IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Keyword Extraction using Natural Language Processing

[1]A Manikanta Lakshmi Narayana, [2]Mrs K Santoshi

[1]Student, [2]Assistant Professor

[1]Department of Information Technology,

[1]GMR Institute of Technology, Rajam, India

ABSTRACT:

We provide an analysis of the main text mining methodologies for extracting keywords. In addition to their many other uses, keywords are very helpful for rapidly and effectively evaluating vast amounts of text while conducting an online search. A document's keywords and are a group of illustrative words that provide interested readers with a highly detailed description of the content. They can be used for index construction, query refinement, text summarization, author aid, etc. They are heavily employed in the field of computer science, particularly in information retrieval and natural language processing. Additionally, we covered some crucial feature selection metrics that researchers typically use to evaluate potential keywords in terms of relevance.

Keywords: Extraction, Keywords, Text Summarization, Natural Language Processing, Text mining.

INTRODUCTION

The process of automatically identifying keywords that can be used to model topics and represent the text is known as keyword extraction. The amount of textual information we must rapidly scan through these days to identify documents relevant to our interests is increasing every day. Several million web pages and tens of thousands of text files are frequently stored nowadays. Having a subset of words (Keywords) that can provide us the major characteristics, concept, subject, etc. of the document will make it easier for us to analyze such massive amounts of data. Appropriate keywords can act as a very good summary of a document and make it easier for us to organize and find materials based on their content. In academic writing, keywords are utilized to inform the reader about the article's substance. In a textbook, they are helpful for readers to recognize and remember the key ideas in a given section. Keywords can be utilized as a gauge of similarity for text clustering because they capture the essential idea of a document.

Importance of Keyword Extraction: Keyword extraction you can find the most important words and phrases in massive datasets in just seconds. And these words and phrases can provide valuable insights into topics your customers are talking about.

LITERATURE SURVEY

In paper [1] Thiruni D. Jayasiriwardene, Gamage Upeksha Ganegoda mainly focuses on Keyword extraction from Tweets. The methods used in this paper are Dataset, Pre-processing, Extracting keywords, Named Entity Recognition (NER), Part of speech tagging, Stanford Core NLP, Wordnet corpus. It is easy to extract keywords from structured documents by considering their linguistic features. But for unstructured text such as tweets, it has become difficult because linguistic features cannot be found due to character limitations. The method which is used to find synonyms and similarity of two words can be modified to give a more precise output which will lead to an increment of performance.

In paper [2] M. Abulaish, T. Anwar mainly focuses on identifying feasible keyphrases in text documents. These methods mines various lexical and semantic features from texts to learn a classification model. The techniques used in this paper are Keyword extraction, Machine learning and Text mining techniques like phrase identification, Feature vector generation. Instead of applying full or partial parsing of text documents for PoS(Parts of Speech) tag patterning, which is generally not feasible for complex sentences, our method applies n-gram technique for candidate phrase generation.

In paper [3] Himani Shukla, Misha Kakkar mainly focused on Regular Expression Grammar Rule approach to identify the Noun Chunks in the text of the transcript. These techniques are used in this paper are machine learning and data mining techniques like MaxEnt and SVM Classifier, NP Tag patterns. It is a technique to extract keywords from educational video transcripts from MOOC's is discussed. The objective is to extract keywords from the text data such that the keywords summarize the text and the topic discussed in the text. This method is found to be simple and robust in extracting noun phrase from the text as it gives the user control in determining the kind of tag patterns to be matched. NLP tasks can be carried out in a distributed framework for extracting keywords from large datasets like web pages.

In paper [4] S. Uthayashangar, T. Ganesh Aravind, K. Saranidaran, V. Sivapavithran, R. Vijayaram Abishek mainly focused on extracting keywords from the collection of dataset from the Facebook account data. By extracting the keywords from the specific account, we will provide the advertisement with help

of the business organizations, to improve the business growth of each organization. The algorithms used in this paper are Decision Tree Algorithm, Keyword Extraction, Image OCR, Random Forest Classifier. We can able to produce growth to the product sales development and it will help them to know the needs of the people about any product or course application, so it is useful to both people and business organization.

METHODOLOGY

1. Data Preprocessing:

The steps involved in Data Preprocessing:

1. Extract text from the data set.
2. Remove URL, Punctuation, Special Characters.
3. Convert to lower case, remove stop words.
4. Lemmatization and Posterstemmer.
5. Tokenization.
6. Spelling correction.
7. Wordnet corpus.

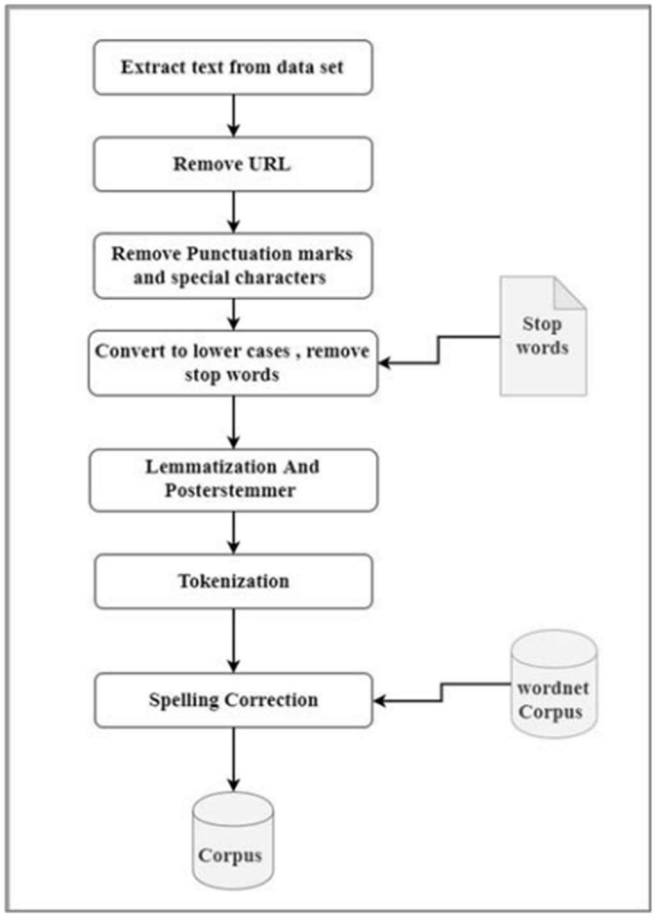

Fig: Data preprocessing

A methodology can be proposed to use TF-IDF statistical method along with NER to find the candidate keywords. TF-IDF is based on the frequency of occurrence of a word in the corpus. This process focuses on the most important factors related to the claimed tweet such as person names, locations, organizations etc.

2. Keyphrase Extraction:

The steps involved:

1. Document preprocessing.
2. Candidate phrase identification.
3. Feature vector generation.

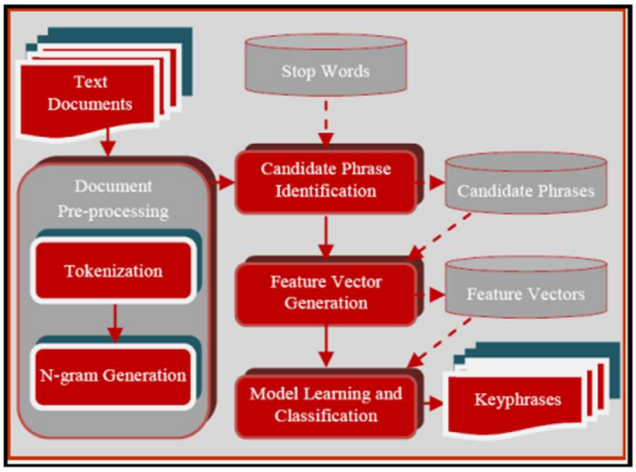

Fig: Functional details of the proposed approach

Tokenization is a method for fragmentize texts into smaller chunks. The phrase processing task consists of removing apostrophes, cleaning numerals associated with a phrase at the boundaries, stemming, case folding, etc. The set of candidate phrases are transformed to equivalent feature vectors comprising eight feature values.

3. Data Mining Technique:

These are the steps involved:

1. Raw Text
2. Sentence Tokenisation
3. Word Tokenisation
4. PoS Tagging
5. NP Chunking

PoS tagging plays a very important role in NP Chunking. The most commonly used NP Chunk grammar is represented as: NP: <DT>?<JJ> *<NN> which defines a NP chunk pattern. After the chunking process, the NP tagged leaves of the chunk tree are extracted and stored as final keywords in the last step.

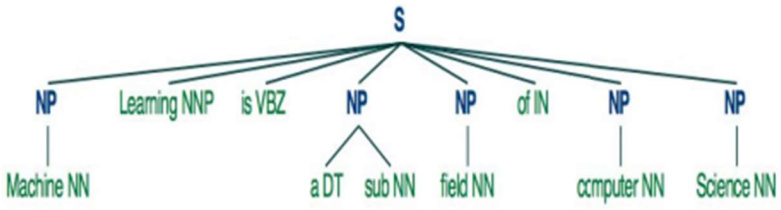

Fig: Chunk Tree

4. Decision Tree Algorithm:

The supervised learning algorithm family that includes the decision tree method. Each leaf node of a tree corresponds to a class label, and each branch node to an attribute. There are two methods are used in it. They are

1. CART (Classification and Regression Tree)
2. Information Gain

It is used to predict the decision from the preferred conditions. To predict how many output labels will be present for analysis, regression is used.

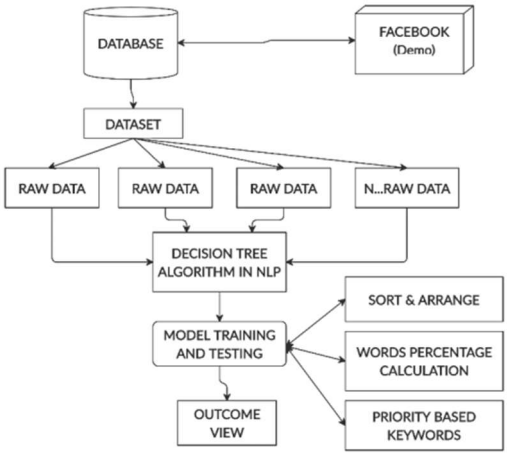

Fig: Architecture Diagram

RESULTS & DISCUSSIONS

The Regular Expression Grammar Rule Based Chunker identified the Noun Phrase Chunks in the text. Chunking is the process of annotating tagged tokens with structures. Since cdhunkers do not analyze complete sentences but form "chunks" of words, the grammar rule based approach is simple, robust and efficient. In the last section, out of the 4 experimented NP tag patterns (Grammar), the following tag.Patten yielded the best keywords.

The sample text for keyword extraction shown in Fig-1 depicts the kind of text data used in this study for chunk parsing. The quality of NLTK Chunker depends on the quality of tag patterns (Grammers) developed. The generated keyword from sample text of Fig-1 are depicted in Fig-2.

```
<text>
Finally we can only observe. Physical parameter will vary with
input and output but if you become an observer the observation
can only be discrete. But whether it is a discrete time signal
or analog time signal the variation of the signals
between these two limits is a continuous variation there is no
defined points by which it can vary. What I am saying is
supposing my initial value is 0V and Vmax is 5V as an
example, as an example I make this 5V I am only allowing this
signal to change from 0 to 5V and within this 0 to 5V it can
take any value.
</text>
```

Fig-1: Sample Text for Keyword Extraction

```
<keywords>
physical parameter,input,output,observer,observation,discrete
time,analog time signal,variation,signals,limits,continuous
variation,points,initialvalue,vmax,example,signal,value,diffic
ulty,limitations
</keywords>
```

Fig-2: Sample Keywords

CONCLUSION

From the analysis conducted in this paper, different algorithms are used. Comparative study performed among the various techniques like Decision tree algorithm and Data Mining Technique to find key words in the given text. The usage of Decision tree algorithm after Data Mining would give result accurately for the user.

REFERENCES

[1] Thiruni D . Jayasiriwardene ; Gamage Upeksha Ganegoda ; " Keyword extraction from Tweets using NLP tools for collecting relevant news " 2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)

[2] M. Abulaish and T. Anwar ; " A supervised learning approach for automatic keyphrase extraction ", International Journal of Innovative Computing,Information and Control, vol. 8, 2012

[3] Himani Shukla and Misha Kakkar ; " Keyword extraction from Educational Video transcripts using NLP techniques " , 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)

[4] S. Uthayashangar ; T.Ganesh Aravind ; K. Saranidaran ; V. Sivapavithran ; R. Vijayaram Abishek " Taxonomy of keyword extraction in Facebook using Decision Tree Algorithm in NLP ", 2020 International Conference on System, Computation, Automation and Networking (ICSCAN)