

1. Your model design and concept

Ans: 我採用了兩種model，第一種是教授提供的n-gram model，我使用的是wiki_5M.arpa，是利用wikipedia的資料作訓練，第二種是我自己train的model，我以助教提供的“A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task”，下載他的資料來作訓練

2. Error Analysis and Discussion

Ans: 沒什麼太大的問題，不過我為了快速把全部在同一個資料夾底下的file讀近來花了一點時間，後來才知道有glob可以用

3. Compare and implement unsupervised method and supervised method

Ans: 沒有完成，抱歉

4. 請描述嘗試過的方法，並且討論曾經遇到的問題以及解決的方法

Ans: 一開始以自己的model做預測，accuracy大概0.41左右，後來上課聽完教授的講解，嘗試使用kenlm套件，並且增加標點符號的簡化，accuracy可以提高到接近0.45。

5. 請討論使用不同訓練資料量訓練n-gram language model對於預測克漏字的效能影響

Ans: 我自己的model有嘗試過兩種資料訓練量，分別是一百萬和五百萬這兩種，一百萬的accuracy大概落在0.39，五百萬大概落在0.41，教授的model也嘗試過兩種，分別是三百萬和五百萬，三百萬的accuracy大概落在0.44，五百萬大概落在0.45。可以得出資料量和accuracy有正相關。

6. 請討論使用不同domain的訓練資料訓練n-gram language model對於預測克漏字的效能影響

Ans: 同樣以五百萬筆資料來說，CNN stories的accuracy是0.41，wikipedia的accuracy是0.45，wikipedia的效果更好

Q1: Data processing

1. Tokenizer and build window

- a. Describe in detail about the tokenization algorithm you use. You need to explain what it does in your own ways

Ans: 我是用word-based tokenization，先將所有字母變小寫，接下來把特殊的標點符號換成空格，最後再把空格都split

2. Answer Option

- a. How did you convert the answer on characters to options on tokens after LM tokenization?

Ans: 我做完tokenization後，所有token都在一個list中，因此我只要挑出我要的token和option一起連再一起變成字串就可以了

- b. After your model predicts the probability of answer, what rules did you apply to determine the final option?

Ans: 選擇probability最高的

Q2: Modeling with Los and the variants








- a. your model (configuration of the model) and details

Ans: 我自己train的：以CNN stories去train

教授提供的：以wikipedia去train

- b. performance:

	wikipedia	CNN
3M/1M	0.44	0.39
5M	0.45	0.41

43	0716231		0.45241	4	7d
44	0716026		0.45119	5	1s
<div>  <div> <p>Your Best Entry!</p> <p>Your most recent submission scored 0.45119, which is an improvement of your previous score of 0.44946. Great job!</p> </div> <div> Tweet this </div> </div>					
45	0711278		0.45102	31	8d
46	0716308		0.44911	11	2h
47	0816038		0.44494	24	1d
	New_baseline		0.44216		