International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Natural Language Processing For analysing and Extracting Insights

Manuraj R , Manoj B R

Department of Computer Application

Dayananda Sagar College of Engineering

Abstract

In this paper we will discuss about Natural Language Processing for Extracting information or insights. Mainly Natural language processing is a subfield of artificial intelligence it allows machines to break down and interpret the human readable language. For this process it uses different methods like Tokenization,

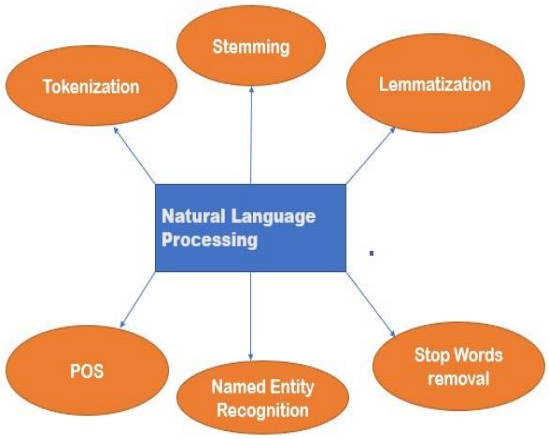
Some interesting applications of natural language processing are Chatbot, sentiment analysis, Speech recognition, Text extraction etc.

Introduction

Natural language processing is concerned with the interactions between human languages (for example: English) and computer .it is important and rapidly developing area of computer application. Natural language processing used as wide range application in different areas like business, daily life applications.

It is the driving force behind things like Speech recognition, machine translation, sentiment analysis, voice assistant like Amazon's Alexa and Siri etc. NLP is a important factor because computer can't understand human languages so we have to use different algorithms or techniques to communicate with computer. NLP also uses different methods and algorithms to fetch data from different data sets or communicate with machine.

Stemming, Stop words, Lemmatization, part of speech etc.



NLP Techniques:

1. **Tokenization:** Tokenization is an NLP method that break down entire text into several words. In these segments each word called tokens. This method eliminates punctuations like question mark, exclamatory, full stop etc.

   For example, if we consider a text

"Tokenization is a method of converting a text into list of tokens" then the list of tokens will be ["Tokenization", "is", "a", "method", "of", "converting", "a", "text", "into", "list", "of", "tokens"].

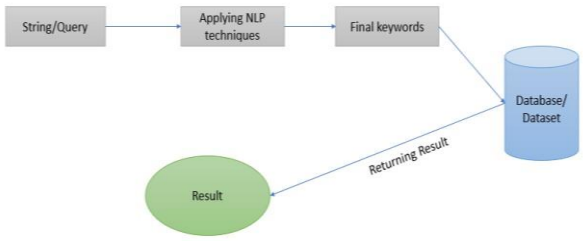
2. **Stemming:** Stemming is a NLP method for reducing the words into their root form. This method based on the principle of having same token for similar words which having different spelling or same meaning. For example, the words "drive", "driving" and "drives" has same token "drive".

3. **Lemmatization:** Lemmatization is much more like to stemming but also reduce the words into their root form which has exact meaning. For example, the word "Hating" will be reduced into "Hat" in stemming which is wrong but in Lemmatization it reduced into "Hate".

4. **Stop words removal:** Stop word removal method is like tokenization but in which it only considers the tokens that as a important meaning for analysing. For example, words like and, a , of etc will be removed form the token list so that it can analyse efficiently.

5. **Named Entity recognition:** It is the method of identify the entity of the text or sentence. It categorize like person , expression, organization etc.

For example " I have joined HARVARD university dated June 2021."
Here Harvard university is organization i.e org NER. And June 2021 is date NER.



It is the information retrieval process from the database or dataset. User query processed through different NLP algorithms. and provides resulted keyword. Those keyword compare with data set and then it fetch the appropriate result .

Chatbot to extract information from trained student dataset. Here I used python for implement NLP methods to extract meaningful information.

Python has built-in libraries for Natural language processing. NLTK package (Natural language toolkit) is a suite of libraries for natural language processing. It is a leading package for building python programs to work with human language. Here we used different techniques like stemming and lemmatization algorithms, Tokenization algorithms, stop words algorithm etc. to fetch exact information from dataset.

main disadvantage of NLP, it is not 100 percent reliable. There are always some errors in information

```
import nltk
from nltk.stem import WordNetLemmatizer
nltk.download('popular', quiet=True)
with open('manu.csv','r', encoding='utf8',
          errors ='ignore') as fin:
    raw = fin.read().lower()
sent_tokens = nltk.sent_tokenize(raw)
word_tokens = nltk.word_tokenize(raw)
print(word_tokens)
```

```
lemmer = WordNetLemmatizer()
print(lemmer)
```

```
<WordNetLemmatizer>
```

```
def LemTokens(tokens):
    return [lemmer.lemmatize(token)
            for token in tokens]
remove_punct_dict = dict((ord(punct), None)
            for punct in string.punctuation)
def LemNormalize(text):
    return LemTokens(nltk.word_tokenize(text.lower()
                    .translate(remove_punct_dict)))
```

```
def greeting(sentence):
    """If user's input is a greeting,
       return a greeting response"""
    for word in sentence.split():
        if word.lower() in GREETING_INPUTS:
            return random.choice(GREETING_RESPONSES)
```

```
flag=True
print("HI GUYS ITS ALL ABOUT DSCE!")
print("DO YOU NEED HELP? (If you dont.,, Enter bye dsce)")

while(flag==True):
    user_response = input()
    user_response=user_response.lower()
    if(user_response!='bye dsce'):
        if(user_response=='thanks' or
           user_response=='thank you' ):
            flag=False
            print("DSCE: You are welcome..")
        else:
            if(greeting(user_response)!=None):
                print("DSCE: "+greeting(user_response))
            else:
                print("DSCE: ",end="")
                print(response(user_response))
                sent_tokens.remove(user_response)
    else:
        flag=False
        print("DSCE: Bye! take care..")
```

```
HI GUYS ITS ALL ABOUT DSCE!
DO YOU NEED HELP? (If you dont.,, Enter bye dsc
e)
hi
DSCE: hello
hi
DSCE: hello pep!
hello
DSCE: hi there
```

Conclusion:

Here we discussed about information retrieval from unstructured or collection of data using several NLP techniques. It improves accuracy of data extraction. The different implementations and methods of NLP can help in different businessplatform.it improves efficiency, and user satisfaction.Some NLP based applications includeautomatic grammar checking chatbot for factextraction, sentimental analysis, insights discovering etc.

retrieval and prediction. So sometimes it shows irrelative information.

References:

[1] Information from:

https://en.wikipedia.org/wiki

[2] Information from:

http://www.google.com

[3] Edward Loper, Steven Bird, Ewan Klein **"**Natural Language Processing with Python" , June 2009.

[4] James H. Martin, Daniel Jurafsky **"**

Speech and Language Processing" , 2000.

[5] "Building Chatbots with Python: Using Natural Language Processing and Machine Learning" Book by Sumit Raj.