Assignment 1 – Part 2

CSCI 5832

Rithik Kumar Athiganur Senthil

How many words are really in the vocabulary of BERT?

After manually going through the text file of Bert’s vocabulary, I could see a lot of meaningless words, numbers, symbols and words that include digits and symbols. So, before I start analyzing how many meaningful words are present, I must clear the file from all these meaningless words.

Once I started reading the text file using an object, I had to first tokenize the file to extract the words and start analysis on them. So, initially I converted all the words to a lower-case and tokenized them. After that I had to strip down meaningless characters or words from the tokenized words, for example “unused987”, #, $, %, &, ##d. There were also characters from other languages apart from English which should also be considered (After performing this the number of words from around 30000 dropped down to 27015).

Also, I performed stemming to all the words to reduce the number of words because the words like cat and cats cannot be considered as two different words and these two must be reduced to cat and stemming does the same for us (and after performing this and getting only the unique words from it the number of words dropped down from 27015 to 13944).

Once stemming is completed, I performed lemmatization technique to the file, by performing this we can convert words like goose and geese to a common word goose because we cannot consider these two words as different, they have a common meaning and after performing lemmatization technique to this (and extracting only the unique words from it, the number of words dropped down from 13944 to **13737**).

After cleaning the file using above methods the resulting file was nearly a clean one and was ready to be considered as a file containing meaningful words and further, I had to only extract the unique words out of it because it will surely have repetitions.

Python code for finding the number of words:

```
import nltk

nltk.download('punkt')

filename ="BERT-vocab.txt"

file=open(filename,encoding="utf8")

text = file.read()                                      # read the file

text = text.replace("\n"," ")                           # remove lines and replace with spaces

text = text.lower()                                     # convert to lower case

words = text.split()                                    # tokenize

print(len(words))                                       # Output: 30522

words = [word.strip('.,!;()[]') for word in words]

words = [word.strip('##') for word in words]

words = [word.strip('[[:digit:]]') for word in words]

words = [word.strip('^.$') for word in words]           # destructor the words by stripping off all the

words = [word.strip('^\[') for word in words]           # meaning less words

words = [word.strip('unused#') for word in words]

words = [word.replace("'s", '') for word in words]

words = [word.strip('^!"#$%&()*+,-./:;<=>?@[\]^_{|}~') for word in words]

words = [idx for idx in words if not re.findall("[^\u0000-\u05C0\u2100-\u214F]+", idx)]

words = [x for x in words if not (x.isdigit() or x[1:].isdigit())]

print(len(words))                                       # Output: 27015


from nltk.stem import PorterStemmer                     # import porterstemmer from nltk library

ps=PorterStemmer()

stemmed_tokens=[]

for i in words:

   stemmed_tokens.append(ps.stem(i))                   # perform stemming and push them to a list

stemmed_set=set(stemmed_tokens)                        # put it in a set to get unique words

print(len(stemmed_set))                                # Output: 13944
```

```
from nltk.stem import WordNetLemmatizer             # import WordNetLemmatizer from nltk library
lemm=WordNetLemmatizer()
lemmed_words=[]
for i in stemmed_set:
  if len(i)!=1:                                     # ignoring single characters
    lemmed_words.append(lemm.lemmatize(i))          # perform lemmatization and push it to a list
lemmed_set=set(lemmed_words)                        # put it in a set to get unique words
print(len(lemmed_set))                              # Output: 13737
print(lemmed_set)
```

References:

https://www.guru99.com/stemming-lemmatization-python-nltk.html