International Journal of scientific research and management
(IJSRM) ||Volume||4||Issue||11||Pages||4902-4906 ||2016||
Website: www.ijsrm.in ISSN (e): 2321-3418

Big Data and Natural Language Processing came together for better information extraction: Text Analytics

Harjit Singh
Punjabi University Neighbourhood Campus, Dehla Seehan (Sangrur)

ABSTRACT

Lot of text content is available in a variety of sources such as blogs, emails, social networking sites, forums etc. that can be analyzed to get some useful information. This type of analysis is called text analytics. Text Analysis is related to Natural Language Processing and now Big Data is a voluminous source of unstructured information for Text Analytics. Big Data analysis is moving towards Text Analytics to get more and more from huge collection of data and use that information for better decisions. Text Analytics techniques include summarization, classification, keyword extraction, information visualization, question answering system, deep learning, clustering, link analysis etc. Varieties of fields such as business, finance, opinion analysis, medical analysis, legal analysis etc. are taking benefits from text analytics. This paper discusses various text analytics techniques in detail and concludes with future research scope of text analytics.

Keywords: Text Analytics, Big Data, Natural Language Processing, Summarization, Classification, Clustering, Keyword Extraction.

I. INTRODUCTION

Natural Language Processing meets Big Data to emerge a new research field Text Analytics. A brief introduction to these terms is given below:

A. Natural Language Processing

Languages help the people to communicate with each other. Natural languages are used for oral and written conversation. Computers do not understand these languages. The languages of computers are the different set. Natural language processing (NLP) is a specific research area where the researchers try to process a natural language to generate some kind of statements that the computers can understand. It makes communication with machines as natural as we talk to other humans. The research work is totally multidisciplinary. Linguistic knowledge is must to process it. Computer science knowledge is obviously needed for coding. Statistics knowledge is also a requirement to analyze the data and the results. These three disciplines need to be aggregated to do the fruitful research in the area of natural language processing. The natural language text is read by the NLP system, individual words are processed to understand the meaning or reach a conclusion. The rules define a natural language's way how it is written.

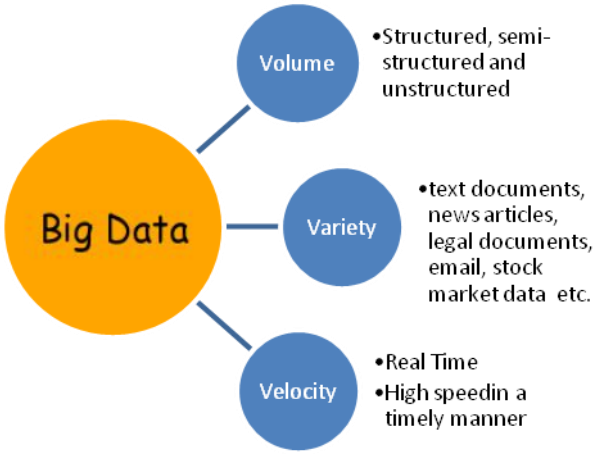
B. Big Data

Big data is a voluminous collection of structured, semi-structured and unstructured collected from a variety of resources such as social media, business transactions, news articles, legal documents and many more. The amount of data is not so important rather the purpose of that data for the organization matters. Big data is analyzed to find unknown information that helps to take better decisions. Doug Laney, an industry analyst conveyed the definition of big data as the 3Vs:

Volume: Big Data is a voluminous collection of structured, semi-structured and unstructured collected from a variety of resources such as social media, business transactions, news articles, legal documents and many more.

Variety: The data exist in various formats. It may be structured like records or can be unstructured like plain text. Even the data can be non-textual such as audio or video. It can be high valued financial transactions or some raw text entered on social media.

DOI: 10.18535/ijsrm/v4i11.15



Velocity: Data flows at very high speed and is dealt in a timely manner.

C. Text Analytics

Text Analytics is a name recently introduced for Natural Language Understanding plus Text and Data Mining. It discovers previously unknown information through automatic extraction of information from several text documents. Data mining tools extract valuable information from data stored in databases (structured data) or the data produced by preprocessing unstructured data. Text analytics in some way is an extension of Data Mining which deals with unstructured and semi-structured data and is not limited to structured data stored in relational databases. Different unstructured resources for text analytics include web pages, news paper articles, blogs, text documents, emails, academic papers, social media text etc. It refers to the extraction of knowledge or non-trivial information from unstructured text. Text analytics which is also known as KDT (Knowledge Discovery in Text) or Text Data Mining is a multidisciplinary research area, based upon Information Extraction, Natural Language Processing, Machine Learning, Data Mining and Statistics. Based upon the interests and methodologies of different expertise working in the field such as computer experts, law experts, financial persons, linguistics, medical experts, psychologists etc., the research is disjointed. Text analytics provide valuable information to decision makers to deal with market trends, fraud detection and risk management.

II. TEXT ANALYTICS TECHNIQUES

The input to text analytics is the written text resources. The text resources are collected and fed as input to text mining tool for processing based on the format and character sets. The output of text mining becomes input for text analysis for information extraction based on the rules. The rule is defined o match a pattern of information from the input resources. Text analysis is repeated until satisfying information is extracted. The components of text analysis include:

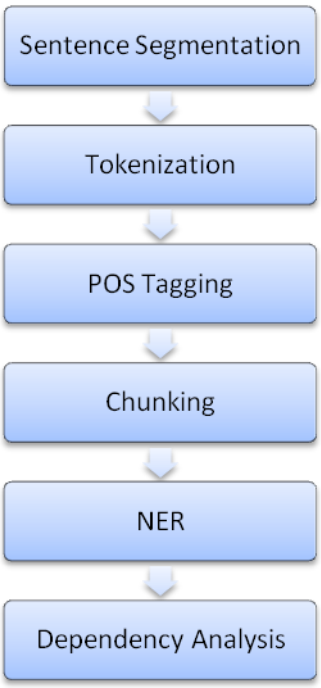
Sentence Segmentation: The sentence segmentation divides text into sentences.

Tokenization: Tokenization divides sentences into words identified using spaces between them.

POS Tagging: Part-of-speech tagging identifies nouns, adjectives, verbs etc. based on lookup and relationship between words.

Chunking: Chunking is used to divide text into subparts using noun phrases, verb phrases and subordinate clauses etc.

NER: Named Entity Recognition component identifies people, places, organizations etc.



Dependency Analysis: Dependency Analysis identifies the references of pronouns such as pronominal anaphora, subordinate clauses etc.

Big Data's role in text analytics is that it provides unstructured written text to text analytics techniques. Now a day, top business strategies need analysis from

almost all the resources of information that are available and the data becomes so big that it is Big Data. Some developments of text analytics techniques include:

A. Summarization

Summarization tools extract important sentences from a document to provide brief summary of the long article. Summarization has long history in the field of text analytics. It is a category of Natural Language Generation. The main purpose of text summarization is to reduce the details in an article to reduce its length but at the same time the overall meaning and main points are preserved. Important sentences are extracted based on statistical ranking of those sentences. Also headings and subheadings are identified to find key points of the article. Deeper analysis method of summarization works on semantic representation of text while shadow analysis works on syntactic representation to extract important parts of text.

B. Classification

Like summarization, classification tools also searches the whole document but only count words. Based on word count the main topics of document are identified and the document is placed in those predefined set of topics. Classification process takes care of synonyms and related terms and ranks the document based on the most content on a predefined topic.

C. Keyword Extraction

Keyword extraction works to identify important words from a document those provide a description of the document. It is useful to get a short summary of the documents such as new articles. Extracting keywords manually is a very cumbersome task. One application of keyword extraction is topic tracking system which identifies keywords from the documents the user views and uses those keywords to present similar documents to the user. Classification system discussed above is very closely related to keyword extraction but keyword extraction has many other applications.

D. Named Entity Recognition

Named Entity Recognition techniques are used to extract entity names from long text documents. These entities may be names of persons, locations, organizations, monetary values, percentages, quantities, stock market values etc. Named Entity Recognition classifies text elements into some categories that are already defined.

E. Question Answering System

Natural Language Processing queries are processed by question answering system to find best possible answer to question. An open source question answering system named OpenEphyra was developed by Nico Schlaefer. It was discontinued later on. YodaQA is another question answering system for general purpose.

F. Clustering

Clustering technique searches similarity among documents by comparing keywords from them. Unlike classification there are no predefined topics (categories), instead unsupervised learning is used by the system to place a document in a category. All the documents are processed and put in some topic or even subtopic as per the search results so no useful document is omitted from clustering. Clustering technique is very useful to divide legal documents in some clusters.

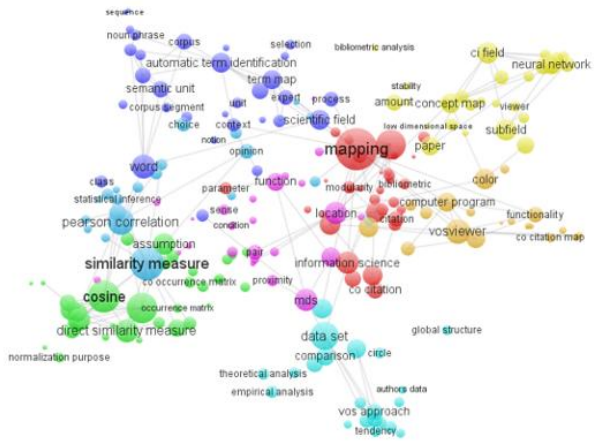
G. Deep Learning Technique

Deep Learning Technique is basically used in Neural Networks and is now being used in Natural Language Processing based on RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network). Using RNNs, the next word in a sentence is predicted based on the words previously written in that sentence. RNNs use memory to store previously calculated information which is then applied for next prediction. CNNs use a number of layers of convolutions to calculate the result. Hundreds or even thousands of filters are used at each layer and the results are combined to form final output. These are applicable for spam detection tools, topic classification techniques, sentiment analysis etc.

H. Information Visualization

Information Visualization presents textual information in a visual form so that the users can easily take an overview of information available. It presents a map of textual information by narrow downing the variety of available documents with related topics. A visual hierarchy of textual resources of information makes it easy to browse those resources as per the need and interests. It provides much more than simple searching techniques.

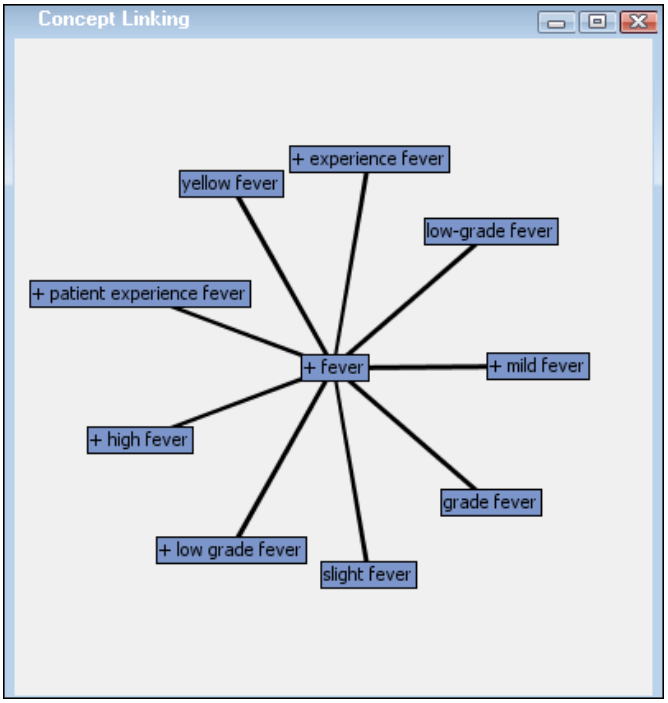
DOI: 10.18535/ijsrm/v4i11.15



Source:http://www.VOSViewer.com

I. Link Analysis

Link Analysis techniques identify common concepts in documents and link them so that users can easily find information in a series of documents connected to each other. When there is a lot of material available and simple searching returns lot of documents which are difficult to browse, link analysis connects related documents together based on the common concepts of these documents. In this way the users can go on reading only related documents and need not to explore all the text material and put associations.



Source: http://www.sas.com

III. TEXT ANALYTICS PROBLEMS

From the viewpoint of text analytics, big data can be defined as a huge collection of natural language data. It may consist of web pages, news paper articles, blogs, text documents, emails, academic papers, social media text etc. Some of these resources provide reliable and valid data that can be directly used for text analysis but some resources such as social media text provide data that is full of noise and distortion. This problem exists in every data set but it becomes very serious problem when we consider big data. Such a huge collection of data cannot be directly viewed by human experts. The only solution is to clean data before using it, but it consumes lot of efforts and even after cleansing the data may not be reliable enough.

Another issue with information extraction is use of irony in a language. When words in a sentence are used in a way that the actual meaning of those words is different than their intended meaning, it is irony. This type of text material becomes very difficult to analyze because the intended meaning is hidden and word meaning is totally different. The solution must exist in the interpretation of text in the context with extra-linguistic knowledge.

The availability of data is another major issue in text analytics. The researchers may have no access to the data from organizations. They may refuse to provide data to researchers. Even if the researchers find big data successfully, the data may not follow some target pattern and it becomes difficult to replicate the analyses. Some data that may not be big enough but may be of good quality such as language corpora, and so it may offer such information that may not be offered by big data.

IV. CONCLUSION

There is variety of fields where text analytics is being used such as business, finance, opinion analysis, medical analysis, legal analysis etc. The information is extracted from a voluminous set of unstructured written text. Big data provides a big platform to text analytics which is although not easily extractable but if extracted successfully is very useful. Text analytics provided enhancements to market research, consumer behavior research, media research, opinion analysis research etc. As more and more text is being generated by general public via social networking, text analytics has a big scope in future research and development. There is a future scope for deep learning in NLP. Emotion analysis, knowledge graphs, speech analysis and article composition such as text messages, emails and

summaries from available text data can get more benefits from text analysis.

REFERENCES

[1] Van Eck, N.J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. ISSI Newsletter, 7(3), 50-54.

[2] Paola Cerchiello, Paolo Giudici, "Big data analysis for financial risk management", Journal of Big Data, 2016

[3] Antonio Moreno1, Teófilo Redondo, "Text Analytics: the convergence of Big Data and Artificial Intelligence", International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 3, No.6

[4] Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain, Natural Language Processing, International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 1, Issue 4

[5] Bharati,Akshar,Chaityanya Vineet and Sangal Rajeev,(1995),Natural Language Processing: A Paninian Perspective,Prentice-Hall of India.

[6] Cini Kurian, A Review of the Progress of Natural Language Processing in India, International Journal of Advances in Engineering & Technology, Volume 7, Issue 5 (Nov. 2014).

[7] http://www.slideshare.net/jhonrehmat/natural language processing.

[8] Natural Language Processing,www.myreaders.info /html/artificial_intelligence.html.

[9] Natural Language Processing-Computer science and engineering, www.cse.unt.edu/~rada/CSCE5290/Lectures/Intro.ppt

[10] NLP, https://www.coursera.org/course/nlp

[11] NLP, research.microsoft.com/en-us/groups/nlp/

[12] Dash, N S and B B Chaudhuri. "Why do we need to develop corpora in Indian languages", International Conference on SCALLA, Banglore, 2001

[13] R M K Sinha. " Machine Translation : An Indian Perspective " , Proceedings of the Language Engineering Conference (LEC'02)

[14] https://en.wikipedia.org/wiki/History_of_natural_language_processing