NLP task

we need to focus how to convert a Text to numbers

this is called as WordEmbeddings

1.Bow

2.tf-idf

3.Word2Vec

4.BERT Transformers

BERT (Bidirectional Encoder Representations from Transformers)

Package name: Transformers

Developed by: Huggingface ==== Google

=========================================

Understanding the BERT models

1.install transformers

2.we need to import

A) BertModel is used to access the Models(1M)

B) BertTokenizer is used to access the vocab and tokens

3. BertModel here we access

model name: 'bert-base-uncased'

4. Input layer === Encoder layer === Pooler layer

5. the total tokens or words= 30522

6. the maximum sentence token size= 512

7. each token vector size = 768

8. Tokenizer has a vocab list

A)which has token and its index

B) 30522 tokens are available

C) [PAD]

D) [CLS]

E) [SEP]

D) [UNK]

E) [MSK]

9) Best part here is

when we want to divide a sentence into tokens

A) first it will search is that word available in vocab list

B) If it is not available

it is try to tokenize the word with the tokens

which is available in vocabulary list

ex: 'moye' ==== > 'mo'+'#ye'

10) Any senetence can tokenize

11) Sentence= 'we are learning BERT Embeddings'

12) ['we','are','learning','BERT','EM','#bed','#dings']

13) Here we are adding some spl tokens [cls],[sep],[PAD]

14) ['[cls]','we','are','learning','BERT',

'EM','#bed','#dings','[SEP]','[PAD]']

15) Attention mask:

where ever [PAD] is make it as 0, other wise 1

16) Attention mask:[1,1,1,1,1,1,1,1,1,0]

17) Token ids: [101, 2057, 2024, 4083, 14324, 7861, 8270, 4667, 2015, 102, 0, 0]

18) In order to get the Vector we need to pass

A) Attention mask of sentence

B) TokenIds of sentence

19) The sentence has 7 words+3 spl tokens , we will get seven vectors

each vector size is 768

shape (10,768)

20) we will apply the pooler layer on 10 vectors finally we will get

one sentence vectors

21) token ids should be in the form of

either pytorch tensors

or tensorflow tensors

A) LIST ======== > [1,2,3,4]

B) numpy array === > array([1,2,3,4])

C) Pytorch ===== > tensor([1,2,3,4])

D) Tensorflow==== > tensor([1,2,3,4],dtype=int64)