Spam Classification

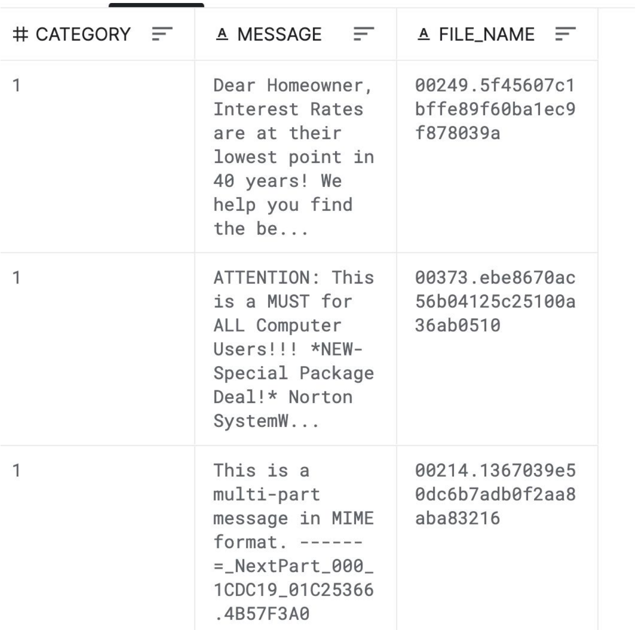
with BERT Feature Extraction

- MSDS631 Haotian Gong Yangzhou Tang

Dataset

Spam Email is harmful for users experiences. By detecting unsolicited and unwanted emails, we can prevent spam messages from creeping into the user’s inbox.

Our data set has 5k+ rows with three two columns, binary indicator specifying whether Email is spam or not, raw text messages, and file name. And each row represents one Email. We will use category as label (1: spam, 0: not spam), and raw text messages to extract features.

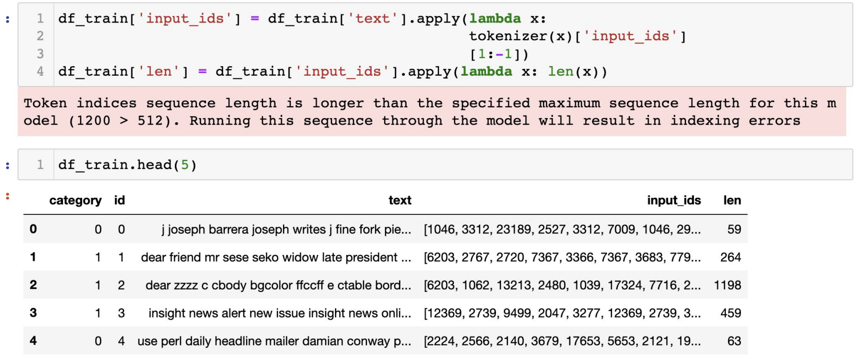


# CATEGORY	MESSAGE	FILE_NAME
1	Dear Homeowner, Interest Rates are at their lowest point in 40 years! We help you find the be...	00249.5f45607c1bffe89f60ba1ec9f878039a
1	ATTENTION: This is a MUST for ALL Computer Users!!! *NEW-Special Package Deal!* Norton SystemW...	00373.ebe8670ac56b04125c25100a36ab0510
1	This is a multi-part message in MIME format. ------=_NextPart_000_1CDC19_01C25366.4B57F3A0	00214.1367039e50dc6b7adb0f2aa8aba83216

Fit BERT - Tokenization

After text cleaning and tokenization, some Email has more than 512 tokens, which is longer than the capacity of base BERT.

So we need to truncate long sentence to several pieces, each having 510 tokens with [CLS] and [SEP] tokens added later.



```
df_train['input_ids'] = df_train['text'].apply(lambda x:
                                               tokenizer(x)['input_ids']
                                               [1:-1])
df_train['len'] = df_train['input_ids'].apply(lambda x: len(x))
```

```
Token indices sequence length is longer than the specified maximum sequence length for this m
odel (1200 > 512). Running this sequence through the model will result in indexing errors
```

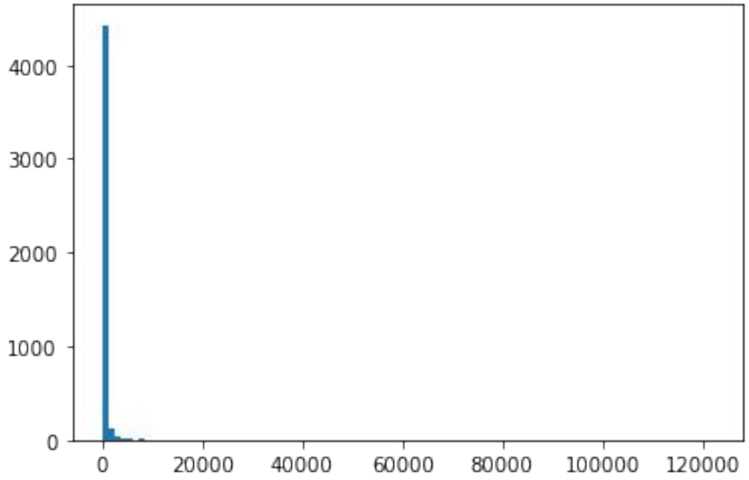
```
df_train.head(5)
```

	category	id	text	input_ids	len
0	0	0	j joseph barrera joseph writes j fine fork pie...	[1046, 3312, 23189, 2527, 3312, 7009, 1046, 29...	59
1	1	1	dear friend mr sese seko widow late president ...	[6203, 2767, 2720, 7367, 3366, 7367, 3683, 779...	264
2	1	2	dear zzzz c cbody bgcolor ffccff e ctable bord...	[6203, 1062, 13213, 2480, 1039, 17324, 7716, 2...	1198
3	1	3	insight news alert new issue insight news onli...	[12369, 2739, 9499, 2047, 3277, 12369, 2739, 3...	459
4	0	4	use perl daily headline mailer damian conway p...	[2224, 2566, 2140, 3679, 17653, 5653, 2121, 19...	63

Fit BERT - Tokenization

After text cleaning and tokenization, some Email has more than 512 tokens, which is longer than the capacity of base BERT.

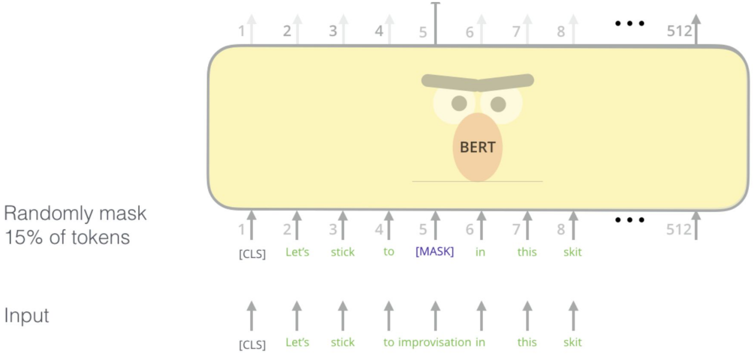
So we need to truncate long sentence to several pieces, each having 510 tokens with [CLS] and [SEP] tokens added later.



Fit BERT - Tokenization

After text cleaning and tokenization, some Email has more than 512 tokens, which is longer than the capacity of base BERT.

So we need to truncate long sentence to several pieces, each having 510 tokens with [CLS] and [SEP] tokens added later.

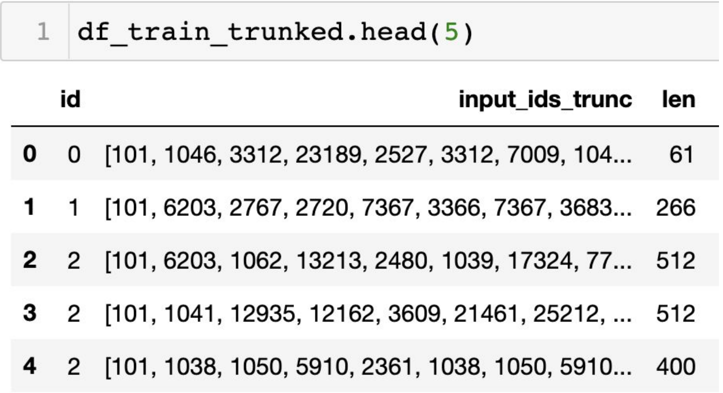


BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Fit BERT - Tokenization

After text cleaning and tokenization, some Email has more than 512 tokens, which is longer than the capacity of base BERT.

So we need to truncate long sentence to several pieces, each having 510 tokens with [CLS] and [SEP] tokens added later.

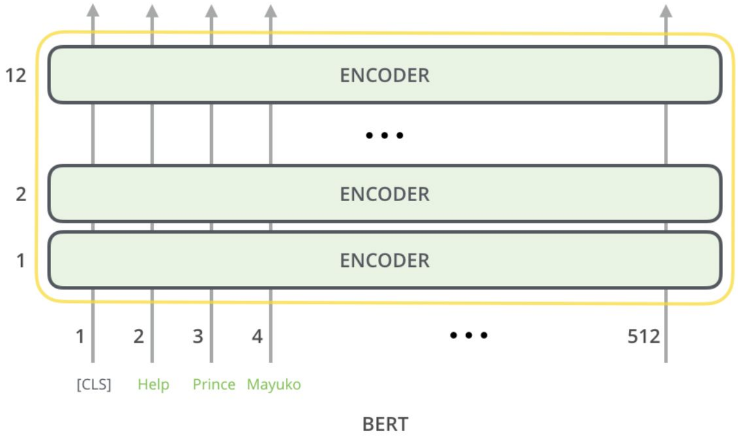


```
df_train_trunked.head(5)
```

	id	input_ids_trunc	len
0	0	[101, 1046, 3312, 23189, 2527, 3312, 7009, 104...	61
1	1	[101, 6203, 2767, 2720, 7367, 3366, 7367, 3683...	266
2	2	[101, 6203, 1062, 13213, 2480, 1039, 17324, 77...	512
3	2	[101, 1041, 12935, 12162, 3609, 21461, 25212, ...	512
4	2	[101, 1038, 1050, 5910, 2361, 1038, 1050, 5910...	400

Fit BERT - Feature Extraction

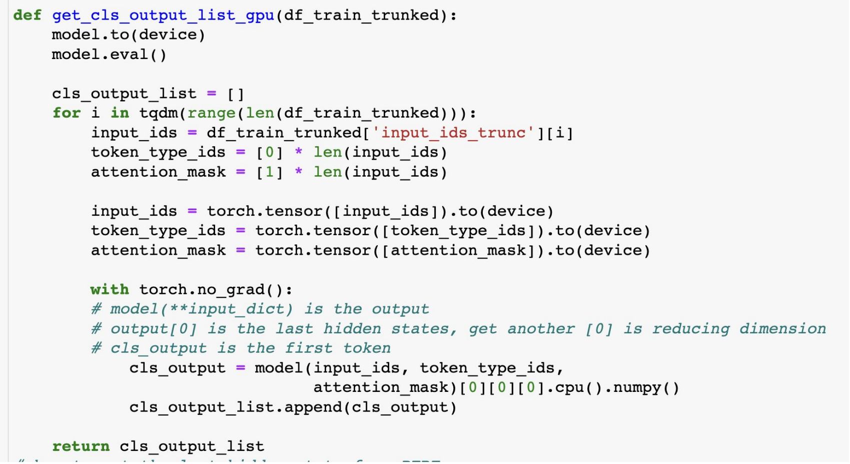
We used the feature of [CLS] position in the last hidden state, a 768-dimensional vector.



Fit BERT - Feature Extraction

We used the feature of [CLS] position in the last hidden state, a 768-dimensional vector.

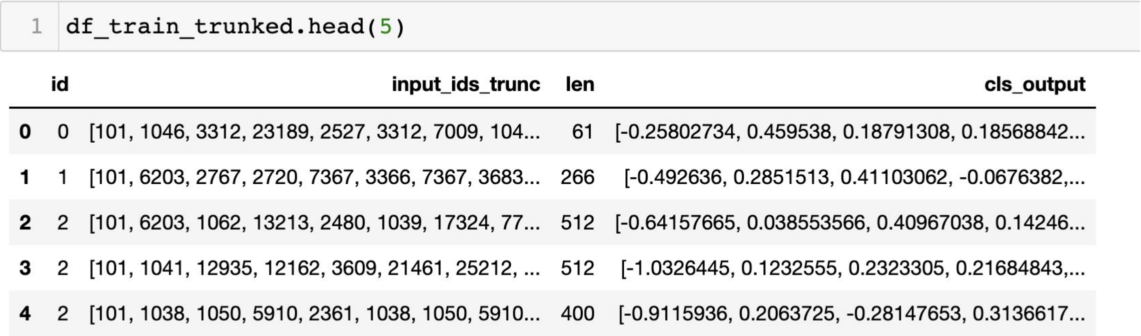
To fit the pre-trained BERT, besides input_id, token_type_ids and attention_mask need to be added.



```python
def get_cls_output_list_gpu(df_train_trunked):
    model.to(device)
    model.eval()

    cls_output_list = []
    for i in tqdm(range(len(df_train_trunked))):
        input_ids = df_train_trunked['input_ids_trunc'][i]
        token_type_ids = [0] * len(input_ids)
        attention_mask = [1] * len(input_ids)

        input_ids = torch.tensor([input_ids]).to(device)
        token_type_ids = torch.tensor([token_type_ids]).to(device)
        attention_mask = torch.tensor([attention_mask]).to(device)

        with torch.no_grad():
        # model(**input_dict) is the output
        # output[0] is the last hidden states, get another [0] is reducing dimension
        # cls_output is the first token
            cls_output = model(input_ids, token_type_ids,
                               attention_mask)[0][0][0].cpu().numpy()
            cls_output_list.append(cls_output)

    return cls_output_list
```

Fit BERT - Max Pooling

As each Email corresponding to one label, the dimension of long Email with multiple chunks needs to be reduced.

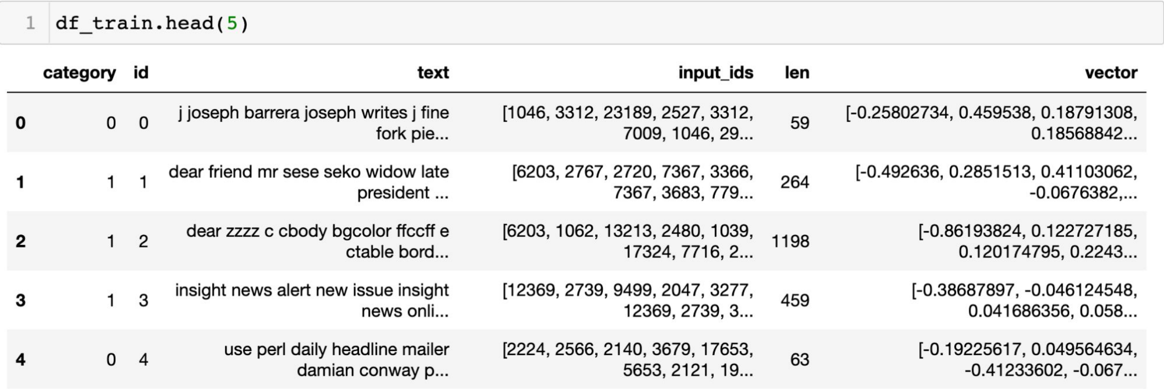


```
df_train_trunked.head(5)
```

	id	input_ids_trunc	len	cls_output
0	0	[101, 1046, 3312, 23189, 2527, 3312, 7009, 104...	61	[-0.25802734, 0.459538, 0.18791308, 0.18568842...
1	1	[101, 6203, 2767, 2720, 7367, 3366, 7367, 3683...	266	[-0.492636, 0.2851513, 0.41103062, -0.0676382,...
2	2	[101, 6203, 1062, 13213, 2480, 1039, 17324, 77...	512	[-0.64157665, 0.038553566, 0.40967038, 0.14246...
3	2	[101, 1041, 12935, 12162, 3609, 21461, 25212, ...	512	[-1.0326445, 0.1232555, 0.2323305, 0.21684843,...
4	2	[101, 1038, 1050, 5910, 2361, 1038, 1050, 5910...	400	[-0.9115936, 0.2063725, -0.28147653, 0.3136617...

Logistic Regression

Extract BERT feature for both training and testing data set.



```
df_train.head(5)
```

	category	id	text	input_ids	len	vector
0	0	0	j joseph barrera joseph writes j fine fork pie...	[1046, 3312, 23189, 2527, 3312, 7009, 1046, 29...	59	[-0.25802734, 0.459538, 0.18791308, 0.18568842...
1	1	1	dear friend mr sese seko widow late president ...	[6203, 2767, 2720, 7367, 3366, 7367, 3683, 779...	264	[-0.492636, 0.2851513, 0.41103062, -0.0676382,...
2	1	2	dear zzzz c cbody bgcolor ffccff e ctable bord...	[6203, 1062, 13213, 2480, 1039, 17324, 7716, 2...	1198	[-0.86193824, 0.122727185, 0.120174795, 0.2243...
3	1	3	insight news alert new issue insight news onli...	[12369, 2739, 9499, 2047, 3277, 12369, 2739, 3...	459	[-0.38687897, -0.046124548, 0.041686356, 0.058...
4	0	4	use perl daily headline mailer damian conway p...	[2224, 2566, 2140, 3679, 17653, 5653, 2121, 19...	63	[-0.19225617, 0.049564634, -0.41233602, -0.067...

Thank you