MLDS HW2-2

組員 : 丁昱升、詹書愷、陳欽安、陳泓均

Data Process

1. data_processing.py
   - 將 clr_conversation 加上 EOS BOS 做成data frame
   - 將 clr_conversation 裡的台詞 轉換成對話形式(第一句對第二句 ; 第二句對第三句以此類推)做成training data set 的 dataframe
   - 後來 : 改用answer.txt, question.txt

Data Process

2. gensim_word_2_vec.py

- 運用gensim 內建的 Word2Vec model將 training dataframe (data_processing.py 做的) 轉成 embedding word to vector dictionary

  embedding size = 250 , window = 3, min_count = 7, batch_words = 256, iter = 10

Data Process

3. dict.py

- 把gensim word2vec的model，丟到nn.Embedding裡面，做成一個embedding layer

  (用於把word壓成vector)

- create word2idx, idx2word dictionary

Data Process

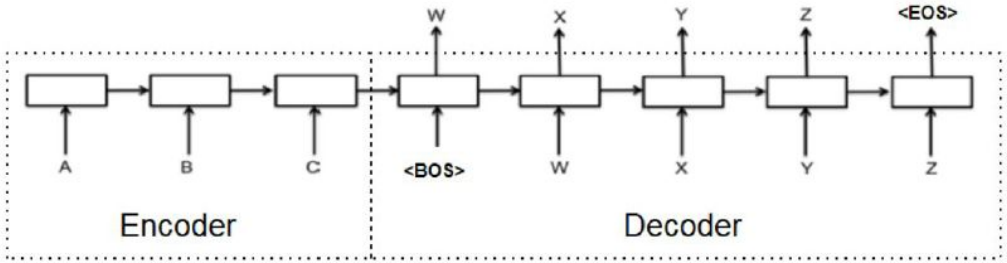
4. train.py

- 每一個epoch都重新作data loader : 把training_data做shuffule後把word改成index
- padding成長度15 後轉成tensor 做成dataloader

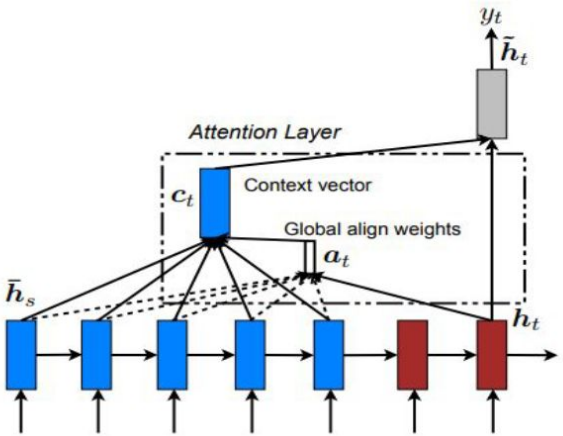
Training data choosing

- All of the training data
- Some of the training data
- Random choose some of the trainging data
- Random sample some data each epoch

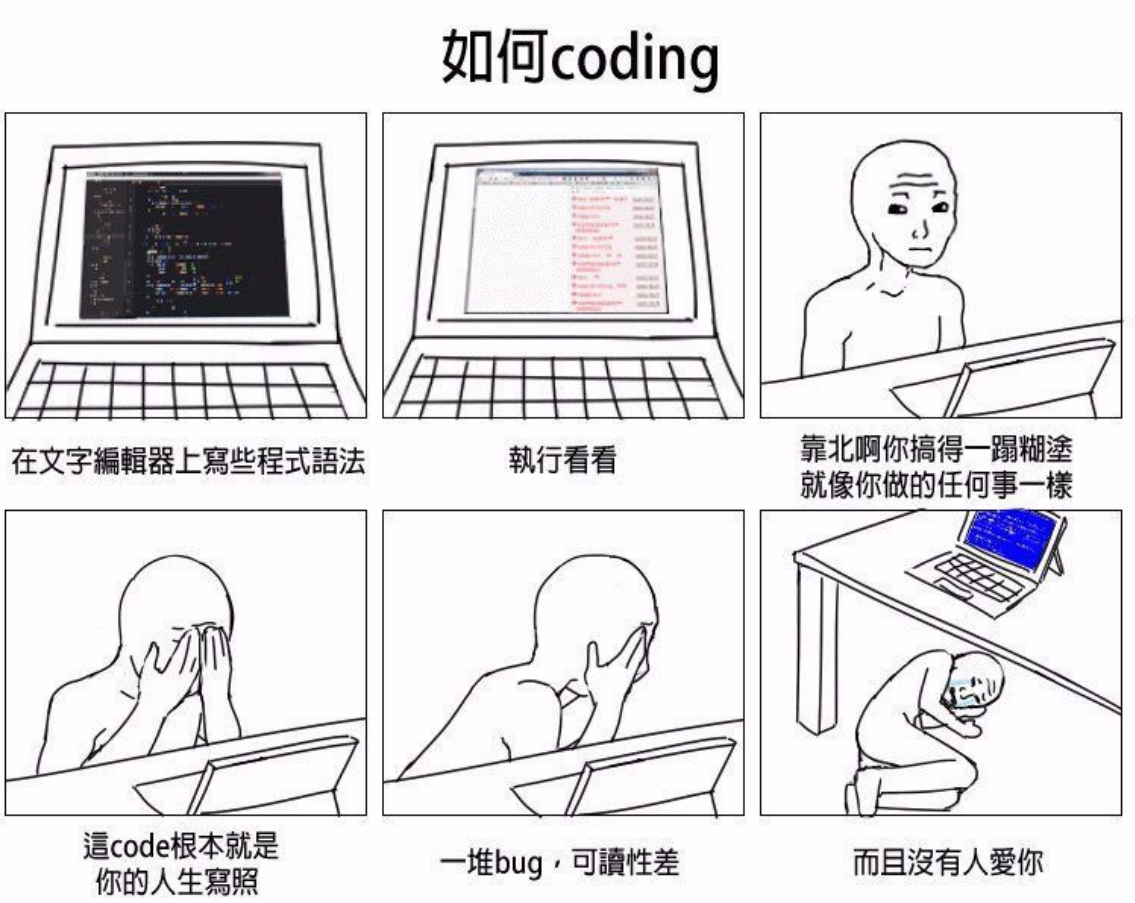
Model



- Embedding layer
- Single-layer LSTM
  - Encoder :
    - Input size : 250
    - Hidden size : 1024
  - Decoder :
    - Attention
    - Input size : 250(embedded word) + 1024(context vector)
- Schedule Sampling
  - ratio = 1 - (epoch_num / EPOCH) ^ 2

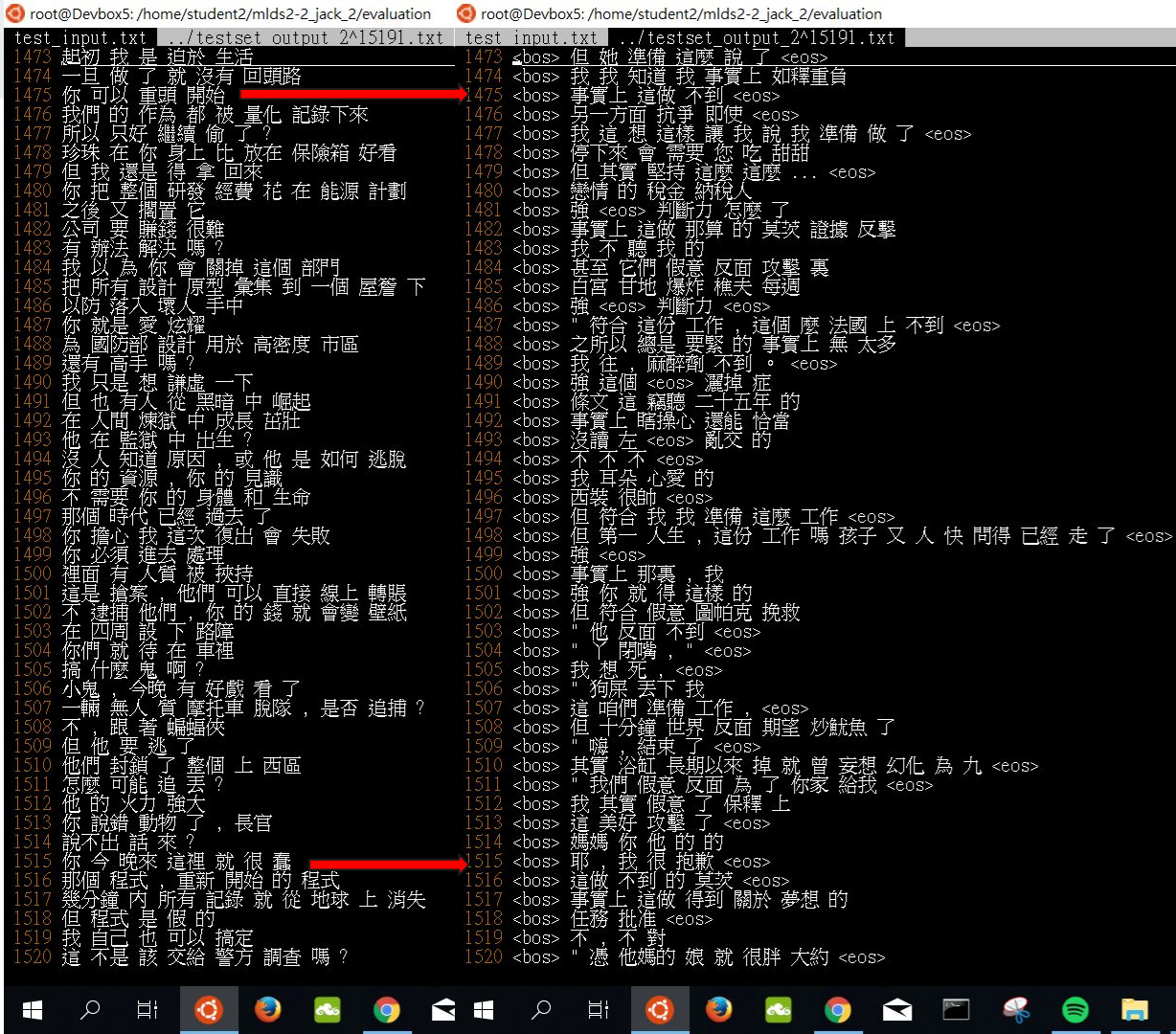


Result


如何coding
在文字編輯器上寫些程式語法
執行看看
靠北啊你搞得一蹋糊塗
就像你做的任何事一樣
這code根本就是
你的人生寫照
一堆bug，可讀性差
而且沒有人愛你

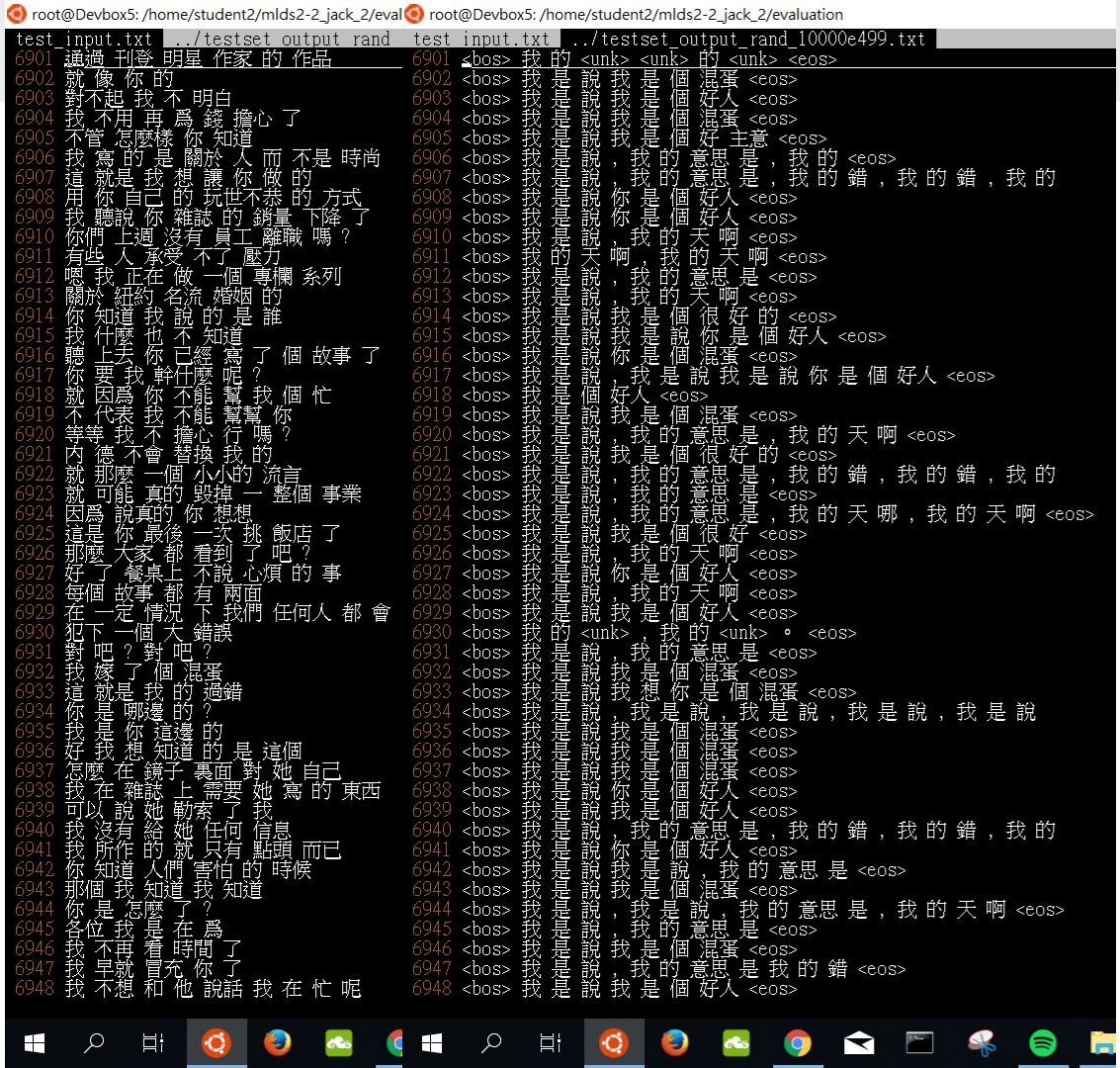
Result

- Perplexity : 697.585920
- Score : 0.21834
- Sample 2^15 trainig data
- 192 epoch



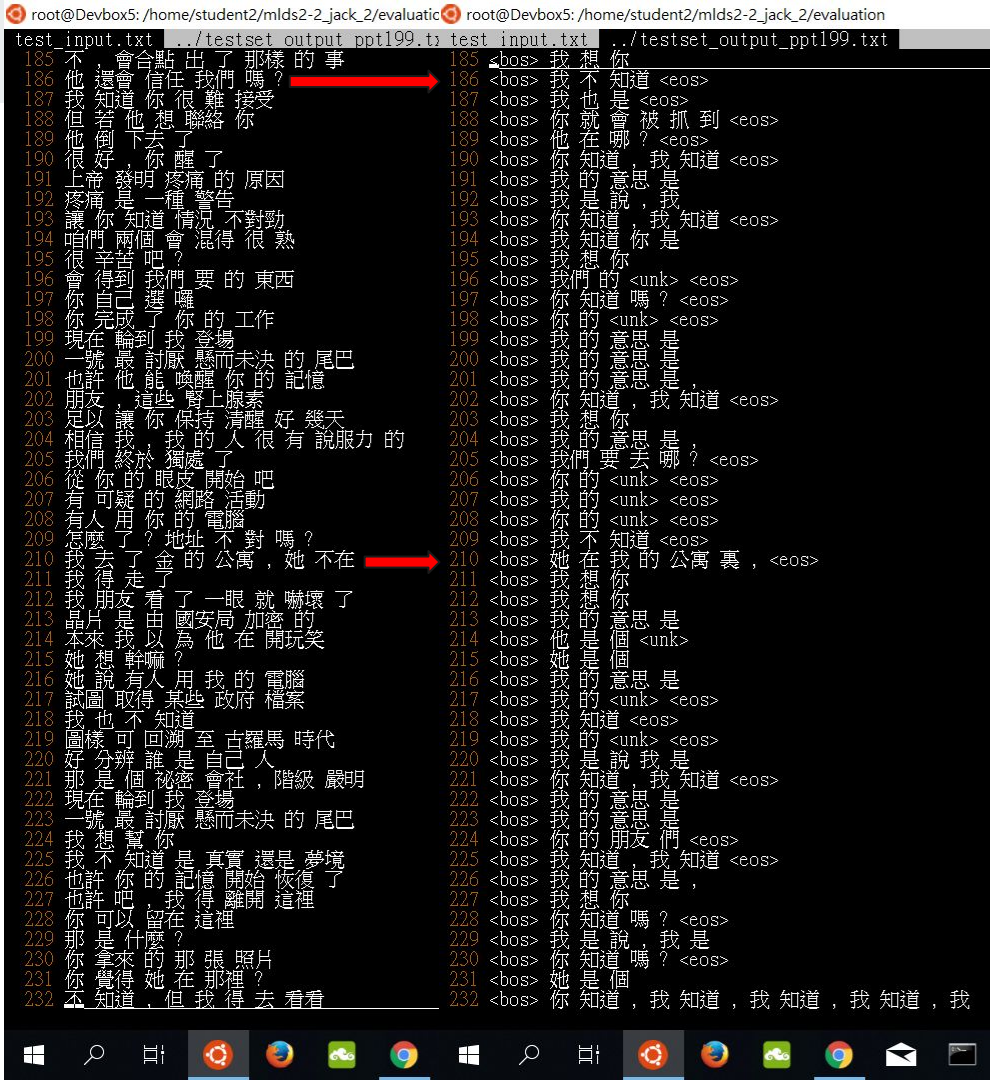
Result

- Perplexity : 188.128145
- Score : 0.17331
- Sample 10000 trainig data
- 500 epoch
- 10000 epoch isn't better



Result

- Perplexity : 1228.416932
- Score : 0.15567
- Sample 8000 training data
  - (every epoch)
- 200 epoch



Result

- Perplexity : 20.542
- Score : 0.55720
- All in
- 20 epoch

root@Devbox5: /home/student2/mlds2-2_jack_2/evaluation

test input.txt ../testset_output_new_19.txt

```
8242 <bos> 好 的 <eos>
8243 <bos> 我 的 天 <eos>
8244 <bos> 帶來 一種 信念 <eos>
8245 <bos> 但 你 可是 超級 酷 的 女孩子 <eos>
8246 <bos> 我 保證 不會 去 看 的 <eos>
8247 <bos> 別 那麼 做 享受 了 , <eos>
8248 <bos> 聽著 , 山姆 <eos>
8249 <bos> 他 看起來 像 他 一樣 <eos>
8250 <bos> 不 , 不 , 我 沒事 <eos>
8251 <bos> 你 知道 就 只有 我 <eos>
8252 <bos> 如果 我 是 在 跳 不 喝 的 , 就 像 <eos>
8253 <bos> 你 和 你 的 家人 都 愛 你 的 父親 <eos>
8254 <bos> 你 是 說 , 她 會 把 所有 的 回憶 都 完美 <eos>
8255 <bos> 福吉 <eos>
8256 <bos> 好好 保重 , 咱們 醫院 見 <eos>
8257 <bos> 好 的 <eos>
8258 <bos> 我 不 知道 。 <eos>
8259 <bos> 馬力 上 , 沒錯 <eos>
8260 <bos> 但 現在 的 情況 是 <eos>
8261 <bos> 我 的 問題 在 哪 ? <eos>
8262 <bos> 他 怎麼 會 知道 那 事 <eos>
8263 <bos> 游泳池 <eos>
8264 <bos> 嘿 , 小 公主 <eos>
8265 <bos> 你 到底 在 想 什麼 ? <eos>
8266 <bos> 但 我們 分手 了 , 你 就 得 稍微 享受 一下 <eos>
8267 <bos> 你 開玩笑 <eos>
8268 <bos> 他 在 大街 上 <eos>
8269 <bos> 我 對 自己 更 有 信心 <eos>
8270 <bos> 應該 不會 是 小孩 <eos>
8271 <bos> 我 在 想 袋子 裏 有沒有 更好 的 事情 <eos>
8272 <bos> 而且 在 這 簍子 的 同時 <eos>
8273 <bos> 我 是 說 , 有個 孩子 <eos>
8274 <bos> 這種 人 不 重視 的 <eos>
8275 <bos> 那 是 誰 ? <eos>
8276 <bos> 我 老爸 <eos>
8277 <bos> 我 剛剛 從 以前 那 地方 就是 想 看到 的 <eos>
8278 <bos> 是不是 因爲 他 是 神 <eos>
8279 <bos> 我 需要 知道 一件 事情 <eos>
8280 <bos> 大概 半個 小時 前 <eos>
8281 <bos> 如果 我 告訴 你 <eos>
8282 <bos> 我 也 不 知道 該 怎麼 做 <eos>
8283 <bos> 好 的 <eos>
8284 <bos> 不要 , 不要 , 不要 了 。 <eos>
8285 <bos> 你 來 開車 <eos>
8286 <bos> 是 什麼 ? <eos>
8287 <bos> 好 吧 , 嗯 ? <eos>
8288 <bos> 彌補 他 的 時間 , <eos>
8289 <bos> 後果 不算 短 <eos>
```

Problem

- Dataset too huge, long training time
- Dataset is not very good
- We don’t know which way is better in choosing dataset
- min_count = 7 (?)

Github link

https://github.com/kaichan1201/MLDS_2019_Spring/tree/master/hw2