Bansilal Ramnath Agarwal Charitable Trust's
Vishwakarma Institute of Technology
(An Autonomous Institute affiliated to Savitribai Phule Pune University)

Department of Information Technology

IT-Bulletin

Nov-2022

'NLP is transforming the way humans and machines interact'

HIGHLIGHTS

'Natural Language Processing(NLP) Basics, Tools and Applications'

I have always been convinced that the only way to get artificial intelligence to work is to do the computation in a way similar to the human brain. That is the goal I have been pursuing and hope to pursue the same with you all ahead. We are making progress, though we still have lots to learn about how the brain actually works hence, let's learn the patterns for the Neurons and build them ourselves.

Not to forget we are cooks not chemist hence, will code it out ourselves.

As there are wide variety of domains in machine learning to deal with like(Regression, classification, Natural Language Processing, Time Series Forecasting, Object Detection, etc).

The machine learning field has developed a lot in past few years wherein many of them are fascinated to know how actually it works.If we specifically talk of Natural language Processing, it has large variety of pre-processing techniques like (tokenization, vectorization, embedding, etc).

'NLP is transforming the way humans and machines interact'

"Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence "

'Natural Language Processing(NLP) Basics, Tools and Applications'

Tokenization:

It is one of the most common pre-processing technique that happen in NLP. Tokenization has a large benefit when we need to deal with large sentences wherein tokenization break the sentence into tokens and assign them a proper numerical value to deal with. Tokenization has three types to deal with namely word, character and sub-word.

Word Tokens for (I Love VIT):

I

Love

VIT

Character Token for (Vishwakarma):

V-i-s-h-w-a-k-a-r-m-a

Sub word token for (Technology):

Techno-logy

Vectorization:

This is one of the pre-processing technique that makes the model building process much more faster and accurate wherein initially it converts the input text into vector of real numbers which is the format that ML models support.

Example:

```
sents=["Hello VIT Pune","I am from IT branch", "Good Morning All"]
```

```
tfidf = TfidfVectorizer()
```

```
transformed = tfidf.fit_transform(sents)
```

```
df = pd.DataFrame(transformed[0].T.todense(),index=tfidf.get_feature_names_out(), columns=["TF-IDF"])
df = df.sort_values('TF-IDF', ascending=False)
```

"Robots will need the ability to perfectly comprehend human speech, making natural language processing more important than ever"

Output:

```
df
```

	TF-IDF
hello	0.57735
pune	0.57735
vit	0.57735
all	0.00000
am	0.00000
branch	0.00000
from	0.00000
good	0.00000
it	0.00000
morning	0.00000

Embedding:

Embedding is the process of converting high-dimensional data to low-dimensional data in the form of a vector in such a way that the two are semantically similar. In its literal sense, "embedding" refers to an extract (portion) of anything.
Basically embedding is nothing but vector of vector.

Text Classification

One of the amazing type of NLP which helps to classify fake tweets based on some event, helps to solve violation that happen now a days on the internet, it also may help to prevent some wrong practise that is going happen somewhere in the world like (Terror attack, etc).

The steps using which we could deal with it would be to initially collect the data from the internet and go-ahead with pre-processing techniques like tokenization, vectorization and embedding then put it into the model and try getting a good accuracy of about 95%.

(Recurrent Neural Network.). The pre-processed text would be initially. passed to the Naïve model and the accuracy would be taken into account and the further models

The models with which we could go ahead would be Naïve bayes, Dense, BERT, RNN (Recurrent Neural Network.). The pre-processed text would be initially. passed to the Naïve model and the accuracy would be taken into account and the further models would be trained to just beat the accuracy that we have got using Naïve model.

Later an Inference model would be built to check the output of the built model on the testing sentences and custom given sentences.

Text Translation:

Now, describing about the second application of NLP, that is Text Translation. Let us consider the translation of the Marathi sentence into English. This is the sequence to sequence prediction problem we have to predict the translated English sentence, given an input of Marathi sentence.

For the sequential data, the Recurrent Neural Network (RNN) is mostly preferred. So we will be using Bidirectional LSTM which is one of the way of implementing RNN. The flow of the entire project is mentioned below:-

1. Checking if the data has any Nan values and removing them as well.
2. Converting the text into Lower case.
3. Adding Starting and Ending token to the target data:
4. Creating the vocabulary of all unique Marathi and English:
5. Creating the vocabulary of all unique Marathi and English words and also calculating the maximum vocabulary size of both Marathi and English words which will be passed as a parameter to the Embedding layer in the Encoder-Decoder model.
6. Tokenization
7. Post-padding the sentences:
8. Data Splitting: Splitting the data into two parts the training data and the testing data, in the case of translation dataset is split in the ratio of 80:20, given training data the size of 80% of the entire data and testing data the size of 20% of the entire data.

'Natural Language Processing(NLP) Basics, Tools and Applications'

9. Building the Model (Encoder and model with attention layer): Encoder Decoder model for the sequence to sequence problem: Sequence to Sequence is a machine translation and language processing approach based on encoder-decoders that converts a sequence input to a sequence output. LSTM layers are used in both the encoder and the decoder. The encoder decoder model is built for text translation as the encoder model releases two states the hidden state and the cell state these are also known as the context vectors. The decoder is an LSTM layer where it outputs are set to the Encoder model's final states. The decoder model generates the output sequence which will be converted into English text by using the vocabulary which we previously created for English.

10. Making Inference model to get the predictions.

Student Editor:

Tejas Gadi
TY-IT

Adwait Bhosale
TY-IT

HOME TOP