

## APACHE OPEN NLP (NATURAL LANGUAGE PROCESSING) LIBRARY INTERFACE FOR HUMAN INTELLIGENCE TO MACHINE LEARNING PERSPECTIVES

Mr.S.S.Aravinth, AP/CSE, Knowledge Institute of Technology, Salem

Mrs.C.Vanitha, AP/CSE, Knowledge Institute of Technology, Salem

R.J .Vigneswaran/III – CSE /Knowledge Institute of Technology, Salem

M.Vijayakumar/III – CSE /Knowledge Institute of Technology, Salem

R.Sabarish / III – CSE/Knowledge Institute of Technology, Salem

Mail ID: [ssacse@kiot.ac.in](mailto:ssacse@kiot.ac.in) Mobile: 098944 – 48683

### ABSTRACT:

The open NLP is a toolkit used to learn about machine language. It will act as an interface between human beings to things such as machines. The human generated codes, symbols & texts will be converted into machine understandable words. This natural language processing tool will play a major role in machine intelligence and learning process. The natural language processing is well advanced area in theoretical computer science and mathematics. This NLP gives the support of communication

between man to machine learning and interaction.

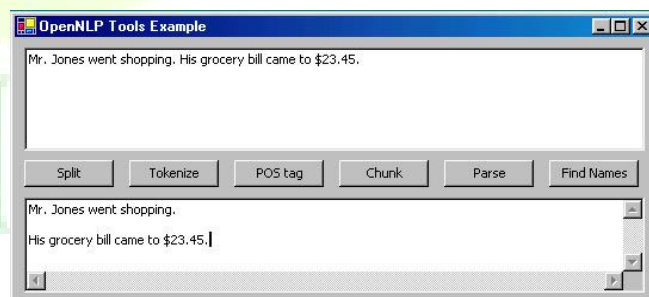
Apache Open NLP provides a strong linguistic support to natural language processing researchers and practitioners. The full package contains tokenization, part of speech, sentence segmentation, sentence detection, estimation of parameters, chunking etc.

### KEYWORDS:

Apache Open Natural Language Processing, tokenizer, NER labels, Chunking, PoS tag, MaxEnt, Parser component, Lexically-based rules

### I. INTRODUCTION:

Jason Baldridge and Gann Bierner were invented the openNLP toolkit in the year of 2000. They were graduate students in the division of informatics at the University of Edinburgh. The apache openNLP is developed for natural language processing. It is developed in various software packages such as java and high level languages.



## **II.OVERVIEW OF APACHE NLP:**

Apache openNLP is a library or software, which used for language processing. It is like a tool kit used for learning about machine language. Apache software foundation has found this apache openNLP. This apache openNLP is written in java language. NLP means Natural Language Processing. Apache 2.0 is the license of apache openNLP. Using the concept of openNLP we can build the more efficient of text processing task. If we want to work on apache toolkit the worker or user must be familiar with java program. The apache openNLP supports all the common NLP tasks.

## **III.MOST COMMON NLP TASKS:**

### **A. TOKENIZATION:**

A tokenizer is a segment and it is used to give an input in the form of tokens. The tokens are segments that has numbers, words and the special characters etc

There are three ways to implement a tokenizer:

- Whitespace Tokenizer
- Simple Tokenizer
- Learnable Tokenizer

There are two stage processes:

- Boundary of a sentence is first identified.
- Next, each sentence of a token are identified.
- The most common method is used for try out the tokenizers are command line tools.

## **B.SENTENCE SEGMENTATION:**

Sentence is the language such as c and c++ and the java languages. There are two sentences segmentation present in the apache openNLP. If the sentence such as code is executed successfully then it is called successful sentence segmentation. Otherwise, it is called unsuccessful sentence segmentation. Written languages cannot be used in this segmentation since it is very difficult to build. In this sentence segmentation, whitespace is used for breaking up the sentence, and it also used to separate the phrases and clauses.

## **C.PART-OF-SPEECH TAGGING:**

The part of speech tagger is a part of software. This tagger read the data in text manner such as noun, verb, pronoun, adjective etc. In this tagger is used for increase the speed, performance, and supports to other languages. Java 1.8+ need for system to run the part of speech software. The both 32 and 64 bit support in this tagger model. More than 60 mb to 200 mb required for to run this tagger software. This tagger software is written in java language.

## **D.NAMED ENTITY RECOGNIZER:**

The named entity recognizer is implemented by java language. It is also known as NER. NER labels are sequence of sentence or words present in the text. The NER labels represent a person names, company names, or gene and protein

names. NER consist of more option to defining the feature extractors. The NER can be recognized only for English that have three classes such as person, organisation, and location. The named entity recognizer available for different languages and circumstances. It should be measured the performance of a word processing.

### **E.CHUNKING:**

Chunking is a group of verbs or nouns in sentences. Chunker is used to split the text into group of words. PoS tag text is the form of input given in the chunker. The different and new languages are trained to deal with a chunker.

There are three columns consist of a training data:

- Word
- PoS tag
- Chunk tag

### **F.PARSING:**

The process of analyzing each word through its constituent parts and determining its structure which is either in natural language or computer language is called parsing. It abides the rules of formal grammar. The process of parsing mainly holds two components such as parser and grammar. Parser component is a procedural component which is a computer program where as Grammar component is a declarative component. Parser does not change

with respect to the languages pursued whereas grammar changes with respect to the languages pursued. Thus a system enables the parsing of various languages by changing the grammar. Here, grammar formalism is the most notified thing since both the parser and grammar depends related to each other.

### **G.COREFERENCE RESOLUTION:**

Coreference is the concept of identifying multiple expressions that refers to the same meaning in a context. This task involves the understanding of natural languages, like document summarization, information extraction and question answering. It is an important process in high level NLP tasks. Coreference resolution system is available in online based on java implementation. Here is an example for coreference resolution. "Peter drove to Amy's apartment. He prepared her lunch". In this sentence both Peter and He refer to the same person and similarly Amy and Her are the same, different entities. One cannot expect openNLP to get this 100% correct. It is a simple example, and also it is a difficult problem.

### **IV.APACHE OPENNLP LIBRARY STRUCTURE:**

#### **A.COMMAND LINE INTERFACE (CLI):**

The Command-Line Interface (CLI), is also known as command-line user interface. This CLI is used to interact with computer programs, also these commands are been accepted in the form of text. It is also named as console user interface and character user interface (CUI).These are been used

in the earlier days on UNIX system, MS-DOS, CP/M and apple DOS. The CLI is not used by ordinary PC's users only by the advance computer users as it is easy to automate via scripting.

#### **B. SENTENCE DETECTION:**

A sentence detector is used to detect the punctuation character that makes the end of a sentence (or) not. A long single sentence is been divided into two sentences using the punctuation character, in which the first and the last character has the longest whitespace. The first non-whitespace character is consider to begin the sentence and the last non-white space character is consider to be the end of the sentence. If the long sentence is given in jumbled sentence (or) word they are been arranged in the proper sentence using this sentence desctor. After detecting the sentence boundaries each sentence is written in its own line. The easy way to work with sentence detector is by using the command line tool. The text are been given in the console, one file read the input and the output file is been redirected to another file.

#### **C. BOUNDARY DETECTION:**

Sentence boundary detection is a way of classifying a task. Characters of ".", "?", "!" is valid character for sentence boundary. The list of using these character are not Sufficient (or) limited, a full-stop "." has many different uses like (decimal point, abbreviations, email etc.). Punctuation character can be used in the place of exitement or in the place of stress, so as per the situvation we are using these characters. So, by Lexically-based rules, debugging rules we can detect the place where these classification methods can be used. By using the detection tool this can been done, we can

identify a sentence and can modify them automatically.

#### **D. MAXIMUM ENTROPY CLASSIFIER:**

Maximum entropy classifier is also known as MaxEnt. It is a discriminative classifier that is used in NLP. It can be implemented using commonly used languages such as java, c++. The MaxEnt is very well used for several text classification problems such as sentiment analysis. It enables minimization of commitments, and models everything that is known.

#### **E. MAXIMUM ENTROPY PRINCIPLE:**

It guarantees the consistency and uniqueness of assignments of probability acquired by various methods. It makes use of different forms of prior data. Beyond the prior data being stated it admits the most ignorance. This principle of maximum entropy when applied to testable information become explicitly useful.

#### **F. PARAMETER ESTIMATION:**

The Parameter estimation is the process of estimating the amount of selected distribution whether the probability will be success or failure using the given sample data (the data that is been taken some years before). By taking the analysis of the previously used data these parameter estimation is been taken. Several methods are available Probability Plotting (it is the easy and simple way) Maximum Likelihood Estimation and Bayesian Estimation methods. For taking the survey report this Parameter Estimation is been used.

#### **CONCLUSION:**

This paper will give you an clear idea about apache openNLP tool which act as the good interface between the machine and the human. It has different methods to detect the sentence and gives the that is given by the user in the text field it automatically detects the sentence in that field whether the data is in number manner or in word and also check that the sentence has noun, verb, pronoun, adjective etc. And by named entity recognizer it recognize whether it is a name, place or location and it is done by word processing. And the white spaces in this consider as the breaking point, and the sentence is automatically detected user desire output. This tool has many option for example, if the user enters the large text or any instance and corrected by using this apache openNLP. And by this we can even also estimate the probability of success or failure with the given data. Really an interested user can get an clear idea on this paper and what are the methods that is been followed in this.

#### **REFERENCE:**

- 1)<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>
- 2)<http://www.programcreek.com/2012/05/opennlp-tutorial/>
- 3)<http://www.ibm.com/developerworks/data/downloads/uima/requirements.html>