

# 專題報告

9/4

b1228006張鴻毅  
b1228022梁釗豪  
b1228029蘇琪文  
b1228032張紋菱



NURSING RECORD COMPLETION



GPT2  
MODEL



"GPT2, your best choice!"

- MIMIC-III

# GPT2 護理記錄補全

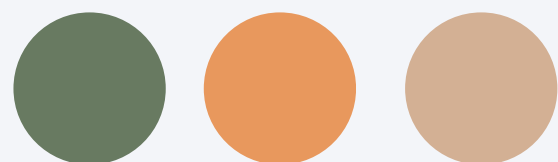
這個專案使用 Hugging Face 的 GPT-2 模型

訓練於 MIMIC-III 的 Nursing Notes

並提供一個互動式補全介面

讓使用者逐字選擇生成內容

探索模型的語言能力與醫療語境掌握



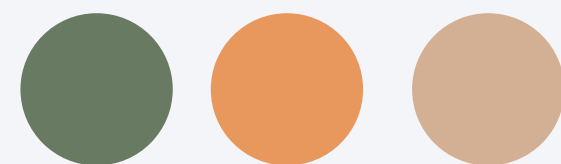
# 資料集介紹

## MIMIC - III

MIMIC - III

→ Category

→ Nursing(223556筆)



**下載方式** (gdown)

pip install gdown

gdown 1QFIWLsFP6\_MzCNe8euuupAK6hFxaueXl

CATEGORY

Nursing/other 822497

Radiology 522279

**Nursing 223556**

ECG 209051

Physician 141624

Discharge summary 59652

Echo 45794

Respiratory 31739

Nutrition 9418

General 8301

Rehab Services 5431

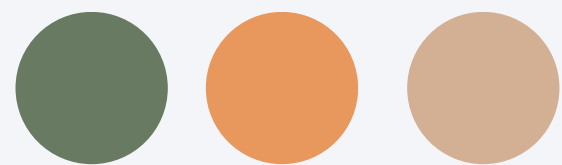
Social Work 2670

Case Management 967

Pharmacy 103

Consult 98

Name: count, dtype: int64



# 不同版本比較

[Github link](#)



**gpt2\_5%.py**



**gpt2\_5000.py**



**gpt2\_all.py**



**gpt2\_ok.py**



**gpt2\_ok\_all.py**

# 主要功能對比

項目	gpt2_all.py	gpt2_5000.py	gpt2_5%.py	gpt2_ok.py	gpt2_ok_all.py
資料規模	全部資料	前1000筆文本， 最多5000句對	5%抽樣	全部資料	
訓練方式	完整fine-tuning	僅載入預訓練模型	完整fine-tuning		
輸出方式	pipeline自動生成	互動式逐詞選擇生成5個建議&機率			互動式選擇生成 5句建議&機率

# 資料處理對比

項目	gpt2_5%.py	gpt2_5000.py	gpt2_all.py	gpt2_ok.py
篩選條件	Nursing類別	Nursing類別	Nursing類別	Nursing類別
預處理	移除標點符號	保留基本標點，小寫轉換	簡單換行處理	去除多餘空白、換行
資料分割	train/test split (9:1)	無分割	無分割	無分割
儲存格式	Dataset物件	記憶體中的句子對	文字檔	文字檔

# 模型訓練對比

項目	gpt2_5%.py	gpt2_5000.py	gpt2_all.py	gpt2_ok.py
基礎模型	gpt2	gpt2	gpt2	gpt2
訓練策略	標準訓練	用 LSTM 模型做淺層的下一詞預測	標準訓練	標準訓練
Epochs	最多10 (early stop)	10	1 (可調)	2
Batch Size	2	N/A	2	4
評估機制	驗證集loss監控	驗證集loss監控	無	驗證集loss監控
模型保存	新增資料夾儲存	新增資料夾儲存	新增資料夾儲存	新增資料夾儲存



# 文字生成對比

項目	gpt2_all.py	gpt2_5000.py	gpt2_5%.py	gpt2_ok.py	gpt2_ok_all.py
生成方式	Pipeline自動生成	手動逐詞選擇			選擇句子建議
互動性	無	高度互動			
選項數量	單一生成結果Top	Top-5詞彙選擇			5句子建議
用戶控制	僅能改prompt	可重新生成、停止			
輸出長度	最大100 tokens	用戶決定			input+20
輸出結果	-	在專有名詞斷句 預測範圍較多	只要不是連接詞等等能預測 幾個字與樣本一樣		第一個字跟樣本一樣 其他不同

### 資料處理策略差異：

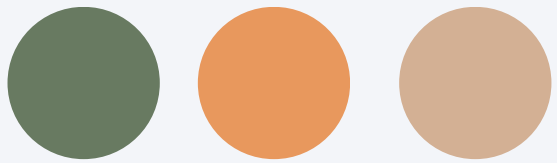
- gpt2\_5%.py: 標點符號清理，取5%樣本
- gpt2\_5000.py: 精細的句子分割和配對，保留語義結構
- gpt2\_all.py: 分批處理大數據，記憶體友善
- gpt2\_ok.py: 標準清理流程，適合一般使用

### 模型訓練方法差異：

- gpt2\_5%.py: 防過擬合的驗證導向訓練
- gpt2\_5000.py: 創新的LSTM預訓練+GPT2載入架構
- gpt2\_all.py: 生產級的標準訓練流程
- gpt2\_ok.py: 簡化的研究訓練流程

### 文字補齊方法差異：

- 逐詞補齊: gpt2\_ok.py, gpt2\_5%.py, gpt2\_5000.py - 精細控制(?)
- 句子補齊: gpt2\_ok\_all.py - 快速生成，適合演示
- 自動生成: gpt2\_all.py - 效率導向



Thank You

