F74056247 曾大瑋

# 多媒體系統與應用 HW3 – Chatbot

程式部份:
程式為 Formal Chatbot.ipynb

**需要的 library**

```
import jieba
from gensim.models import word2vec
import numpy as np
import math
```

**斷詞**

```
output = open('./stopword.txt', 'w', encoding='utf-8')
with open('./Dataset.txt', 'r', encoding='utf-8') as content :
    for texts_num, line in enumerate(content):
        line = line.strip('\n')                    # 去除換行符號
        words = jieba.cut(line, cut_all=False)     # 用 jieba 斷詞
        for word in words:                         # 如果斷詞的字是 stopwords 將它去除
            output.write(word + ' ')
        output.write('\n')
        if (texts_num + 1) % 10000 == 0:           # 每 10000 行顯示進度
            print("已完成前 %d 行的斷詞" % (texts_num + 1))
```

**Training**

```
sentences = word2vec.LineSentence('./stopword.txt')
model = word2vec.Word2Vec(sentences,size = 400, window = 20, workers = 3, sg =
1, min_count=0, iter=300)
model.save('stupid.bin')
```

**QA**

```
model = "stupid.bin" # 載入模組
model_w2v = word2vec.Word2Vec.load(model)

# 設定 output
outputfile = open('F74056247.csv', 'w+', encoding='utf-8')
```

```
# 讀入題目
with open("PPT_test_corpus.txt", encoding='utf-8')as inputline:
    for line in inputline:
        line = line.strip('\n')
        output = line.split("\t", 1)
        text = output[0]
        answer = output[1].split("\t")


        words = list(jieba.cut(text.strip()))
        word = []   # 當前的題目儲存在這裡
        for w in words:     # 去除 stopword
            if w in model_w2v.wv.vocab:
                word.append(w)

        eachans = []    #每題的四個選項儲存在這裡
        # 以 jieba 切割每個選項，並去除 stopword 之後再儲存成 list
        for everyans in answer:
            answercut = []
            temp1 = "".join(everyans.split(')')[1])
            answercuts = jieba.cut(temp1, cut_all=False)
            for checkvocab in answercuts:
                if checkvocab in model_w2v.wv.vocab:
                    answercut.append(checkvocab)
            eachans.append(list(answercut))

        score = []
        score.append(model_w2v.wv.n_similarity(word, eachans[0]))
        score.append(model_w2v.wv.n_similarity(word, eachans[1]))
        score.append(model_w2v.wv.n_similarity(word, eachans[2]))
        score.append(model_w2v.wv.n_similarity(word, eachans[3]))
        choose = np.argmax(score) + 1

        outputfile.write('[' + str(choose) + ']\n')

outputfile.close()
```

**Check 正確率**

```
# get incorrect line count
incorrect = ! diff -y --suppress-common-lines ./F74056247.csv
./correct_answer_file.txt | grep '^' | wc -l
incorrect = int(incorrect[0])

print(incorrect)
```

```
# calculate rate
print( str((500-incorrect)/500*100) + '%')
```