仿 PTT 鄉民問答 Bot

108321030 劉承熙
108321032 王廷郡

動機

- 因為平常喜歡看 PTT 的鄉民梗，雖然有時候PTT鄉民發言時常政治不正確，但是有些用語很有趣而且很特別，算是一種次文化
- 剛好 Kaggle 上有人整理 PTT 中文語料庫 PTT-Gossiping-Corpus

Dataset - 資料來源

- PTT-Gossiping-Corpus 的資料太舊，最新的資料還停留在2019年
  - 網路流行語變化很快，可能不太適用
  - GitHub 爬蟲程式無法運作
- 自己寫爬蟲程式
  - 爬取八卦板中分類為 ”[問卦]” 的「標題」及「推文」
  - 每 2.5 秒爬一個頁面，爬了大約五天
  - 取 2022-02-01 ~ 2022-05-28 共計 108900 篇文章

2018年度10大PTT鄉民流行用語揭曉

排名	用語	網友怎麼說	網路聲量
1	韓導	「被政治耽誤的導演」	72,189
2	1124滅東廠	「年底用選票給民進黨教訓」	34,245
3	發大財	「OO進的來，OO出的去，OO發大財！」	31,849
4	又老又窮	「台灣有哪裡不是又老又窮？」	25,601
5	越想越不對勁	「台男要好好保護自己」	20,043
6	主流民意	「大家好我們是對同志不友善的中國台北人」	17,817
7	喜韓兒	「韓粉出征，寸草不生」	16,914
8	貪婪老人	「1500=台北價值」	16,765
9	擊潰丁守中	「擊潰丁守中！」	13,304
10	關三天	「政客、名嘴要不要先關一關？」	9,973

資料分析：DailyView網路溫度計 透過 KEYPO大數據關鍵引擎(keypo.tw)，
以國際級的語意分析架構、先進的機器學習技術與人工智慧推論引擎，
感知網友語意脈絡與情緒，分析時事網路大數據。
分析期間: 2017/12/17~2017/12/16

網路溫度計

Dataset - QA對應

- 每篇問卦的文章有不定數量的推文，要找出適當的回應當作回應
  - 「標題」: 問題(Q)
  - 「推文」: 回應(A)
- 使用詞出現的數量當作 weight，每個推文計算加權分數，取最高者

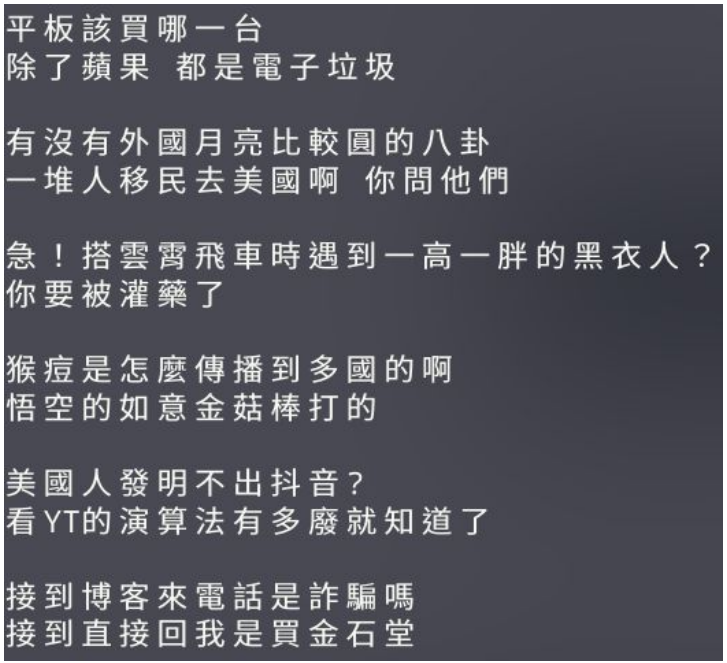
{‘貓’: 4, ‘可愛’: 1, ‘這’: 1, ‘隻’: 1, ……}

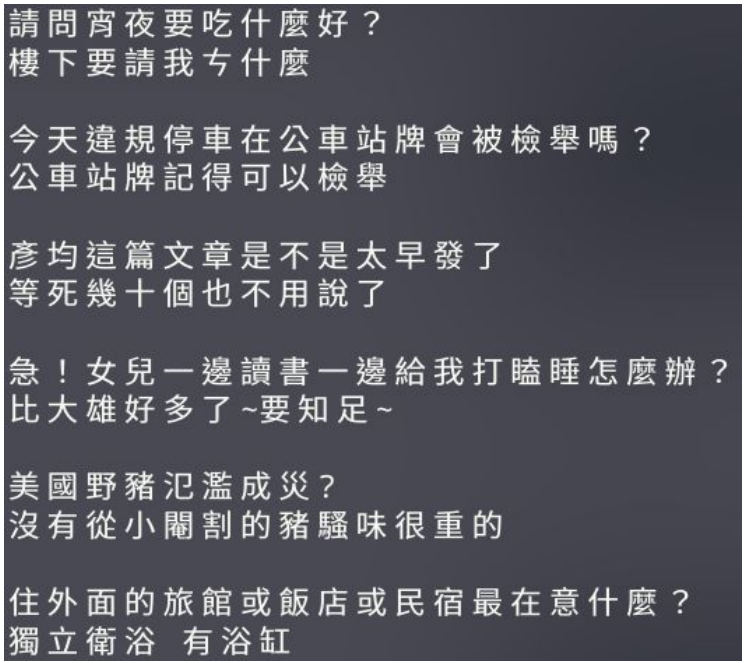
推 TPDC: 貓貓可愛
推 horseorange: 這隻貓太肥了吧
噓 iamshiba: 拿貓騙推

$$\frac{4 + 4 + 1}{3}$$

Dataset - QA對應

- 無法作到完美，但是大多數都看得出來問答對應關係





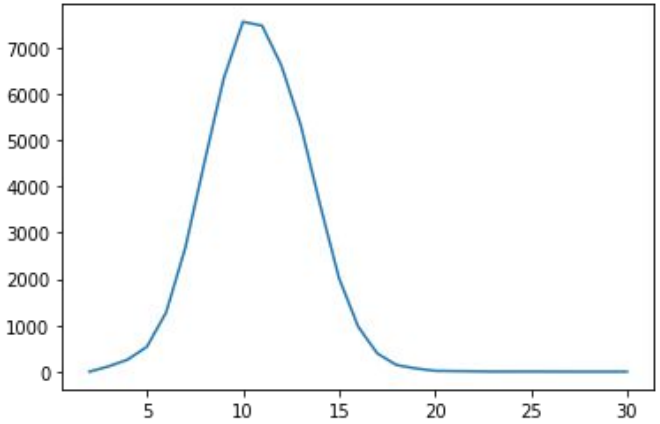
Dataset - PTT 常見詞

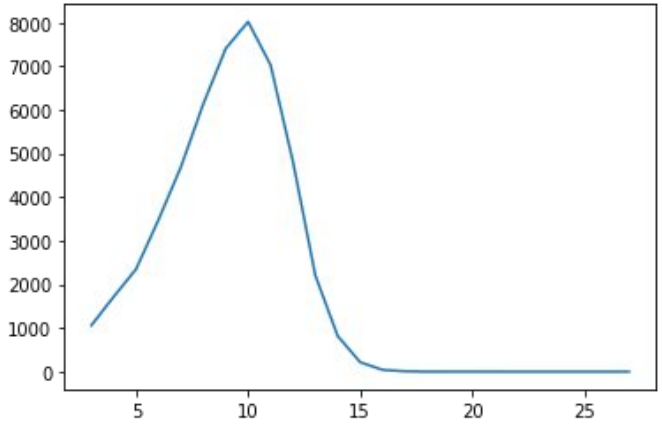
- Python WordCloud 套件

Preprocess - Word-based

- 為所有詞建立 Tokenizer，總共 56130 個詞
- 分析 Q 和 A 的 token 數量分佈，大都落在 7~13 左右
  - QA 最長都可能到 30，因此補 padding token 到長度 30

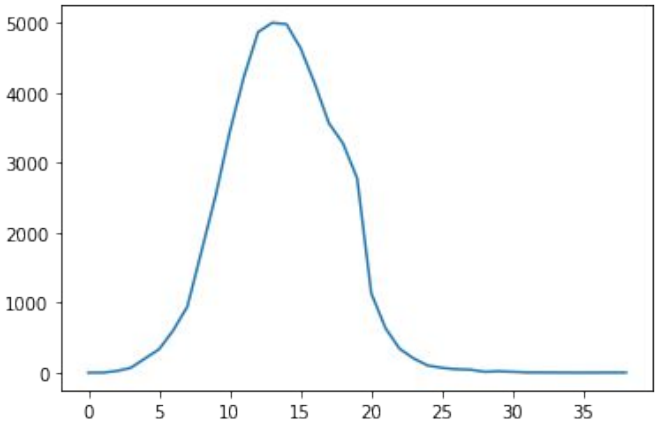
['<sos>', '平板', '該', '買', '哪', '一', '台', '<end>', '<pad>', '<pad>', '<pad>', '<pad>' ... … … …]

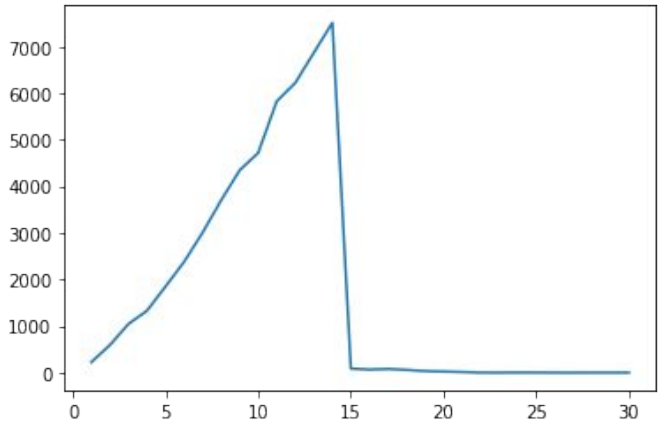




Preprocess - Char-based

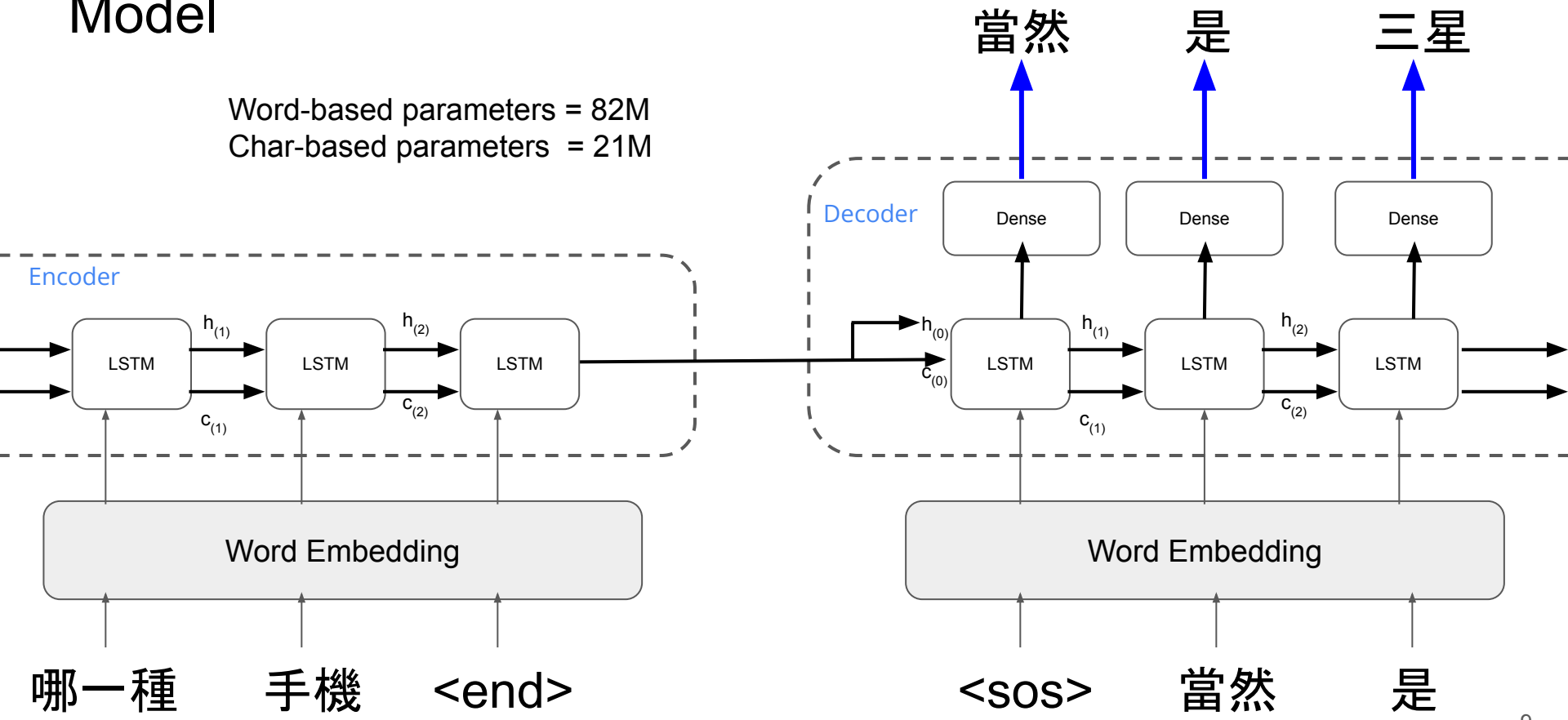
- 為所有字元建立 Tokenizer，總共 5434 個詞
- 分析 Q 和 A 的 token 數量分佈
  - Q 最長都可能到 37，因此補 padding token 到長度 40
  - A 最長都可能到 30，因此補 padding token 到長度 30





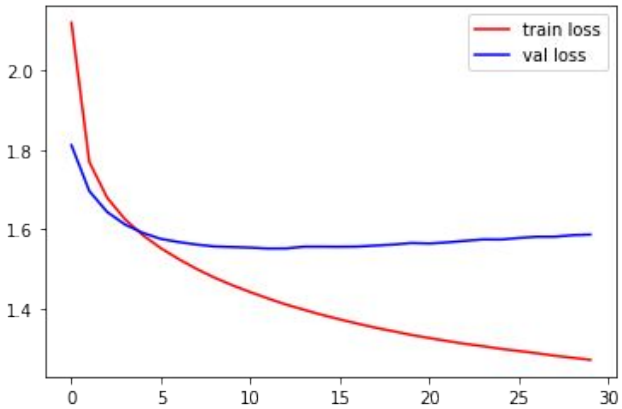
Model

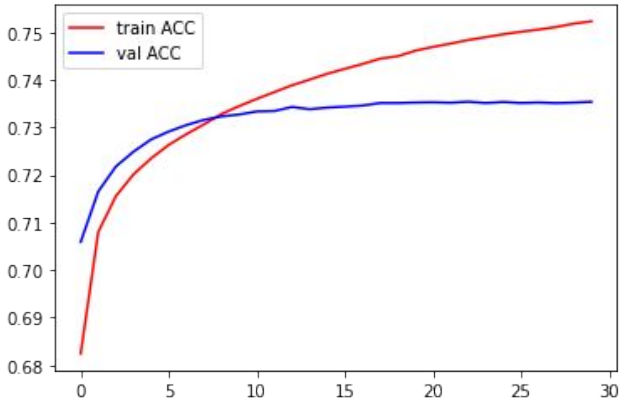
Word-based parameters = 82M
Char-based parameters = 21M



Char-based Training

- Epoch: 30
- Loss function: Cross Entropy
- Optimizer: Adam
- Train:Val = 10000:8900





Word-based Predict

- 確診了體溫只有35度多這樣正常嗎？
  - 現在都低加盟金了
- 死很多的定義是幾個啊？
  - 台灣的人情味是世界前段班
- 清明節大家都有吃潤餅捲嗎？
  - 我都用傳送回家的卷軸

Char-based Predict

- 清心的飲料點全冰會出現多少冰塊？
  - 我都喝冰冰冰冰冰塊冰塊冰塊
- 排隊群聚搶打疫苗？
  - 打疫苗的人都是打疫苗的
- 現在還敢生小孩的人都是怎樣的人
  - 你的人生都是老人的人

Telegram Bot

- 使用 python-telegram-bot 開發，開發前要先跟 Telegram 的 BotFather 申請 Token
- 執行在 i5-8265U、MX150 顯示卡的筆電上，需要 1~2 秒的預測時間

Example 1 - 暨大


暨南大學在哪裡 9:56 PM
台南的學校 9:56 PM
暨南大學是怎樣的一間學校 9:57 PM
學校學校都是中國人 9:57 PM

Example 2 - 投資


股票跌了怎麼辦 9:48 PM
我知道你還有錢 9:48 PM
台積電現在多少錢 9:48 PM
300萬 9:48 PM
應該繼續買嗎？ 9:48 PM
你買不到 9:48 PM
該買虛擬貨幣嗎？ 9:48 PM
買高端 9:48 PM
該買比特幣嗎 9:49 PM
你要買快篩的 9:49 PM

Example 3 - 食物


晚餐要吃什麼 10:13 PM
樓下都吃什麼 10:13 PM
端午節要吃什麼 10:13 PM
麥當勞吃到飽 10:13 PM
北部粽跟南部粽選一個 10:13 PM
南部粽 南部粽 10:13 PM

Example 4 - 食物


KFC不賣蛋塔怎麼辦 10:14 PM
你的錢買不到 10:14 PM
去肯德基該買什麼 10:14 PM
買麥當勞啊 10:14 PM
你喜歡麥當勞還是肯德基 10:14 PM
麥當勞 我都吃麥當勞 10:14 PM

Example 5 - 防疫


誰買得到快篩 10:17 PM
五樓都去228公園買到 10:17 PM
你有買保險嗎 10:17 PM
保險公司倒了 10:17 PM
不打疫苗的人在想什麼 10:17 PM
打疫苗死的比較多 10:17 PM
疫情失守怎麼辦？ 10:17 PM
自主應變 自主應變 10:17 PM
PCR 跟快篩哪個強 10:17 PM
pcr 10:17 PM

Example 6 - 防疫


確診了怎麼辦 10:19 PM
不篩就沒確診 10:19 PM
出事了誰負責 10:19 PM
負責 我負責了 我負責了 10:19 PM
連假你敢出去玩嗎 10:19 PM
玩啊 哪次不玩 10:19 PM

Conclusion

- Char-based 成果比 Word-based 好推測是斷詞不夠好的關係
  - 有許多詞會被錯誤的切割開
  - 尤其在新詞彙、特殊詞彙比較多的 PTT
- Improvement
  - 使用 CKIP 的中文 Pretrain Word Embedding
  - Attention
  - BERT