Homework2: Question Answering

108201017 梁致銓

I. Data Preprocessing

1. 用 tf-idf 和 cosine similarity 選取和問題最相近且包含答案的兩個句子
2. 將兩個句子合併為一文本，作為 QA 的答案來源
3. Tokenized 問題和答案
4. 找出答案在文本中的位置，並給予 start 和 end index
5. 將句中答案的位置轉換成詞的位置

II. Model architecture

1. 模型採用："nreimers/MiniLM-L6-H384-uncased"
2. 模型架構：pretrained model 加上一層 linear layer

```
from transformers import BertModel

class QAModel(torch.nn.Module):

    def __init__(self):

        super(QAModel, self).__init__()

        self.bert = BertModel.from_pretrained("nreimers/MiniLM-L6-H384-uncased")
        self.fc = torch.nn.Linear(384, 2)


    def forward(self, input_ids, attention_mask, token_type_ids):

        output = self.bert(input_ids=input_ids, attention_mask=attention_mask, token_type_ids=token_type_ids, retu
        logits = output[0]
        out = self.fc(logits)

        return out
```

III. Training Process

1. Training epoch = 10
2. Batch size = 16
3. loss function = CrossEntropyLoss()
4. optimizer = AdamW with learning rate: 1e-4

III. Evaluation Scores

(valid data)

1. Longest Common Subsequence: accuracy: 0.796631
2. F1 score: 0.767724891293049

V. Result

	sentences	question	answer	start	end	predict
0	free flashcards authors studystack spain 1959...	spain 1959 , wrote dangerous summer , story r...	hemingway	101	110	hemingway
1	california facts , map state symbols enchante...	valley 282 feet sea level state lowest point ...	california	1	11	california
2	price convenience? atm surcharge debate jul 1...	like banks , many grocery stores dispensing c...	atms	127	131	atms
3	steamboat willy classic cartoons pinterest st...	voice mickey mouse steamboat willie	walt disney	135	146	walt disney
4	eastern europe see section 2 2 6 nation state...	eastern european capital city 2 2 million	bucharest	231	240	bucharest
5	us state longest shoreline?? guide humans ala...	6 , 640 miles coast , state longest shoreline	alaska	43	49	alaska
6	north dakota pumps 1 million barrels oil day ...	pumps one million barrels oil day , state	texas	167	172	north dakota
7	day day npr hear day day program march 20 , 2...	day day things considered among programs goin...	npr	9	12	npr
8	daffy known voice mel blanc 1937 1989 daffy d...	voice daffy duck first 50 years	mel blanc	19	28	mel blanc
9	region 4 russ nelson 's home page glossary ba...	braced framework carrying railroad chasm	trestle	93	100	russ nelson
10	twinkle chubbins astonishing adventures natur...	name laura bancroft , wrote twinkle chubbins ...	l frank baum	198	210	l frank baum
11	november 14 , 2008 eagerly crave rhymes wave ...	considered healthiest state 2006 , 's also ho...	minnesota	128	137	minnesota
12	cia headquarters named president bush cia hea...	headquarters compound langley , virginia name...	cia	1	4	center
13	puss boots shrek wikipedia puss boots fiction...	voiced puss boots shrek 2	antonio banderas	222	238	antonio
14	benet , william rose sgeresultat aalborg bibl...	william rose benet pulitzer dust god , brothe...	stephen vincent benet	117	138	stephen vincent benet
15	life begins 40 boeing blogs aug 17 , 2007 thi...	boeing manufacturing plant everett world 's l...	washington	151	161	airplanes
16	tnt apologises airing 'castle' bomb episode d...	tv cable network , explosive bombs	tnt	1	4	katic trip
17	mike judge beavis butthead voices google docs...	provided voices beavis butthead	mike judge	1	11	mike judge
18	aegisthus greek mythology greek mythology , a...	son agamemnon , avenged father 's death killi...	orestes	187	194	orestes
19	san antonio texas history timeline history st...	1718 texas town founded martin de alarcon fat...	san antonio	1	12	san antonio