1. After your model predicts the probability of answer span start/end position, what rules did you apply to determine the final start/end position? (the rules you applied must be different from the sample code)

1. 如果輸入文章大於 self.max_paragraph_len，Sample code 會使用 doc_stride 技術，將文章拆分成許多小的 windows 彼此會有 overlap 區域然後在每個 window 中找出答案 start 的機率與答案 end 的機率，兩者相加，並選擇其中一組機率相加最大的 windows 做為正確 Answer，但我另外增加了兩個條件分別是:
   1. start index 必須小於 end index
   2. 答案不能太長 → end index - start index 必須小於 25。

如下圖:

```
# Replace answer if calculated probability is larger than previous windows
if prob > max_prob and (end_index - start_index) < 25 and end_index > start_index:
        max_prob = prob
        # Convert tokens to chars (e.g. [1920, 7032] --> "大 金")
        answer = tokenizer.decode(data[0][0][k][start_index : end_index + 1])
```

2. Try another type of pretrained model which can be found in huggingface's Model Hub (e.g. BERT -> BERT-wwm-ext, or BERT -> RoBERTa), and describe

● the pretrained model you used

● performance of the pretrained model you used

● the difference between BERT and the pretrained model you used (architecture, pretraining loss, etc.)

1. 我有嘗試過 hfl/chinese-roberta-wwm-ext-large 以及 hfl/chinese-macbert-large。
2. hfl/chinese-roberta-wwm-ext-large: 0.78822
   hfl/chinese-macbert-large: 0.81605
3. Bert base : 做 Mask LM pretrained 任務以及 NSP(Next Sentence Prediction)任務，hidden size=768 ，num_attention_heads=12，num_hidden_layers=12。
   Roberta-wwm-ext-large : 與 bert 相比用更多的數據更大的 batch size 更久的時間 pretrained，並去掉 NSP 任務，訓練句子長度更長，並且採用動態 Mask 的方式，wwm 為使用 Whole Word Masking 方式，ext 為使用了更多的預訓練數據，而相比於 base model，Large model 的架構為 hidden size=1024，num_attention_heads=16，num_hidden_layers=24。
   Macbert-large : 對於 Mask LM 任務，macbert 做了一些修改，使用全詞 masked 以及 Ngram masked 策略，並對 15%比例的輸入詞進行 masking，其中 80%替換為同義詞，10%替換為隨機的單詞，剩下 10%則不變，並執行類似於 albert 所做的 NSP 預訓練任務，其架構為 hidden size=1024，num_attention_heads=16，num_hidden_layers=24。