1. Data processing (1%)
Describe how do you use the data for `extractive.sh`, `seq2seq.sh`, `attention.sh`:
    a. How do you tokenize the data?
    如助教給的code，利用spacy的en_core_web_sm為我的語言模型，再利用spacy.tokenizer把每個英文句子拆成每個英文單詞
    b. Truncation length of the text and the summary.
    關於text，只取前面300個word
    關於summary，只取前面80個word
    c. The pre-trained embedding you used.
    用 glove.840b.300d 這個當作我的 pre-trained embedding model

2. Describe your extractive summarization model. (2%)
    a. your model
    $e_t = Embed(x_t)$
    $o_t,\ h_t = GRU(e_t,\ h_{t-1})$
    $\widehat{y}_t = Linear(o_t)$
    where
    x_t *is the input word of the text*
    o_t *is the t_th predict output from GRU*
    e_t *is the word embedding of the t_th token*
    h_t *is the t_th hidden layer of GRU*
    $\widehat{y}_t$ *is the probability of the words that* $0 < \widehat{y}_t < 1$
    b. performance of your model. (on the validation set)

	Rouge-1	Rouge-2	Rouge-L
Mean * 100	19.03	3.20	12.99
Std * 100	8.82	4.38	6.37

    c. the loss function you used.
    $loss = BCEWithLogitsLoss(\widehat{y}_t,\ y_t)$ *where*
    $\widehat{y}_t$ *is the probability of the words that* $0 < \widehat{y}_t < 1$
    y_t *is the ground truth for the probability of words that* $y_t = 0$ *or* 1
    d. The optimization algorithm (e.g. Adam), learning rate and batch size.
    Optimization: Adam
    Learning Rate: 0.001
    Batch Size: 64
    e. Post-processing strategy.

根據每個句子中預測多少 1 來當作該句子是否該選擇的機率：

$P(sentence) = \frac{(predict\ 1\ in\ sentence)}{(length\ of\ sentence)}$

如果該句子大於 threshold 則選擇該句子

舉例：

如果該句子預測出來為 [0, 0, 1, 1, 1]， 句子長度為 5， 故該句子被選擇的機率為 $\frac{3}{5}$， 而若預設的 threshold 為 0.5， 則選擇該句子。

3. Describe your Seq2Seq + Attention model. (2%)
   a. your model

      // encoder

      $e^x{}_t = Embed(x_t)$

      $o_t,\ h^e{}_t = EncoderGRU(e^x{}_t,\ h^e{}_{t-1})$

      $h^e{}_t = Tanh(h^e{}_t)$

      $a_t = Softmax(Attention(o_t,\ h^e{}_t))$

      // decoder

      $e^y{}_t = Embed(y_t)$

      $\widehat{y}_t,\ h^d{}_t = DecoderGRU(e^y{}_t,\ a_t,\ h^d{}_{t-1})$

      $where$

      $x_t,\ y_t\ is\ the\ input\ word\ of\ the\ text\ and\ summary,\ respectively$

      $e^x{}_t,\ e^y{}_t\ is\ the\ word\ embedding\ of\ the\ t\_th\ token\ for\ x_t\ and\ y_t,\ respectively$

      $h^e{}_t,\ h^d{}_t\ is\ the\ t\_th\ hidden\ layer\ of\ EncoderGRU\ and\ DecoderGRU,\ respectively$

      $o_t\ is\ the\ t\_th\ output\ from\ EncoderGRU$

      $a_t\ is\ the\ t\_th\ attention\ from\ Attention$

      $\widehat{y}_t\ is\ the\ t\_th\ predict\ output\ from\ DecoderGRU$
   b. performance of your model. (on the validation set)

	Rouge-1	Rouge-2	Rouge-L
Mean * 100	20.15	4.43	16.82
Std * 100	10.80	6.64	9.50

   c. the loss function you used.

      $loss = CrossEntropy(\widehat{y}_t,\ y_t)$

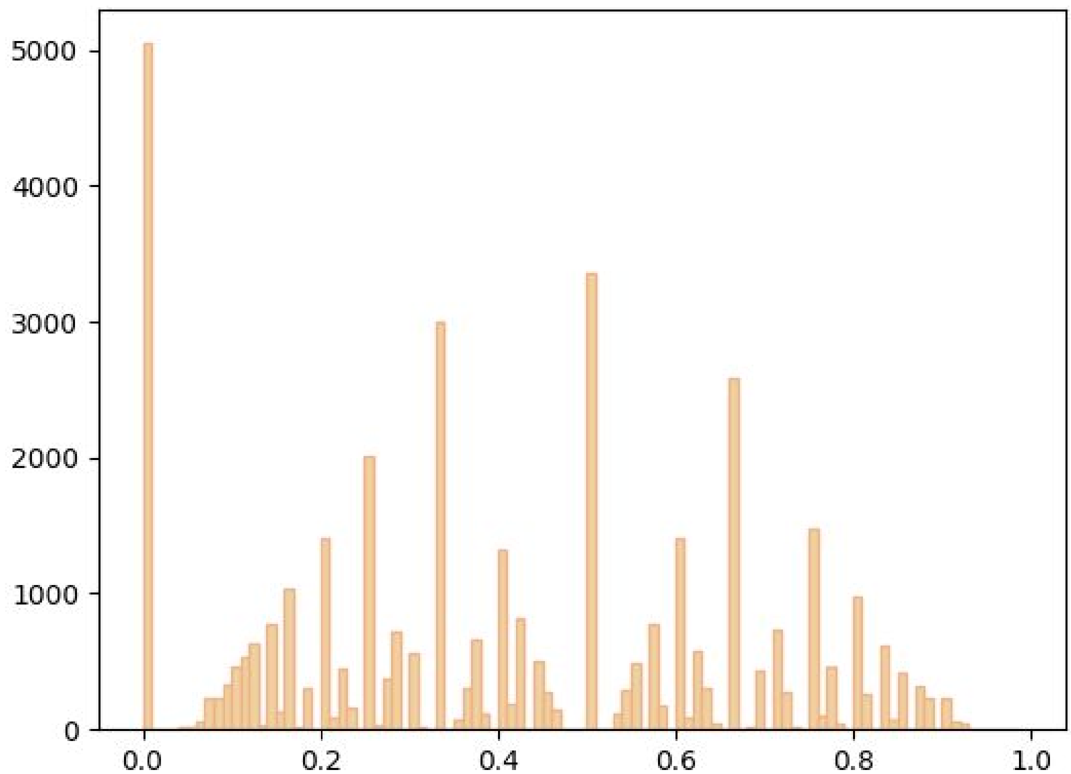
      $where$

      $\widehat{y}_t\ is\ the\ predict\ words$

      $y_t\ is\ the\ ground\ truth\ words$
   d. The optimization algorithm (e.g. Adam), learning rate and batch size.

Optimization: Adam
Learning Rate: 0.001
Batch Size: 64

4. Plot the distribution of relative locations of your predicted sentences by your extractive model, and describe your findings. (1%)
    a. X-axis: Relative Location [0, 1)
    b. Y-axis: Density
    c. For example:
       Prediction = [[0, 1, 3], [3, 6]]
       Relative Location = [[0, .25, .75], [.25, .5]]
       where # sentences in first and second documents are 4 and 12, respectively.



可以看到模型大多選擇開頭，或是靠文章中間的句子。
因為在英文的寫作中大約可分為兩種：
先說結論，也就是文章主旨的部分，之後才慢慢解釋原因，
或是先描述一段背景，中間再來個轉折，點出重點，
故這樣的統計我覺得非常合理。

5. Visualize the attention weights (2%).
    a. Take one example in the validation set and visualize the attention weights (after

softmax)

i. Readable text on the two axises. (0.5%)

pass

ii. Colors that indicate the value. (0.5%)

pass

b. Describe your findings. (1%)

發現幾乎都在關注開頭的字或是中間的字，或是一些轉折詞為重點，這感覺符合常見英文寫作的方式，先說結論，或是轉折時帶出重點。

6. Explain Rouge-L (1%)

a. Explain the way Rouge-L is calculated. (You don't need to explain what is covered in the ADA (algorithm design and analysis) course).

Rouge-L利用共同最長子序列(LCS)，計算兩個句子間的相似度

公式如下：

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

其中

$$R_{lcs} = \frac{LCS(X, Y)}{len(X)}$$

$$P_{lcs} = \frac{LCS(X, Y)}{len(Y)}$$

β是一個自己調的參數