# TEXT MINING: OPEN SOURCE TOKENIZATION TOOLS – AN ANALYSIS

Dr. S.Vijayarani[1] and Ms. R.Janani[2]
[1]Assistant Professor, [2]Ph.D Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore.

## ABSTRACT

*Text mining is the process of extracting interesting and non-trivial knowledge or information from unstructured text data. Text mining is the multidisciplinary field which draws on data mining, machine learning, information retrieval, computational linguistics and statistics. Important text mining processes are information extraction, information retrieval, natural language processing, text classification, content analysis and text clustering. All these processes are required to complete the preprocessing step before doing their intended task. Pre-processing significantly reduces the size of the input text documents and the actions involved in this step are sentence boundary determination, natural language specific stop-word elimination, tokenization and stemming. Among this, the most essential and important action is the tokenization. Tokenization helps to divide the textual information into individual words. For performing tokenization process, there are many open source tools are available. The main objective of this work is to analyze the performance of the seven open source tokenization tools. For this comparative analysis, we have taken Nlpdotnet Tokenizer, Mila Tokenizer, NLTK Word Tokenize, TextBlob Word Tokenize, MBSP Word Tokenize, Pattern Word Tokenize and Word Tokenization with Python NLTK. Based on the results, we observed that the Nlpdotnet Tokenizer tool performance is better than other tools.*

## KEYWORDS:

*Text Mining, Preprocessing, Tokenization, machine learning, NLP*

## 1. INTRODUCTION

Text mining is used to extract interesting information, knowledge or pattern from the unstructured documents that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques [5]. It is the analysis of data which contained in natural language text. If the text mining techniques are used to solve business problems then it is called as text analytics. Figure 1 shows the general steps of text mining.

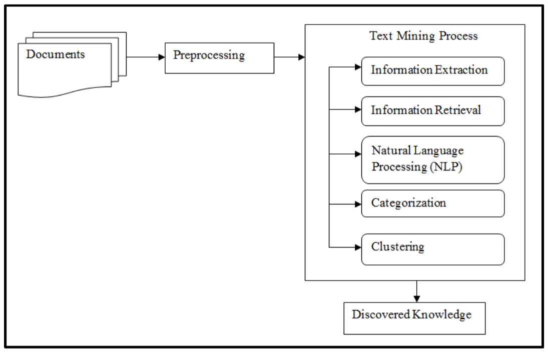


Fig 1: Text Mining Steps

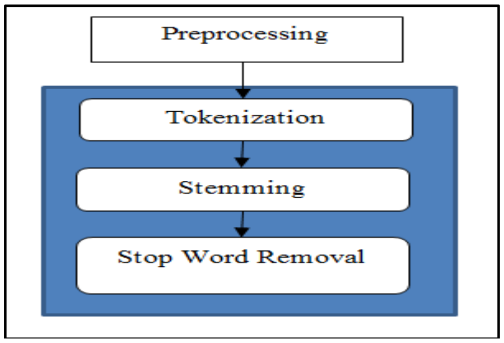


Fig 2: Preprocessing Operations

Figure 2 shows the various operations performed during preprocessing. Stemming is the process for reducing modulated words to their word stem, root or base. Stop word is word which is filtered out before or after the processing of natural language text [6]. Tokenization is the process of breaking a stream of textual content up into words, terms, symbols, or some other meaningful elements called tokens. In this research work, we have analyzed the performance of seven open source tokenization tools. The list of tokens turns into input for in additional processing including parsing or text mining. Tokenization is beneficial both in linguistics and in computer science, where it forms the part of lexical analysis [1]. Generally, the process of tokenization occurs at the word level. But, it is sometimes tough to define what is meant by a "word". Regularly a tokenizer commits on simple heuristics, for example:

- Punctuation and whitespace may or may not be included in the resulting list of tokens.
- All contiguous strings of alphabetic characters are part of one token; in the same way with numbers.
- Tokens are separated by the way of whitespace characters, such as a space or line break, or by punctuation characters.

Tokenization process is mainly complicated for languages written in ‘scriptio continua’ which reveals no word limits such as Ancient Greek, Chinese, or Thai [3]. A Scriptio continuum, also known as scriptura continua or scripta continua, is a style of writing without any spaces or other marks in between the words or sentences. The main use of tokenization is to identify the meaningful keywords [2]. The disadvantage of tokenization is difficult to tokenize the document without any whitespace, special characters or other marks.

## 1.1 Need for Tokenization

Generally textual data is only a set of characters at the initial stage. All processes in text analysis will require the words which is available in the data set. For this reason, the requirement for a parser is a tokenization of documents [1]. This process may trivial as the text is already stored in machine-readable formats. But, some problems are still left, for example the punctuation mark removal, end of line hyphen removal [5]. But characters like brackets, hyphens, etc. are processing well. Tokenizers also provide the reliability for the documents.

## Example

Input: Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use.

Output: Data, Mining, is, the, process, to, extract, hidden, predictive, information, from, database, and, transform, it, into, understandable, structure, for, future, use.

The paper is organized as follows. Section 2 gives the details about the tokenization tools. Section 3 explains the performance analysis of different open source tokenization tools and conclusion of this work is given in section 4.

## 2. Tools for Tokenization

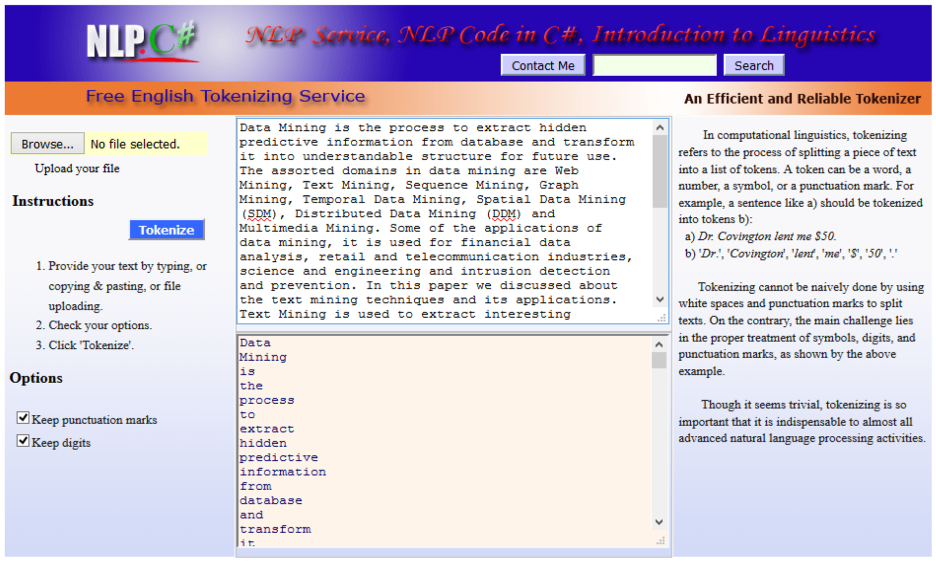
For tokenize a document, many tools are available. The tools are listed as follows,

- Nlpdotnet Tokenizer
- Mila Tokenizer
- NLTK Word Tokenize
- TextBlob Word Tokenize
- MBSP Word Tokenize
- Pattern Word Tokenize
- Word Tokenization with Python NLTK

In order to perform the analysis, we provide an input document and this document are processed by these tools and the output produced by these tools is considered for analysis. Each tool has produced different outputs for the same document input.

### 2.1 Nlpdotnet Tokenizer

In computational linguistics, Nlpdotnet is a Python library for Natural Language Processing tasks which is based on neural networks. Currently, it performs tokenization, part-of-speech tagging, semantic role labelling and dependency parsing. Though it seems trivial, tokenizing is so important that it is vital to almost all advanced natural language processing activities [8]. Figure 3 shows the input and output of Nlpdotnet tokenizer.


NLP.C#
NLP Service, NLP Code in C#, Introduction to Linguistics
Contact Me
Search
Free English Tokenizing Service
An Efficient and Reliable Tokenizer
Browse...
No file selected.
Upload your file
Tokenize

**Instructions**

1. Provide your text by typing, or copying & pasting, or file uploading.
2. Check your options.
3. Click 'Tokenize'.

**Options**

☑ Keep punctuation marks
☑ Keep digits

```
Data Mining is the process to extract hidden
predictive information from database and transform
it into understandable structure for future use.
The assorted domains in data mining are Web
Mining, Text Mining, Sequence Mining, Graph
Mining, Temporal Data Mining, Spatial Data Mining
(SDM), Distributed Data Mining (DDM) and
Multimedia Mining. Some of the applications of
data mining, it is used for financial data
analysis, retail and telecommunication industries,
science and engineering and intrusion detection
and prevention. In this paper we discussed about
the text mining techniques and its applications.
Text Mining is used to extract interesting
```

```
Data
Mining
is
the
process
to
extract
hidden
predictive
information
from
database
and
transform
it
```

In computational linguistics, tokenizing refers to the process of splitting a piece of text into a list of tokens. A token can be a word, a number, a symbol, or a punctuation mark. For example, a sentence like a) should be tokenized into tokens b):

a) *Dr. Covington lent me $50.*
b) '*Dr.*', '*Covington*', '*lent*', '*me*', '*$*', '*50*', '*.*'

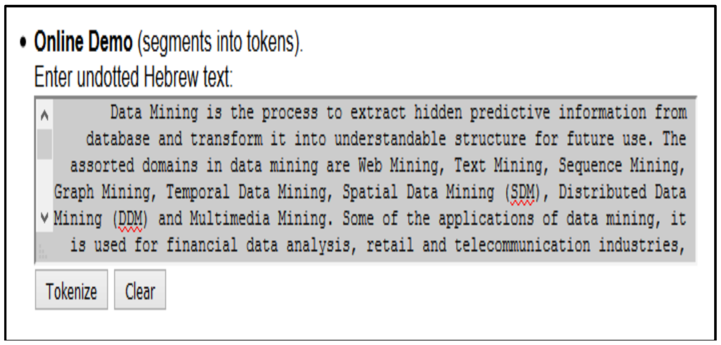
Tokenizing cannot be naively done by using white spaces and punctuation marks to split texts. On the contrary, the main challenge lies in the proper treatment of symbols, digits, and punctuation marks, as shown by the above example.

Though it seems trivial, tokenizing is so important that it is indispensable to almost all advanced natural language processing activities.

Fig 3: Nlpdotnet tokenizer output

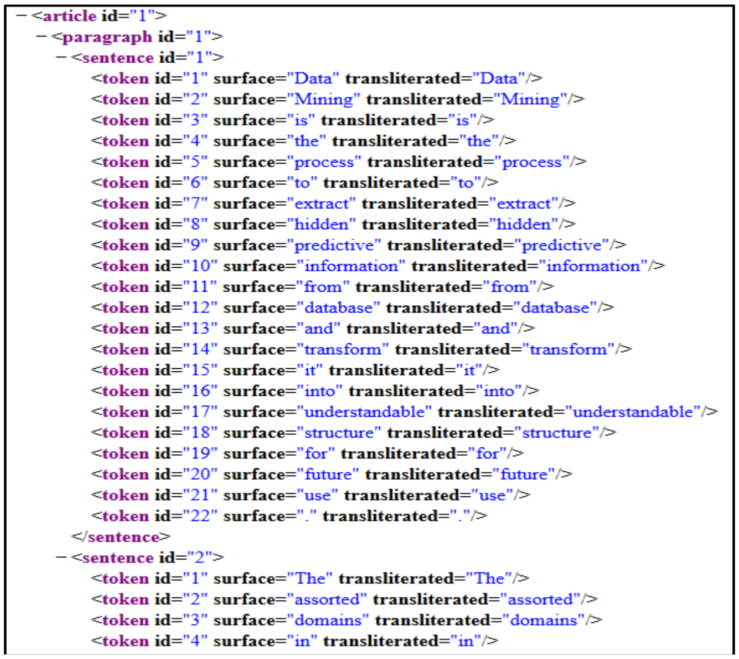
## 2.2 Mila Tokenizer

MILA was developed in 2003 by the Israel Ministry for Science and Technology. Its mission is to create the infrastructure necessary for computational processing of the Hebrew language and make it available to the research community as well as to commercial enterprises. The MILA Hebrew Tokenization Tool divides inputted undotted Hebrew text (right to left) into tokens, sentences, and paragraphs and converts these into XML format. [9]. Figure 4 and 5 shows the input and output of Mila tokenizer.


• Online Demo (segments into tokens).
Enter undotted Hebrew text:
Tokenize
Clear

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries,

Fig 4: Mila Tokenizer Input

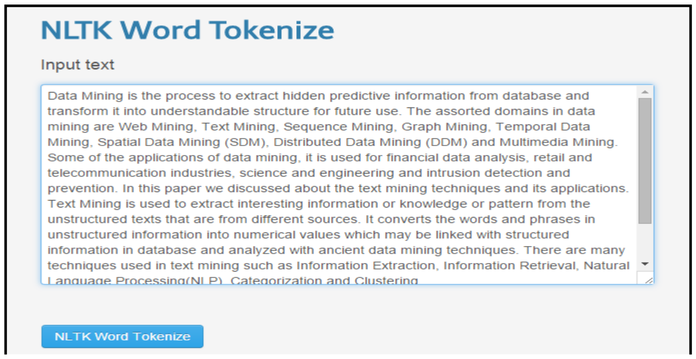


```
– <article id="1">
  – <paragraph id="1">
    – <sentence id="1">
        <token id="1" surface="Data" transliterated="Data"/>
        <token id="2" surface="Mining" transliterated="Mining"/>
        <token id="3" surface="is" transliterated="is"/>
        <token id="4" surface="the" transliterated="the"/>
        <token id="5" surface="process" transliterated="process"/>
        <token id="6" surface="to" transliterated="to"/>
        <token id="7" surface="extract" transliterated="extract"/>
        <token id="8" surface="hidden" transliterated="hidden"/>
        <token id="9" surface="predictive" transliterated="predictive"/>
        <token id="10" surface="information" transliterated="information"/>
        <token id="11" surface="from" transliterated="from"/>
        <token id="12" surface="database" transliterated="database"/>
        <token id="13" surface="and" transliterated="and"/>
        <token id="14" surface="transform" transliterated="transform"/>
        <token id="15" surface="it" transliterated="it"/>
        <token id="16" surface="into" transliterated="into"/>
        <token id="17" surface="understandable" transliterated="understandable"/>
        <token id="18" surface="structure" transliterated="structure"/>
        <token id="19" surface="for" transliterated="for"/>
        <token id="20" surface="future" transliterated="future"/>
        <token id="21" surface="use" transliterated="use"/>
        <token id="22" surface="." transliterated="."/>
      </sentence>
    – <sentence id="2">
        <token id="1" surface="The" transliterated="The"/>
        <token id="2" surface="assorted" transliterated="assorted"/>
        <token id="3" surface="domains" transliterated="domains"/>
        <token id="4" surface="in" transliterated="in"/>
```

Fig 5: Mila Tokenizer output

### 2.3 NLTK Word Tokenize

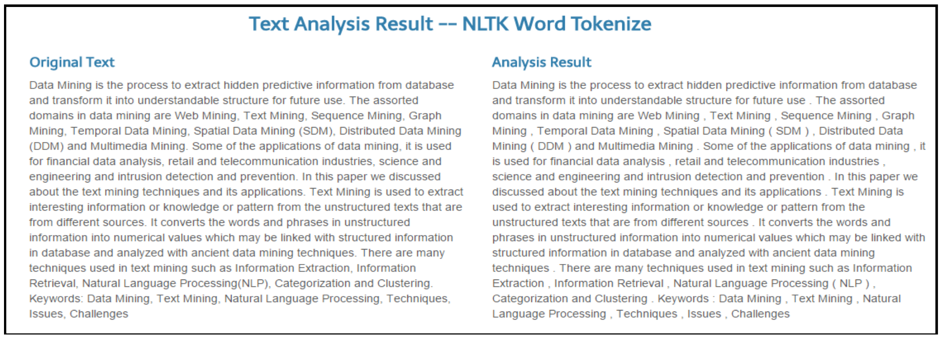
NLTK stands for Natural Language Tool Kit. It is the famous Python Natural Language Processing Toolkit. NLTK is an important platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora (a corpus (plural corpora) is a large and structured set of texts) and lexical resources such as WordNet, along with a set of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning [10]. Figure 6 and 7 shows the input and output of NLTK Word Tokenize.


NLTK Word Tokenize

Input text

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering

Fig 6: NLTK Word Tokenize Input


Text Analysis Result -- NLTK Word Tokenize

Original Text

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering. Keywords: Data Mining, Text Mining, Natural Language Processing, Techniques, Issues, Challenges

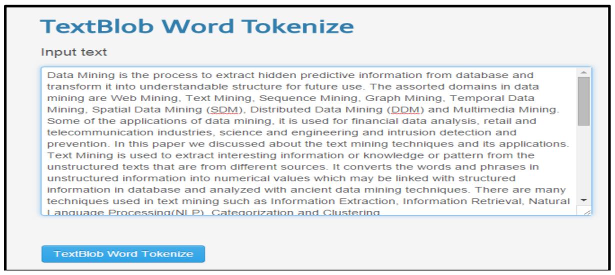
Analysis Result

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use . The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . Some of the applications of data mining , it is used for financial data analysis , retail and telecommunication industries , science and engineering and intrusion detection and prevention . In this paper we discussed about the text mining techniques and its applications . Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources . It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques . There are many techniques used in text mining such as Information Extraction , Information Retrieval , Natural Language Processing ( NLP ) , Categorization and Clustering . Keywords : Data Mining , Text Mining , Natural Language Processing , Techniques , Issues , Challenges

Fig 7: NLTK Word Tokenize output

### 2.4 TextBlob Word Tokenize

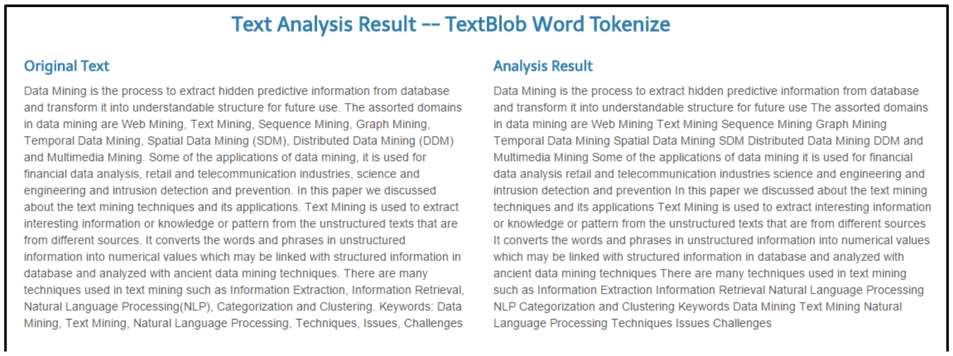
TextBlob is a new python based natural language processing toolkit, which carries the fields like NLTK and Pattern. It provides text mining, text analysis and text processing modules for the python developers. It contains the python library for processing the textual form of data. It provides a simple application program interface (API) for leaping into common natural language processing (NLP) tasks, such as tokenizing, part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation and more [11]. Figure 8 and 9 shows the input and output of TextBlob Word Tokenize.


TextBlob Word Tokenize

Input text

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering

Fig 8: Text Blob Word Tokenize input



Text Analysis Result -- TextBlob Word Tokenize

Original Text

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering. Keywords: Data Mining, Text Mining, Natural Language Processing, Techniques, Issues, Challenges

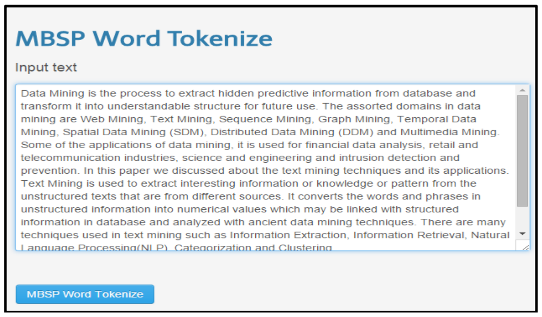
Analysis Result

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use The assorted domains in data mining are Web Mining Text Mining Sequence Mining Graph Mining Temporal Data Mining Spatial Data Mining SDM Distributed Data Mining DDM and Multimedia Mining Some of the applications of data mining it is used for financial data analysis retail and telecommunication industries science and engineering and intrusion detection and prevention In this paper we discussed about the text mining techniques and its applications Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques There are many techniques used in text mining such as Information Extraction Information Retrieval Natural Language Processing NLP Categorization and Clustering Keywords Data Mining Text Mining Natural Language Processing Techniques Issues Challenges

Fig 9: TextBlob Word Tokenize output

## 2.5 MBSP Word Tokenize

MBSP is a text analysis system and it is based on TiMBL (Tilburg Memory-Based Learner) and MBT (Memory Based Tagger) memory based learning applications developed at CLiPS (Computational Linguistics & Psycholinguistics) and ILK (Induction of Linguistic Knowledge). It provides tools for Tokenization and Sentence Splitting, Part of Speech Tagging, Chunking, Lemmatization, Relation Finding and Prepositional Phrase Attachment. The general english version of MBSP has been trained on data from the Wall Street Journal corpus. The Python implementation of MBSP is open source and freely available [12]. Figure 10 and 11 shows the input and output of MBSP Word Tokenize.

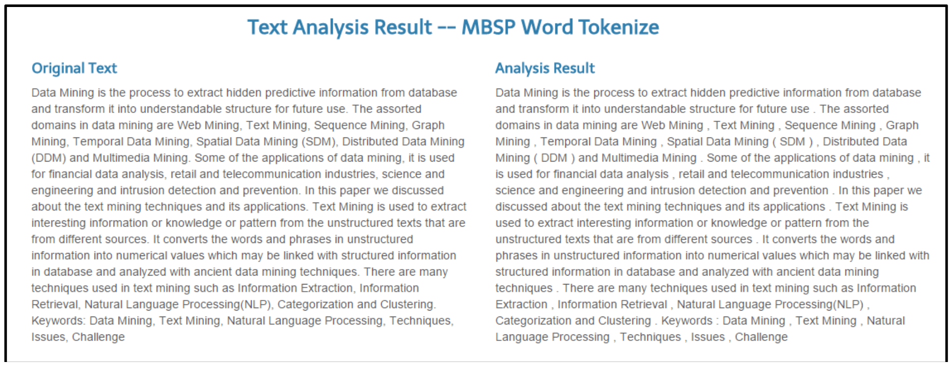


MBSP Word Tokenize

Input text

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering

Fig 10: MBSP Word Tokenize input



Text Analysis Result -- MBSP Word Tokenize

Original Text

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering. Keywords: Data Mining, Text Mining, Natural Language Processing, Techniques, Issues, Challenge

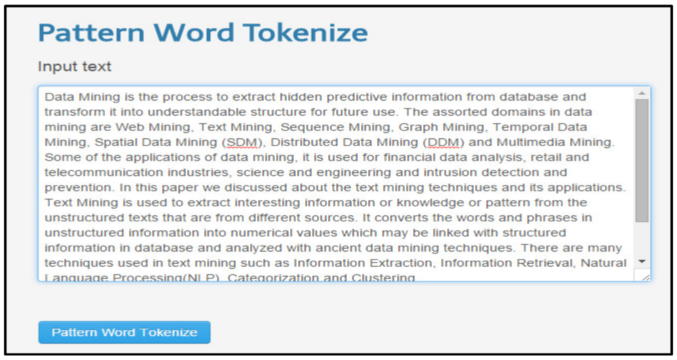
Analysis Result

Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use . The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . Some of the applications of data mining , it is used for financial data analysis , retail and telecommunication industries , science and engineering and intrusion detection and prevention . In this paper we discussed about the text mining techniques and its applications . Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources . It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques . There are many techniques used in text mining such as Information Extraction , Information Retrieval , Natural Language Processing(NLP) , Categorization and Clustering . Keywords : Data Mining , Text Mining , Natural Language Processing , Techniques , Issues , Challenge

Fig 11: MBSP Word Tokenize output

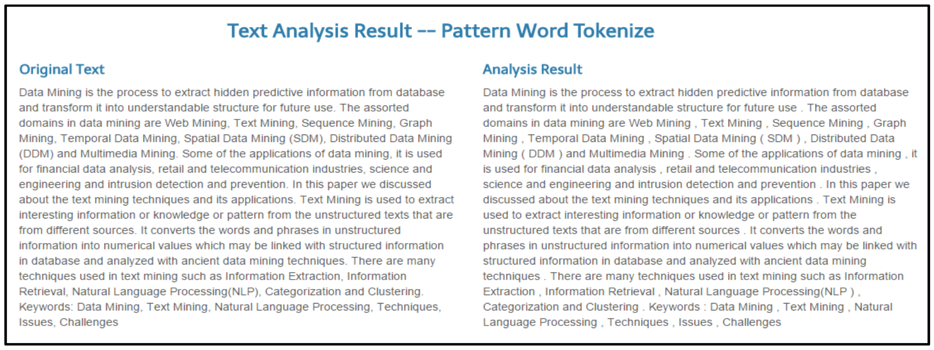
### 2.6 Pattern Word Tokenize

Pattern is a web mining module for the Python programming language. It has tools for data mining (Google, Twitter and Wikipedia API, a web crawler, a HTML DOM parser), natural language processing (part-of-speech taggers, preprocessing, n-gram search, sentiment analysis, WordNet), machine learning (vector space model, clustering, SVM), network analysis and canvas visualization [13].



Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering

Fig 12: Pattern Word Tokenize input



Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use. The assorted domains in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining (SDM), Distributed Data Mining (DDM) and Multimedia Mining. Some of the applications of data mining, it is used for financial data analysis, retail and telecommunication industries, science and engineering and intrusion detection and prevention. In this paper we discussed about the text mining techniques and its applications. Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as Information Extraction, Information Retrieval, Natural Language Processing(NLP), Categorization and Clustering. Keywords: Data Mining, Text Mining, Natural Language Processing, Techniques, Issues, Challenges

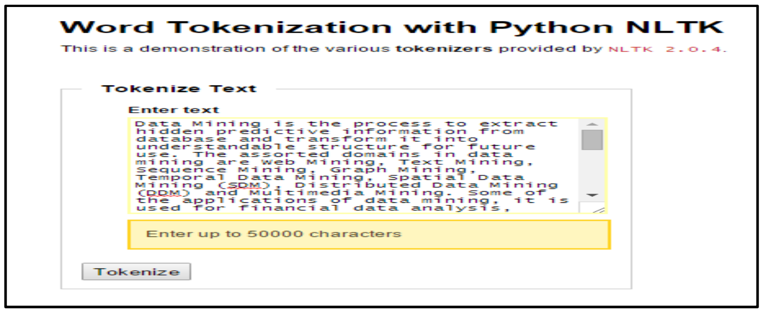
```
Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use . The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . Some of the applications of data mining , it is used for financial data analysis , retail and telecommunication industries , science and engineering and intrusion detection and prevention . In this paper we discussed about the text mining techniques and its applications . Text Mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources . It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques . There are many techniques used in text mining such as Information Extraction , Information Retrieval , Natural Language Processing(NLP ) , Categorization and Clustering . Keywords : Data Mining , Text Mining , Natural Language Processing , Techniques , Issues , Challenges
```

Fig 13: Pattern Word Tokenize output

### 2.7 Word Tokenization with Python NLTK

NLTK provides a number of tokenizers in the module. The text is first tokenized into sentences using the PunktSentenceTokenizer [14]. Then each sentence is tokenized into words using four different word tokenizers:

- TreebankWordTokenizer - This tokenizer uses regular expressions to tokenize text as in Treebank.
- WordPunctTokenizer - This tokenizer divides a string into substrings by splitting on the specified string, which it is defined in subclasses.
- PunctWordTokenizer- This tokenizer divides a text into a list of sentences; by using unsupervised algorithms.
- WhitespaceTokenize - This tokenizer divides text at whitespace.



**Word Tokenization with Python NLTK**

This is a demonstration of the various tokenizers provided by NLTK 2.0.4.

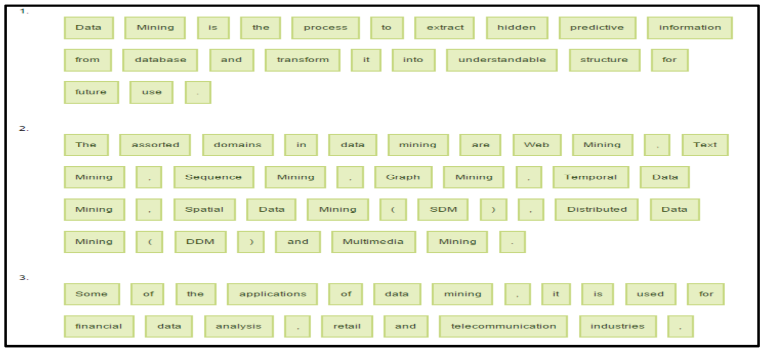
**Tokenize Text**

Enter text

```
Data Mining is the process to extract
hidden predictive information from
database and transform it into
understandable structure for future
use. The assorted domains in data
mining are Web Mining, Text Mining,
Sequence Mining, Graph Mining,
Temporal Data Mining, Spatial Data
Mining (SDM), Distributed Data Mining
(DDM) and Multimedia Mining. Some of
the applications of data mining, it is
used for financial data analysis,
```

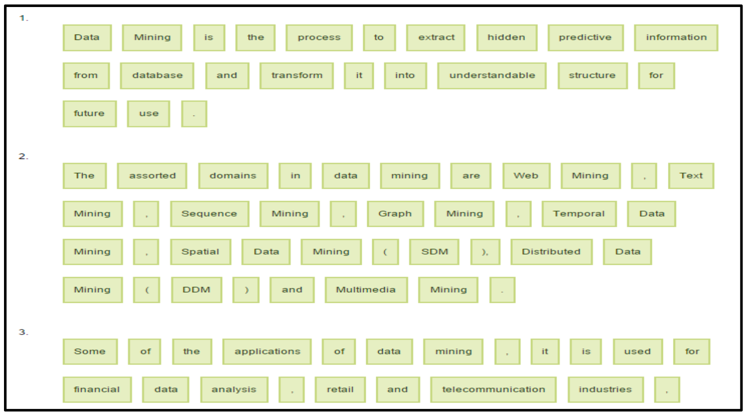
Enter up to 50000 characters

Fig 14: Word Tokenization with Python Input TreebankWordTokenizer



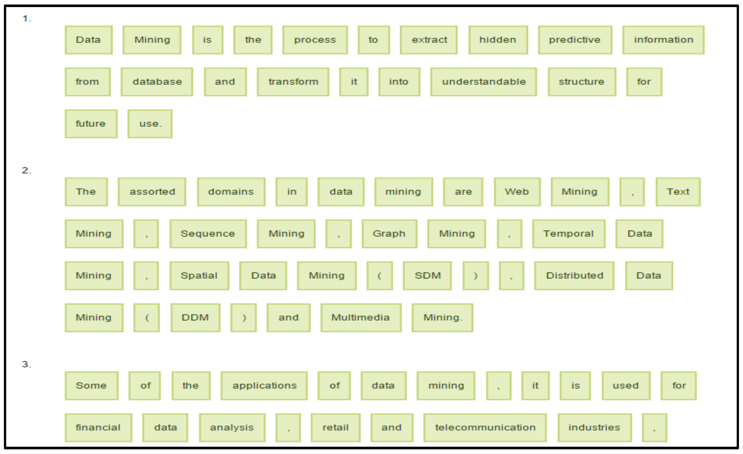
```
1.
Data  Mining  is  the  process  to  extract  hidden  predictive  information
from  database  and  transform  it  into  understandable  structure  for
future  use  .

2.
The  assorted  domains  in  data  mining  are  Web  Mining  ,  Text
Mining  ,  Sequence  Mining  ,  Graph  Mining  ,  Temporal  Data
Mining  ,  Spatial  Data  Mining  (  SDM  )  ,  Distributed  Data
Mining  (  DDM  )  and  Multimedia  Mining  .

3.
Some  of  the  applications  of  data  mining  ,  it  is  used  for
financial  data  analysis  ,  retail  and  telecommunication  industries  .
```

Fig 15: Treebank Word Tokenizer output WordPunctTokenizer



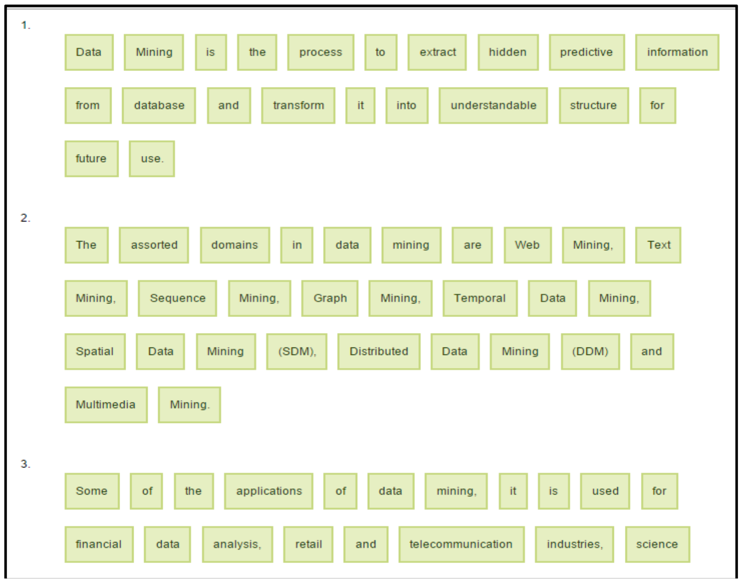
```
1.
Data  Mining  is  the  process  to  extract  hidden  predictive  information
from  database  and  transform  it  into  understandable  structure  for
future  use  .

2.
The  assorted  domains  in  data  mining  are  Web  Mining  ,  Text
Mining  ,  Sequence  Mining  ,  Graph  Mining  ,  Temporal  Data
Mining  ,  Spatial  Data  Mining  (  SDM  ),  Distributed  Data
Mining  (  DDM  )  and  Multimedia  Mining  .

3.
Some  of  the  applications  of  data  mining  ,  it  is  used  for
financial  data  analysis  ,  retail  and  telecommunication  industries  .
```

Fig 16: wordPunct Tokenizer output PunctWordTokenizer



```
1.
Data  Mining  is  the  process  to  extract  hidden  predictive  information
from  database  and  transform  it  into  understandable  structure  for
future  use.

2.
The  assorted  domains  in  data  mining  are  Web  Mining  ,  Text
Mining  ,  Sequence  Mining  ,  Graph  Mining  ,  Temporal  Data
Mining  ,  Spatial  Data  Mining  (  SDM  )  ,  Distributed  Data
Mining  (  DDM  )  and  Multimedia  Mining.

3.
Some  of  the  applications  of  data  mining  ,  it  is  used  for
financial  data  analysis  ,  retail  and  telecommunication  industries  .
```

Fig 17: PunctWord Tokenizer output WhitespaceTokenize

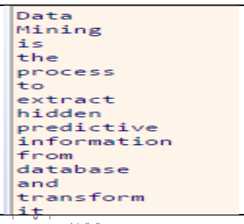
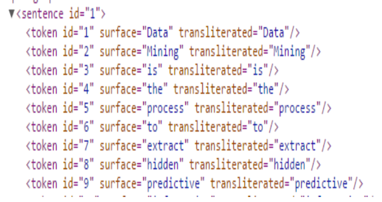
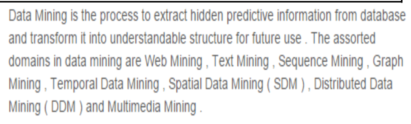
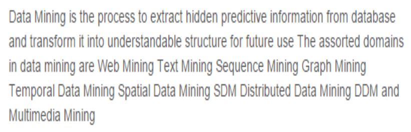
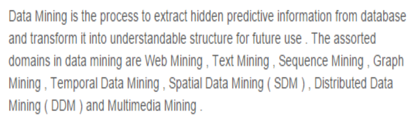
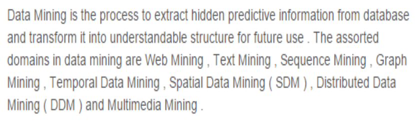
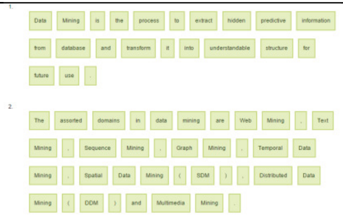


```
1.
Data  Mining  is  the  process  to  extract  hidden  predictive  information
from  database  and  transform  it  into  understandable  structure  for
future  use.

2.
The  assorted  domains  in  data  mining  are  Web  Mining,  Text
Mining,  Sequence  Mining,  Graph  Mining,  Temporal  Data  Mining,
Spatial  Data  Mining  (SDM),  Distributed  Data  Mining  (DDM)  and
Multimedia  Mining.

3.
Some  of  the  applications  of  data  mining,  it  is  used  for
financial  data  analysis,  retail  and  telecommunication  industries,  science
```

Fig 18: Whitespace Tokenizer output

## 3. Performance Analysis

In order to perform the comparative analysis of the tokenization tools, there are two performance measures are used; limitations and output. Limitations describe the uploading file format, language used and maximum number of characters taken for tokenization. Output helps to find how the normal characters, numbers and special symbols are tokenized. Each tool has produced different output for the same input document. Table 1 provides the performance of the seven open source tokenization tools.

| S.No | Name of the Tool | Limitations | Output |
|---|---|---|---|
| 1 | Nlpdotnet Tokenizer | This tool can read only text documents but it can take large number of documents. This output is easy to use and formatted. |  Data<br>Mining<br>is<br>the<br>process<br>to<br>extract<br>hidden<br>predictive<br>information<br>from<br>database<br>and<br>transform<br>it |
| 2 | Mila Tokenizer | In this tool, only we have to enter the input, uploading of the files is not possible. The output is in XML format and it is very difficult to use for future. In this tool we can enter the text from right to left. |  ▼<sentence id="1"><br>&lt;token id="1" surface="Data" transliterated="Data"/&gt;<br>&lt;token id="2" surface="Mining" transliterated="Mining"/&gt;<br>&lt;token id="3" surface="is" transliterated="is"/&gt;<br>&lt;token id="4" surface="the" transliterated="the"/&gt;<br>&lt;token id="5" surface="process" transliterated="process"/&gt;<br>&lt;token id="6" surface="to" transliterated="to"/&gt;<br>&lt;token id="7" surface="extract" transliterated="extract"/&gt;<br>&lt;token id="8" surface="hidden" transliterated="hidden"/&gt;<br>&lt;token id="9" surface="predictive" transliterated="predictive"/&gt; |
| 3 | NLTK Word Tokenizer | In this tool, only we have to enter the input, uploading of the files is not possible. It is a Python based tool. |  Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use . The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . |
| 4 | TextBlob Word Tokenizer | In this tool, only we have to enter the input, uploading of the files is not possible. This tool could not tokenize the special characters. |  Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use The assorted domains in data mining are Web Mining Text Mining Sequence Mining Graph Mining Temporal Data Mining Spatial Data Mining SDM Distributed Data Mining DDM and Multimedia Mining |
| 5 | MBSP Word Tokenizer | In this tool also we have to enter the input, uploading of the files is not possible. It is also a Python based tool. |  Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use . The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . |
| 6 | Pattern Word Tokenize | File uploading is not possible. Input is to be entered and it is a Python based tool. |  Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use . The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . |
| 7 | Word Tokenization with Python NLTK | This tool can take up to 50000 characters at a time. But it gives the best output |  1.<br>Data Mining is the process to extract hidden predictive information from database and transform it into understandable structure for future use .<br>2.<br>The assorted domains in data mining are Web Mining , Text Mining , Sequence Mining , Graph Mining , Temporal Data Mining , Spatial Data Mining ( SDM ) , Distributed Data Mining ( DDM ) and Multimedia Mining . |

## 4. Conclusion

Pre-processing activities plays an important role in the various application domains. The domain specific applications are more proper for text analysis. This is very important and it improves the performance of the information retrieval system. This research paper analyzes the performance of seven open source tokenization tools. From this analysis, some tools have read text documents only and considered the limited number of characters. By analyzing all the measures, compared to other tools Nlpdotnet tokenizer gives the best output. Normally, Tokenization process can be used for the documents which are written in English and French, because these languages use white space to separate the words. Tokenization process could not be used in other languages, for example, Chinese, Thai, Hindi, Urdu, Tamil, etc. This is one of the open research problem for text mining researchers. There is a need to develop a common Tokenization tool for all the languages.

## References

[1] C.Ramasubramanian , R.Ramya, Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

[2] Dr. S. Vijayarani , Ms. J. Ilamathi , Ms. Nithya, Preprocessing Techniques for Text Mining - An Overview, International Journal of Computer Science & Communication Networks,Vol 5(1),7-16

[3] I.Hemalatha, Dr. G. P Saradhi Varma, Dr. A.Govardhan, Preprocessing the Informal Text for efficient Sentiment Analysis, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 1, Issue 2, July – August 2012

[4] A.Anil Kumar, S.Chandrasekhar, Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 5, July - 2012 ISSN: 2278-0181

[5] Vairaprakash Gurusamy, SubbuKannan, Preprocessing Techniques for Text Mining, Conference paper- October 2014

[6] ShaidahJusoh , Hejab M. Alfawareh, Techniques, Applications and Challenging Issues in Text Mining, International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November -2012 ISSN (Online): 1694-0814

[7] Anna Stavrianou, PeriklisAndritsos, Nicolas Nicoloyannis, Overview and Semantic Issues of Text Mining, Special Interest Group Management of Data (SIGMOD) Record, September-2007,Vol. 36, No.3

[8] http://nlpdotnet.com/services/Tokenizer.aspx

[9] http://www.mila.cs.technion.ac.il/tools_token.html

[10] http://textanalysisonline.com/nltk-word-tokenize

[11] http://textanalysisonline.com/textblob-word-tokenize

[12] http://textanalysisonline.com/mbsp-word-tokenize

[13] http://textanalysisonline.com/pattern-word-tokenize

[14] http://text-processing.com/demo/tokenize

**Dr.S.Vijayarani,** MCA, M.Phil, Ph.D., is working as Assistant Professor in the Department of Computer Science, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

**Ms. R. Janani,** MCA., M.Phil., is currently pursuing her Ph.D in Computer Science in the Department of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining, Text Mining and Natural Language Processing.