Обработка языка. Практика 4
Маскирование слов

Михаил Пожидаев

17 октября 2023 г.

Установка

```
pip install "tensorflow==2.8.0"
pip install torch
pip install transformers
```

Подготовка

```
from transformers import BertTokenizer, BertForMaskedLM
from torch.nn import functional as F
import torch
```

Инициализация

```
name = \
  'bert-base-multilingual-uncased'
tokenizer = \
  BertTokenizer.from_pretrained(name)
model = BertForMaskedLM.\
  from_pretrained(name, return_dict = True)
```

Вычисления

```
text = \
  "В университете студенты " + tokenizer.mask_token \
  + " целый день."
input =\
   tokenizer.encode_plus(text, return_tensors = "pt")
mask_index = torch.where(\
  input["input_ids"][0] == tokenizer.mask_token_id)
output = model(**input)
```

Вывод

```
logits = output.logits
softmax = F.softmax(logits, dim = -1)
mask_word = softmax[0, mask_index[0], :]
top = torch.topk(mask_word, 10)
for token in top[-1][0].data:
    print(tokenizer.decode([token]))
#проходят, провели, имеют, используют, прошли, на, вел
```

Спасибо за внимание!

Всё о курсе: https://marigostra.ru/materials/nlp.html
E-mail: msp@luwrain.org
Канал в Телеграм: https://t.me/MarigostraRu