



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

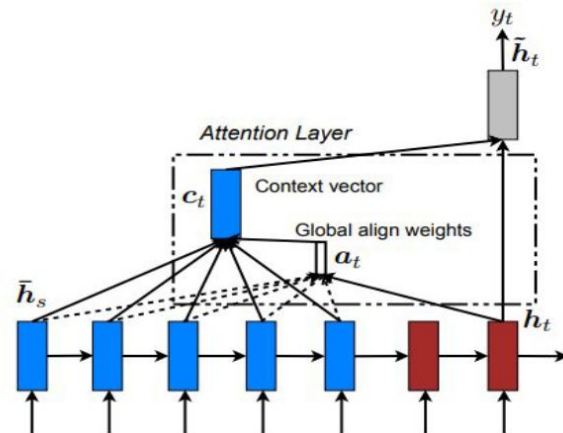
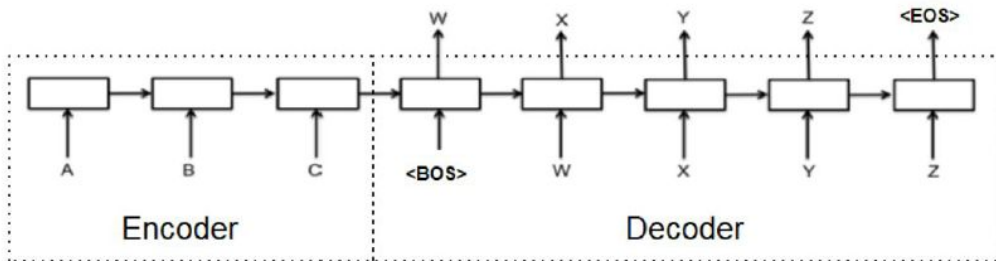


Training data choosing

- All of the training data
- Some of the training data
- Random choose some of the training 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
 - $\text{ratio} = 1 - (\text{epoch_num} / \text{EPOCH})^2$



如何coding

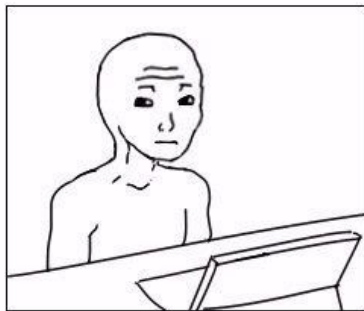
Result



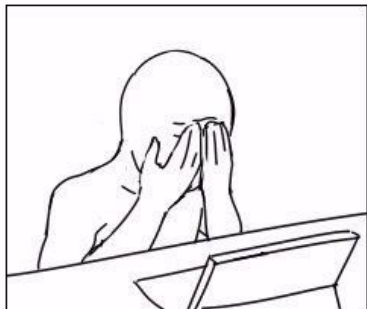
在文字編輯器上寫些程式語法



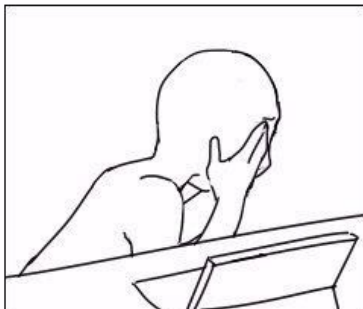
執行看看



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

root@Devbox5:/home/student2/mlids2-2_jack_2/evaluation

```
test input.txt ../testset output 2^15191.txt
1473 起初我是迫於生活
1474 一旦做了就沒有回頭路
1475 你可以重頭開始
1476 我們的作為都被量化記錄下來
1477 所以只好繼續偷了？
1478 珍珠在你身上比放在保險箱好看
1479 但我還是得拿回來
1480 你把整個研發經費花在能源計劃
1481 之後又擱置它
1482 公司要賺錢很難
1483 有辦法解決嗎？
1484 我以為你會關掉這個部門
1485 把所有設計原型彙集到一個屋簷下
1486 以防落入壞人手中
1487 你就是愛炫耀
1488 為國防部設計用於高密度市區
1489 還有高手嗎？
1490 我只是想謙虛一下
1491 但也有人從黑暗中崛起
1492 在人間煉獄中成長茁壯
1493 他在監獄中出生？
1494 沒人知道原因，或他是如何逃脫
1495 你的資源，你的見識
1496 不需要你的身體和生命
1497 那個時代已經過去了
1498 你擔心我這次復出會失敗
1499 你必須進去處理
1500 裡面有人會被挾持
1501 這是搶案，他們可以直接線上轉賬
1502 不逮捕他們，你的錢就會變壁紙
1503 在四周設下路障
1504 你們就待在車裡
1505 搞什麼鬼啊？
1506 小鬼，今晚有好戲看了
1507 一輛無人質摩托車脫隊，是否追捕？
1508 不，跟著蝙蝠俠
1509 但他要逃了
1510 他們封鎖了整個上西區
1511 怎麼可能追丟？
1512 他的火力強大
1513 你說錯動物了，長官
1514 說不出話來？
1515 你今晚來這裡就很蠢
1516 那個程式，重新開始的程式
1517 幾分鐘內，所有記錄就從地球上消失
1518 但程式是假的
1519 我自己也可以搞定
1520 這不是該交給警方調查嗎？
```

root@Devbox5:/home/student2/mlids2-2_jack_2/evaluation

```
test input.txt ../testset output 2^15191.txt
1473 <bos>但她準備這麼說了<eos>
1474 <bos>我我知道我事實上如釋重負
1475 <bos>事實上這做不到<eos>
1476 <bos>另一方面抗爭即使<eos>
1477 <bos>我這想這樣讓我說我準備做了<eos>
1478 <bos>停下來會需要您吃甜甜
1479 <bos>但其實堅持這麼這麼...<eos>
1480 <bos>戀情的稅金納稅人
1481 <bos>強<eos>判斷力怎麼了
1482 <bos>事實上這做那算的莫茨證據反擊
1483 <bos>我不聽我的
1484 <bos>甚至它們假意反面攻擊裏
1485 <bos>百官甘地爆炸樵夫每週
1486 <bos>強<eos>判斷力<eos>
1487 <bos>"符合這份工作，這個麼法國上不到<eos>
1488 <bos>之所以總是要緊的事實上無太多
1489 <bos>我往，麻醉劑不到。<eos>
1490 <bos>強這個<eos>灑掉症
1491 <bos>條文這竊聽二十五年的
1492 <bos>事實上暗操心還能恰當
1493 <bos>沒讀左<eos>亂交的
1494 <bos>不不不<eos>
1495 <bos>我耳朵心愛的
1496 <bos>西裝很帥<eos>
1497 <bos>但符合我準備這麼工作<eos>
1498 <bos>但第一人生，這份工作嗎孩子又人快問得已經走了<eos>
1499 <bos>強<eos>
1500 <bos>事實上那裏，我
1501 <bos>強你就得這樣的
1502 <bos>但符合假意圖帕克挽救
1503 <bos>"他反面不到<eos>
1504 <bos>"丫閉嘴，"<eos>
1505 <bos>我想死，<eos>
1506 <bos>"狗屎丟下我
1507 <bos>這咱們準備工作，<eos>
1508 <bos>但十分鐘世界反面期望炒魷魚了
1509 <bos>"噯，結束了<eos>
1510 <bos>其實浴缸長期以來掉就曾妄想幻化為九<eos>
1511 <bos>"我們假意反面為了你家給我<eos>
1512 <bos>我其實假意了保釋上
1513 <bos>這美好攻擊了<eos>
1514 <bos>媽媽你他的的
1515 <bos>耶，我很抱歉<eos>
1516 <bos>這做不到的莫茨<eos>
1517 <bos>事實上這做得到關於夢想的
1518 <bos>任務批准<eos>
1519 <bos>不，不對
1520 <bos>"憑他媽的娘就很胖大約<eos>
```

Result

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

```
root@devbox5:/home/student2/mlsd2-2_jack_2/eval test input.txt ../testset output rand
6901 通過刊登明星作家的作品 <unk> <unk> <eos>
6902 就像你的 <bos> 我是說我是個混蛋 <eos>
6903 對不起我不明白 <bos> 我是說我是個好人 <eos>
6904 我不再為錢擔心了 <bos> 我是說我是個混蛋 <eos>
6905 不管怎麼樣你知道 <bos> 我是說我是個好主意 <eos>
6906 我寫的是關於人而不是時尚 <bos> 我是說，我的意思是，我的 <eos>
6907 這就是我想讓你做的 <bos> 我是說，我的意思是，我的錯，我的錯，我的
6908 這用你自己的玩世不恭的方式 <bos> 我是說，你是個好人 <eos>
6909 我聽說你雜誌的銷量下降了 <bos> 我是說，你是個好人 <eos>
6910 你們上週沒有員工離職嗎？ <bos> 我是說，我的天啊 <eos>
6911 有些人承受不了壓力 <bos> 我是說，我的天啊 <eos>
6912 喝我正在做一個專欄系列 <bos> 我是說，我的意思是 <eos>
6913 關於紐約名流婚姻的 <bos> 我是說，我的天啊 <eos>
6914 你知道我說的是誰 <bos> 我是說，我是個很好的 <eos>
6915 我什麼也不知道 <bos> 我是說，你是個好人 <eos>
6916 聽上去你已經寫了個故事了 <bos> 我是說，你是個混蛋 <eos>
6917 你要我幹什麼呢？ <bos> 我是說，我是說我是說你是個好人 <eos>
6918 你因為我不能幫我個忙 <bos> 我是說，我是個好人 <eos>
6919 你代表我不能幫幫你 <bos> 我是說，我是個混蛋 <eos>
6920 等等我不擔心行嗎？ <bos> 我是說，我的意思是，我的天啊 <eos>
6921 內德不會替換我的 <bos> 我是說，我是個很好的 <eos>
6922 就那麼一個小小的流言 <bos> 我是說，我的意思是，我的錯，我的錯，我的
6923 就可能真的毀掉一整個事業 <bos> 我是說，我的意思是 <eos>
6924 因為說真的你想想 <bos> 我是說，我的意思是，我的天哪，我的天啊 <eos>
6925 這是你最後一次挑飯店了 <bos> 我是說，我是個很好 <eos>
6926 那麼大家都看到了吧？ <bos> 我是說，我的天啊 <eos>
6927 好，餐桌上不說心煩的事 <bos> 我是說，你是個好人 <eos>
6928 每個故事都有兩面 <bos> 我是說，我的天啊 <eos>
6929 在特定情況下我們任何人都會 <bos> 我是說，我是個好人 <eos>
6930 犯下一個大錯誤 <bos> 我是說，我的 <unk>。 <eos>
6931 對吧？對吧？ <bos> 我是說，我的意思是 <eos>
6932 我嫁了個混蛋 <bos> 我是說，我是個混蛋 <eos>
6933 這就是我的過錯 <bos> 我是說，我想你是個混蛋 <eos>
6934 這是哪邊的？ <bos> 我是說，我是說，我是說，我是說
6935 我是指這邊的 <bos> 我是說，我是個混蛋 <eos>
6936 好我想知道的是這個 <bos> 我是說，我是個混蛋 <eos>
6937 怎麼在鏡子裏面對她自己 <bos> 我是說，我是個混蛋 <eos>
6938 我在雜誌上需要她寫的東西 <bos> 我是說，你是個好人 <eos>
6939 可以說她勒索了我 <bos> 我是說，我是個好人 <eos>
6940 我沒有給她任何信息 <bos> 我是說，我的意思是，我的錯，我的錯，我的
6941 我所作的就只有點頭而已 <bos> 我是說，你是個好人 <eos>
6942 你知道人們害怕的時候 <bos> 我是說，我的意思是 <eos>
6943 那個我知道我知道 <bos> 我是說，我是個混蛋 <eos>
6944 你是怎麼了？ <bos> 我是說，我是說，我的意思是，我的天啊 <eos>
6945 各位我是在為 <bos> 我是說，我的意思是 <eos>
6946 我不再看時間了 <bos> 我是說，我是個混蛋 <eos>
6947 我早就冒充你了 <bos> 我是說，我的意思是我的錯 <eos>
6948 我不想和他說話我在忙呢 <bos> 我是說，我是個好人 <eos>
```

Result

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

```
test input.txt ../testset_output_ppt199.txt test input.txt ../testset_output_ppt199.txt
185 <bos> 我想你 185 <bos> 我想你
186 <bos> 我 知道 <eos> 186 <bos> 我 知道 <eos>
187 <bos> 我 不 是 <eos> 187 <bos> 我 不 是 <eos>
188 <bos> 我 就 會 被 抓 到 <eos> 188 <bos> 我 就 會 被 抓 到 <eos>
189 <bos> 他 在 哪 ？ <eos> 189 <bos> 他 在 哪 ？ <eos>
190 <bos> 你 知 道 ； 我 知 道 <eos> 190 <bos> 你 知 道 ； 我 知 道 <eos>
191 <bos> 的 意 思 是 191 <bos> 的 意 思 是
192 <bos> 我 是 說 ； 我 知 道 <eos> 192 <bos> 我 是 說 ； 我 知 道 <eos>
193 <bos> 你 知 道 ； 我 知 道 <eos> 193 <bos> 你 知 道 ； 我 知 道 <eos>
194 <bos> 我 知 道 ； 你 是 194 <bos> 我 知 道 ； 你 是
195 <bos> 我 想 你 195 <bos> 我 想 你
196 <bos> 我 們 的 <unk> <eos> 196 <bos> 我 們 的 <unk> <eos>
197 <bos> 你 知 道 嗎 ？ <eos> 197 <bos> 你 知 道 嗎 ？ <eos>
198 <bos> 你 的 <unk> <eos> 198 <bos> 你 的 <unk> <eos>
199 <bos> 的 意 思 是 199 <bos> 的 意 思 是
200 <bos> 我 的 意 思 是 200 <bos> 我 的 意 思 是
201 <bos> 我 的 意 思 是 201 <bos> 我 的 意 思 是
202 <bos> 你 知 道 ； 我 知 道 <eos> 202 <bos> 你 知 道 ； 我 知 道 <eos>
203 <bos> 我 想 你 是 199 <bos> 我 想 你 是
204 <bos> 的 意 思 是 204 <bos> 的 意 思 是
205 <bos> 我 們 要 去 哪 ？ <eos> 205 <bos> 我 們 要 去 哪 ？ <eos>
206 <bos> 你 的 <unk> <eos> 206 <bos> 你 的 <unk> <eos>
207 <bos> 的 <unk> <eos> 207 <bos> 的 <unk> <eos>
208 <bos> 你 的 <unk> <eos> 208 <bos> 你 的 <unk> <eos>
209 <bos> 我 不 知 道 <eos> 209 <bos> 我 不 知 道 <eos>
210 <bos> 她 不 在 我 的 公 寓 裏 ； <eos> 210 <bos> 她 不 在 我 的 公 寓 裏 ； <eos>
211 <bos> 我 想 你 199 <bos> 我 想 你
212 <bos> 我 想 你 199 <bos> 我 想 你
213 <bos> 我 的 意 思 是 213 <bos> 我 的 意 思 是
214 <bos> 他 是 個 <unk> 214 <bos> 他 是 個 <unk>
215 <bos> 她 是 個 215 <bos> 她 是 個
216 <bos> 我 的 意 思 是 216 <bos> 我 的 意 思 是
217 <bos> 我 的 <unk> <eos> 217 <bos> 我 的 <unk> <eos>
218 <bos> 我 知 道 <eos> 218 <bos> 我 知 道 <eos>
219 <bos> 我 的 <unk> <eos> 219 <bos> 我 的 <unk> <eos>
220 <bos> 我 是 說 我 是 220 <bos> 我 是 說 我 是
221 <bos> 你 知 道 ； 我 知 道 <eos> 221 <bos> 你 知 道 ； 我 知 道 <eos>
222 <bos> 我 的 意 思 是 222 <bos> 我 的 意 思 是
223 <bos> 我 的 意 思 是 223 <bos> 我 的 意 思 是
224 <bos> 你 的 朋 友 們 <eos> 224 <bos> 你 的 朋 友 們 <eos>
225 <bos> 我 知 道 ； 我 知 道 <eos> 225 <bos> 我 知 道 ； 我 知 道 <eos>
226 <bos> 的 意 思 是 226 <bos> 的 意 思 是
227 <bos> 我 想 你 199 <bos> 我 想 你
228 <bos> 你 知 道 嗎 ？ <eos> 228 <bos> 你 知 道 嗎 ？ <eos>
229 <bos> 我 是 說 ； 我 是 229 <bos> 我 是 說 ； 我 是
230 <bos> 你 知 道 嗎 ？ <eos> 230 <bos> 你 知 道 嗎 ？ <eos>
231 <bos> 你 是 個 231 <bos> 你 是 個
232 <bos> 你 知 道 ； 我 知 道 ； 我 知 道 ； 我 知 道 ； 我 232 <bos> 你 知 道 ； 我 知 道 ； 我 知 道 ； 我 知 道 ； 我
```

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> 福音 <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> 而且在這簽字的同時 <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> 是不是因為他是神 <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