Basic Idea of Text Mining

Author
* [1]Amulya Sukeerthy NT, Department of Computer Science and Engineering, Presidency University
[2]Bhavana G, Department of Computer Science and Engineering, Presidency University

Abstract

Individuals and organizations generates tons of raw data everyday, among the existing text data 80% of them are unstructured. The process of extracting valuable insights from unstructured data or the process of transforming unstructured data into structured data, to make the data accessible and useful is known as Text mining or text analytics.

The fields of text mining are NLP, Information Retrieval, Data mining and information Extraction. Text Transformation, Feature Selection, Data mining, Text pre-processing, Applications, Evaluate these are the processes of Text mining.

Keywords: Text Mining, NLP, Data mining

I. INTRODUCTION

Text mining is outlined as ―the non-trivial extraction of Hidden, antecedent unknown, and doubtless helpful data from (large quantity of) matter data''. Text Mining can also be a new field that is trying to extract understandable data from text language. It may be outlined because the method of analysing text to extract data that's helpful for a selected purpose. Compared with the sort of knowledge keep in databases, text is unstructured, ambiguous, and troublesome to method. Notwithstanding, in trendy culture, text is that the most communal means for the formal exchange of knowledge. Text mining typically deals with texts whose perform is that the communication of actual data or Opinions, and therefore the stimuli for attempting to extract data from such text mechanically is fascinating – though success is just partial.

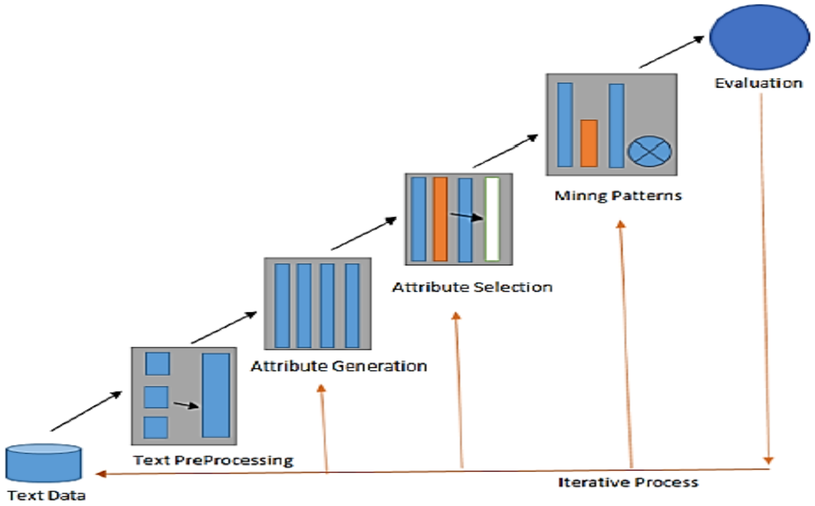
II. BACKGROUND

Text mining is analogous to data processing, except that data processing tools area unit designed to handle structured knowledge from Databases, however text mining also can work with unstructured or Semi-structured knowledge sets like emails, text documents and mark-up language files etc. As an output, text mining can be a much better solution. Text mining is one of the method of structuring the Text input (usually parsing, in joining with the addition of few derived options which are Linguistic and thereby removing others, and Insertion future into a database), patterns which are explained inside the knowledge structure, and the final analysis is the interpretation of Output. The term ―text mining is usually accustomed denote any System that analyses massive quantities of tongue text and detects lexical or linguistic usage patterns in a shot to extract most likely helpful (although solely most likely correct) data.

III. LITERATURE

Process / Activities of Text Mining

- Text pre-processing
  Text Clean-up
  Speech Tagging
  Tokenization
- Attribute Generation
- Attribute Selection
- Mining Patterns



Text pre-processing may be a methodology to wash the text knowledge and build it able to feed knowledge to the model. Text knowledge contains noise in varied forms like emotions, punctuation, text may be a completely different case. once we state Human Language then, there square measure other ways to mention identical factor, and this is often solely the most drawback we've to influence as a result of machines won't perceive words, they have numbers thus we'd like to convert text to numbers in associate economical manner.

Attribute Generation is additionally called text Transformation. The text document is described by the words (features) it contains and their occurrences. 2 main approaches of document illustration square measure a) Bag of words b) Vector area.

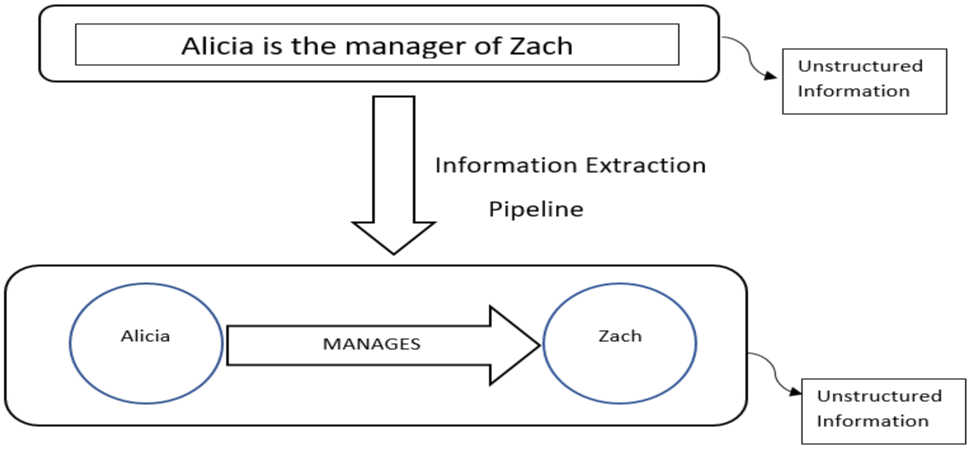
Feature choice additionally called variable choice that is nothing however Attribute choice, is that the method of choosing a set of necessary options to be used in model creation. The most assumption once employing a feature choice technique is that the information contain several redundant or extraneous options. Redundant options square measure the one that provides no further data. extraneous options give no helpful or relevant data in any context. Feature choice technique may be a set of the additional general field of feature extraction.

Techniques used in Text Mining

- Information Extraction
- Information retrieval
- Categorization
- Clustering

- Summarisation

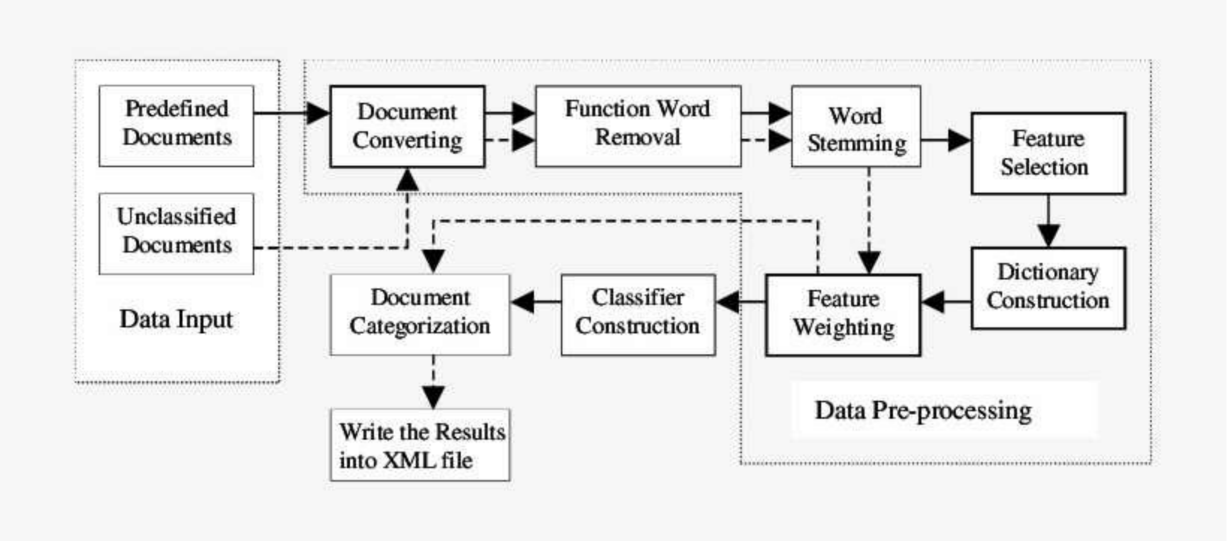
Information extraction is the task of automatically extracting structured information from unstructured or semi-structured, machine-readable data in the form of text using natural language processing. It is the starting step where the system will be able to convert the unstructured data by discovering key phrases and relationships within the text and involves tasks such as identification of named entities, tokenization, part-of-speech, and sentence segmentation. For this purpose, Information Extraction Systems are practiced to bring out some specific information, entities, and attributes from the text and recognize their relationship. After this process, the extracted information is well-organized and stored in the databases for further process. The process which is used to check and evaluate the relevance of results is known as 'Precision and Recall'.



Information Retrieval is the process of extracting relevant information based on a set of specific words or phrases. In text mining techniques, Information retrieval systems make use of different algorithms, which track and monitor user behaviours to discover the relevant data accordingly.
The information retrieval process begins as soon as the user enters a query into the system. User queries are matched against the database information, and the results returned may or may not match the query, hence the results are typically ranked. The information retrieval system provides access to documents, and stores and manages the documents. Web search engines are Information retrieval applications.

Information Retrieval is the process of extracting relevant information based on a set of specific words or phrases. In text mining techniques, Information retrieval systems make use of different algorithms, which track and monitor user behaviours to discover the relevant data accordingly.
The information retrieval process begins as soon as the user enters a query into the system. User queries are matched against the database information, and the results returned may or may not match the query, hence the results are typically ranked. The information retrieval system provides access to documents, and stores and manages the documents. Web search engines are Information retrieval applications.

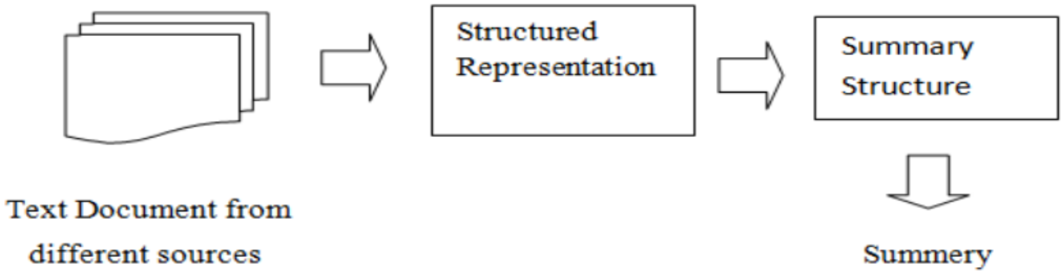
A categorization is a form of supervised learning because it is based on input-output examples to classify new documents. Thus categorization, also known as natural language processing is a process of gathering text documents, processing, and analysing them. The purpose of text categorization is to increase the detection of information which leads to a better decision. It involves methods such as indexing, pre-processing, dimensionality reduction, and classification. The main goal of categorization is to train the classifier based on known and unknown examples. Statistical classification techniques like Nearest Neighbour Classifier, Naïve Bayesian Classifier, Support Vector Machines, and Decision Tree can be used to categorize the text.



It is one of the most crucial mining techniques. This method is an unsupervised method that classifies the text documents into small groups by applying various clustering algorithms. Clustering means similar terms or patterns are organized and extracted from various text documents, which clustering can be performed in top-down and bottom-up ways. So as a result, unique patterns called clusters are generated and each of these clusters has a number of documents. Contents in a single cluster of each document are similar and the contents in different clusters are not similar, which generates a better quality of clustering. The main advantage of clustering, for multiple class's text content can be relevant. There are different types of clustering techniques, k-means clustering, hierarchical, distribution, and density centroid which are used for analysing the unstructured text documents.

Text summarization refers to the process of automatically generating the compressed version of a specific data by breaking down the long publications into manageable sentences, which contain valuable information.
The summarization process includes the following steps,

- Pre-processing obtains a structured representation of the original text.
- An algorithm is applied to transform the text structure into a summary structure in the next processing step.
- The final summary is obtained from the summary structure at the invention step.


Structured Representation
Summary Structure
Text Document from different sources
Summery

IV. CONCLUSION

The process of extracting valuable information from unstructured data is known as Text mining. Text mining algorithms reduce time and cost by providing useful and structured data. Data mining techniques are applied in text extraction to get useful patterns from the documents. And the relevant data is produced from the corpus, which is known as summarization. Clustering is an unsupervised technique and classification is a supervised technique. The successful implementation of the text mining techniques helps us to identify the category of documents and where it fits best.

Reference

[1] Likes Kumar et al, Journal of Global Research in Computer Science, 4 (3), March 2013, 36-39
[2] Text Mining Summit Conference Brochure, http://www.textminingnews.com/, 200
[3] http://en.wikipedia.org/wiki/Text_analytics
[4] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay. ―A tutorial review on Text Mining Algorithms‖, in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, 2012.
[5] http://www.scism.lsbu.ac.uk/inmandw/ir/jaberwocky.htm
[6] Johannes C. Scholtes. ―Text-Mining: The next step in search technology‖, DESI-III Workshop Barcelona, 2009.