OpenNLP Interface. R package version 0.1, URL http: //CRAN.R-project.org/package=openNLP.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY