拯救托福聊天機器人

Rescue your TOEFL

組員：李振維、黃大瑋、藍璟誠

01

專題摘要

02

專題簡介

03

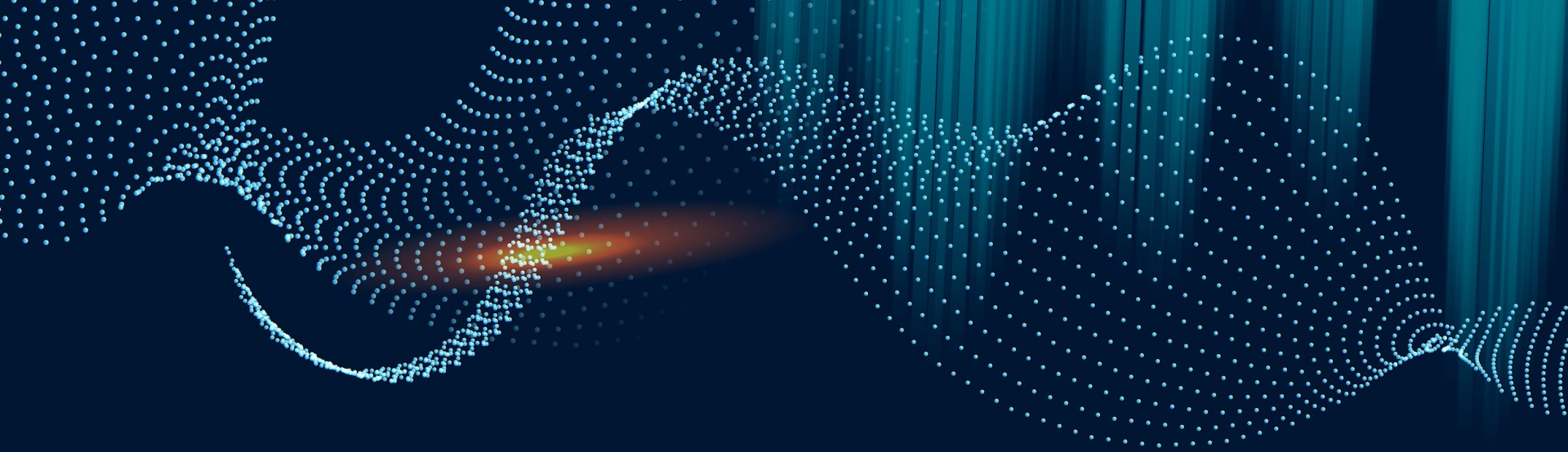
方案架構
實作說明

04

Demo

05

方案評估
未來發展

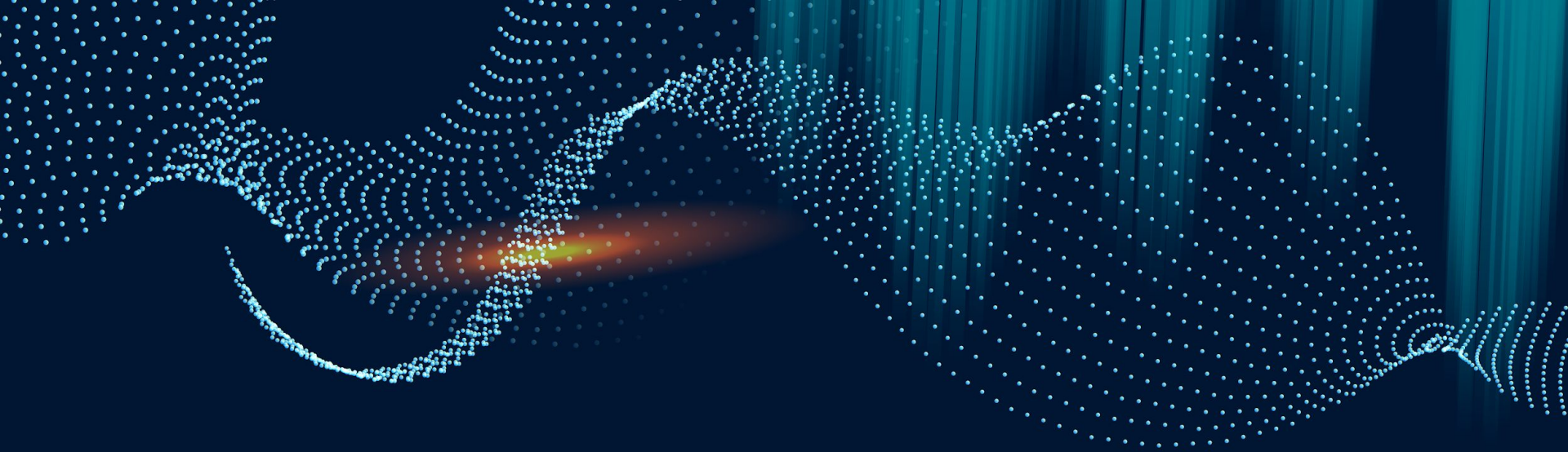


01 專題摘要

Project Digest

專題題目＆成果

“ 目的希望為托福考生建立一個能夠快速找到適合自己的學習方法，同時讓考生每日都能建立自己的學習日記，紀錄自己的成長，作為學習之追蹤 ”



02 專題簡介

Project Introduction

專題動機

資料來源：國際及兩岸教育司- 110年度世界各主要國家之我留學生人數統計（更新日期2021/12/30）

60,307

總留學人數

40%

美、加留學人口佔總留學人數之比例

23,563

美、加留學人數

專題動機

疫情趨緩，需求上升

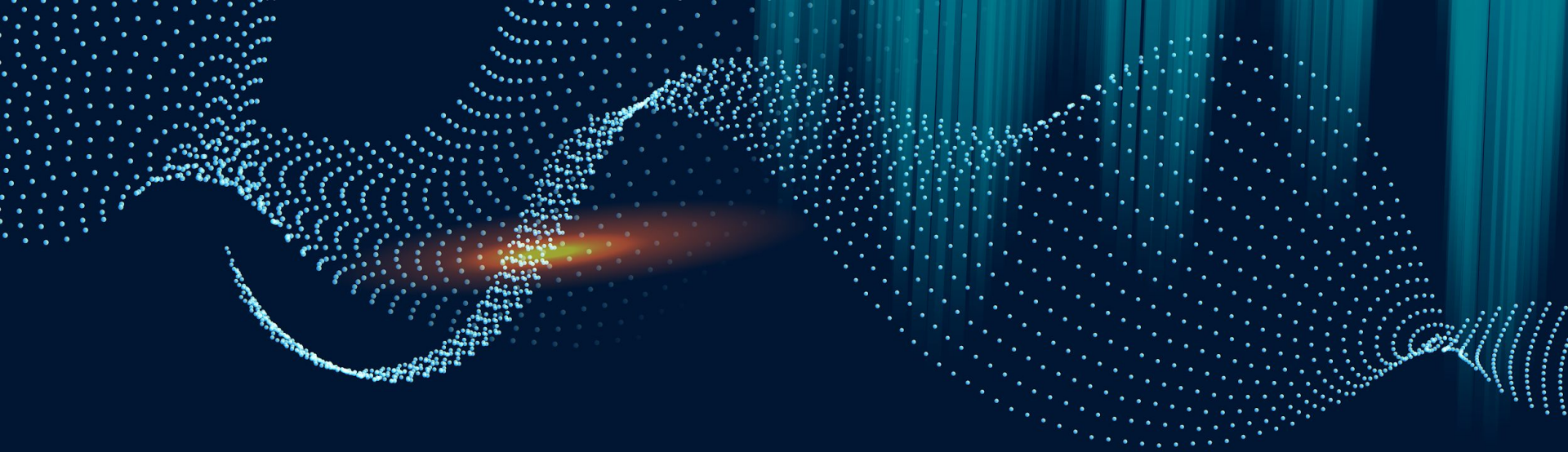
前幾年由於疫情的影響，需求銳減，但近期疫情趨緩後，相信出國留學的人會開始回流，準備考試的人也會相對上升

深受其擾

由於個人需要準備托福考試，但心得文百百種，也不一定符合自身的需求，而且閱讀起來不夠方便有效率，

專題目的

- **提供考生閱讀、聽力、口說、寫作之準備方法摘要**
  - 快速找到適合的學習方法，在茫茫心得海中，幫你做重點摘要
- **提供考生記錄各科學習歷程的功能**
  - 追蹤自己的進度，並且可以回顧過去的學習狀況，找出自己的弱點，並加以改進



03 方案架構

Solution Structure

方案架構圖


資料集生成
資料前處理
心得摘要
學習歷程
Dcard 爬蟲
PTT 爬蟲
CSV 檔案
心得分割
考生背景分割
目標分數分割
特殊符號去除
摘要生成
條件判斷
問答設計
資料輸入 / 輸出
分類模型訓練
問答設計
建立資料庫
各科分佈

資料集生成&資料前處理

Data collection & Preprocessing

Dcard 爬蟲

Selenium

由於 Dcard API 無法使用，只好改用傳統的 Selenium 進行網站行為模擬

於網址設定關鍵字：托福、由新到舊

層層抽取資訊

1. Article（貼文）
2. Title（標題）
3. Content（內文）

批次讀取，每次會紀錄最後一筆資料之唯一值，下次讀取從上次的最後一筆開始讀取

條件判斷 & 寫入 CSV

- 心得 in title
- 一戰 in title
- 二戰 in title
- 分享 in title
- 雅思 not in title

每次會將讀取的的資訊，依序匯入 CSV 檔案中保存

PTT 爬蟲

Beautiful Soup

使用bs4模擬使用者點擊網站內的超連結，進而閱讀文章

觀察網址特徵，得到文章及上一頁的連結

層層抽取資訊

1. 作者、看版、標題、日期
2. 內文

- 使用find_all獲取html中的欄位內容
- 使用split獲取正確的內文陣列位置

條件判斷 & 寫入 CSV

- 心得 in title

PTT文章標題具有一致性，可利用中括號內的文字判斷文章性質

爬蟲&前處理範例

原架構

.....
考生背景:XXXX
目標分數:XXXX
.....
閱讀:
XXXXX
聽力:
XXXXX
口說:
XXXXX
寫作:
XXXXX
.....

background	學測 14 級、指考 90 分、多益 900
targetScore	R:28 L27 S:25 W:29 T109
rContent	閱讀最重要的除了速度和單字(這兩點除了勤練之外也沒什麼撇步)，邏輯也很重要……
lContent	聽力我本身比較強，並沒有聽不懂的問題，但我覺得老師上課教的筆記很重要……
sContent	我的口說也算弱，因為本身口才就不好。所以上課的模板對我來說非常有用……
wContent	寫作一直都是我的罩門，因為我不是屬於很懂的考官要什麼的那種學生……

心得摘要

Preparation Digest

心得摘要

套件介紹

- Spacy
- Neuralcoref - 用來做共指代消(避免代詞的權重下降)
- Transformer - 加載 custom 語言模型(bert-base-multilingual-cased)
  - autoConfig:儲存model或tokenizer的超引數
  - autoTokenizer:將原始的語料編碼成適配模型的輸入
  - autoModel:載入模型
- Bert-extractive-summarizer - Summarizer model

Spacy & Neuralcoref

- **共指代消**：指代消解（Coreference Resolution）一般被應用於處理資訊檢索中的前處理部份，主要是找回原先被替換過的字詞，為了避免重要的字詞因指代的因素而造成權重計算降低的問題，若沒有加以處理，在權重計算上會產生因為北極熊此字詞出現次數過於稀少，而導致資訊檢索系統誤判，因此，透過指代消解的處理，可以將被替換過的字詞還原成原有的意思，以提高權重計算的次數，增加檢索的正確性。

Ex：北極熊又稱白熊，是在北極裡生長的熊，牠是陸上最龐大的肉食動物。在牠生存的空間裡，牠是食物鏈最頂層。牠擁有極厚的脂肪及毛髮來保暖，其白色的外表在雪白的雪地上是良好的保護色，而且牠可以在陸上及海上捕捉食物，因此牠能在北極這種極嚴酷的氣候裡生存。

bert-extractive-summarizer

來源：https://github.com/dmmiller612/bert-extractive-summarizer

bert-extractive-summarizer 是一個使用 Bert 加上 Clustering 進行抽取式摘要的模型

- **Clustering**：聚類演算法，不斷迭代找出相似的句子聚集在一起
- **Summarizer**：藉由在文章裡找尋重要的句子，並把找出來的句子接在一起，當作這篇文章的summary。在模型訓練上，則是當做sentence classification 的任務，藉由分類每個句子是否為重要句子，來得到要做為summary 的句子。

Example（摘要前）

口說三次分別是20/23/26，口說我覺得在缺乏環境的狀況下，是最難練的。最後一次能考到 26也是我完全沒想到的，因為26達到了所有我想申請學校的門檻，考之前也很沒信心，不覺得自己有比第二次進步多少。仔細回想準備口 ．說的過程，在一戰前，我非常晚才開始認真準備，前期一直不想面對，因為口 說沒有對象、環境真的很難練，開口就很容易結巴。而且一開始常常是腦袋想中文，要再經過腦中翻譯這個過程，講的語速很慢。錄音起來的 內容很多文法錯誤、結巴，很容易越練越不想練，所以一戰只有 20。二戰前兩週報名了克雷英文，我覺得這個課程對我來 說的幫助，在於每堂課都有對應的作業（三題TPO），而且作業是要錄到最完美的版本。可能我在這方面有點強迫症，為了錄到最好交出去，我花了很大量時間反覆說、一題會錄好幾十次，直到不卡、在時間 內剛好講完，才交出去。然後會有專業老師幫改，老師改的速度滿快的，會錄影一邊播音檔，告訴你哪邊發音、用字可以再更好。除了錄作業那三題，我還會自己額外練很多題 TPO上傳，一樣用交作業的標準去錄到最完美的版本。雖然二戰只有 23，但我覺得這對我三戰 26打下了一些基礎。三戰前，我回想最大的進步是，我在生活中更多地跟自己 說英文，走路時、洗澡時，真的發聲講出來，而不是講在心裡。一開始聽到自己講英文，因為很不習慣，再加上發音、文法、講的 內容都有一定的不完美，其實覺得滿恥的。但就真的要盡量講，讓自己習慣講英文，越講會越快，腦中翻譯的過程會慢慢消失，到後面我自己也沒有意識到，我可以省略中翻英，直接用英文表達出我的想法。另外考試過程一定要放鬆，緊張也會影響發揮，三戰第一題其實我自認講得很差，可能太緊張一直鬼打牆，舉例也沒有很具體、完整。但第二三四題，自己有比較放鬆，該講的點都有在時間 內講完，也滿流暢的，就得到 26分了。最後，推薦兩個幫助練口說的app。口說助手，完全針對托福考試設計的，有計時 +錄音的功能，頁面簡潔好操作，平常練習就讓自己習慣看著秒數倒數，知道剩下幾秒至少要講到哪個部分，對於時間控管也有幫助。 SK2 TOEFL DAVID托福口說，裡面有各種分類的題目，很適合拿來練第一題，有老師錄好的答案，可以多聽學習老師怎麼舉例。

Example（摘要後）

最後一次能考到26也是我完全沒想到的，因為 26達到了所有我想申請學校的門檻，考之前也很沒信心，不覺得自己有比第二次進步多少。仔細回想準備口 說的過程，在一戰前，我非常晚才開始認真準備，前期一直不想面對，因為口 說沒有對象、環境真的很難練，開口就很容易結巴。 二戰前兩週報名了克雷英文，我覺得這個課程對我來 說的幫助，在於每堂課都有對應的作業（三題 TPO），而且作業是要錄到最完美的版本。 然後會有專業老師幫改，老師改的速度滿快的，會錄影一邊播音檔，告訴你哪邊發音、用字可以再更好。 一開始聽到自己講英文，因為很不習慣，再加上發音、文法、講的 內容都有一定的不完美，其實覺得滿恥的。 但就真的要盡量講，讓自己習慣講英文，越講會越快，腦中翻譯的過程會慢慢消失，到後面我自己也沒有意識到，我可以省略中翻英，直接用英文表達出我的想法。 但第二三四題，自己有比較放鬆，該講的點都有在時間內講完，也滿流暢的，就得到 26分了。

Example（csv 架構）

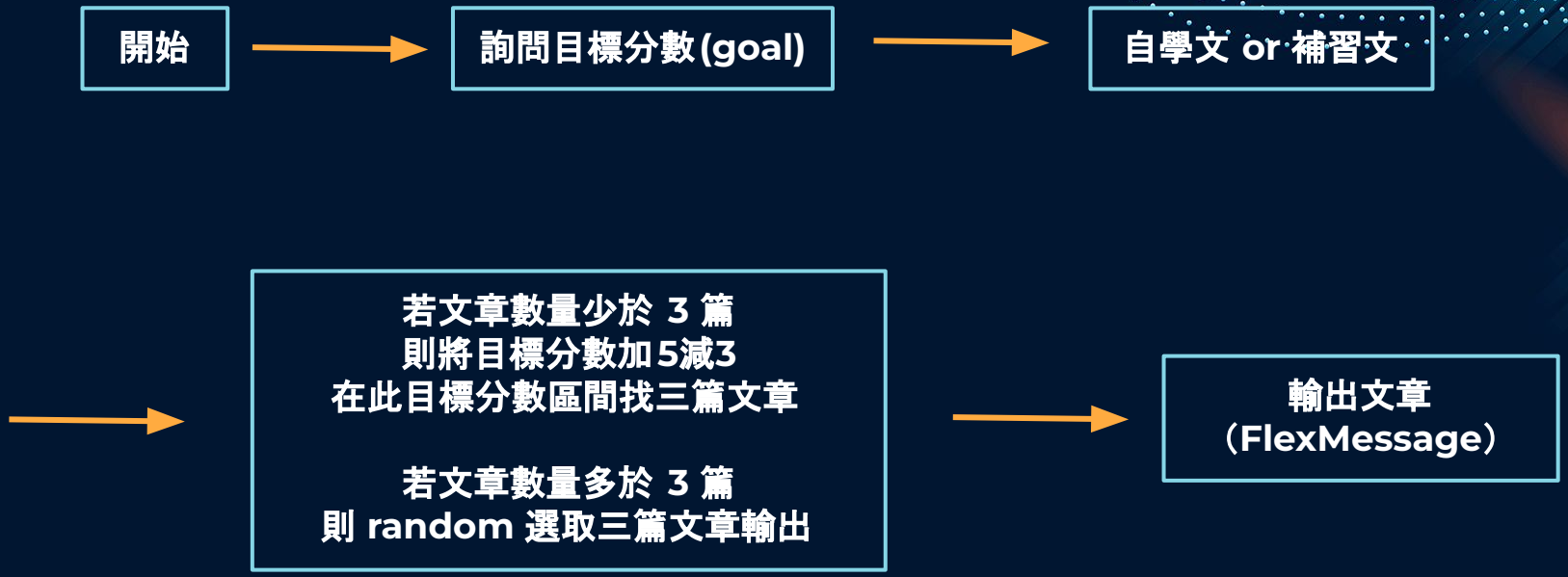
原架構

rContent	閱讀最重要的除了速度和單字（這兩點除了勤練之外也沒什麼撇步），邏輯也很重要……
lContent	聽力我本身比較強，並沒有聽不懂的問題，但我覺得老師上課教的筆記很重要……
sContent	我的口說也算弱，因為本身口才就不好。所以上課的模板對我來說非常有用……
wContent	寫作一直都是我的罩門，因為我不是屬於很懂的考官要什麼的那種學生……

新增各科摘要欄位

rSummary	若超過 50 個字，則摘要成一半
lSummary	若超過 50 個字，則摘要成一半
sSummary	若超過 50 個字，則摘要成一半
wSummary	若超過 50 個字，則摘要成一半

Example（LineBot）



學習歷程

Learning Journey

學習歷程

套件介紹

- Jieba - 斷詞、關鍵字提取
- Sklearn - 分類模型訓練（多項式分類）
- SQLite3 - 資料庫存取

學習歷程

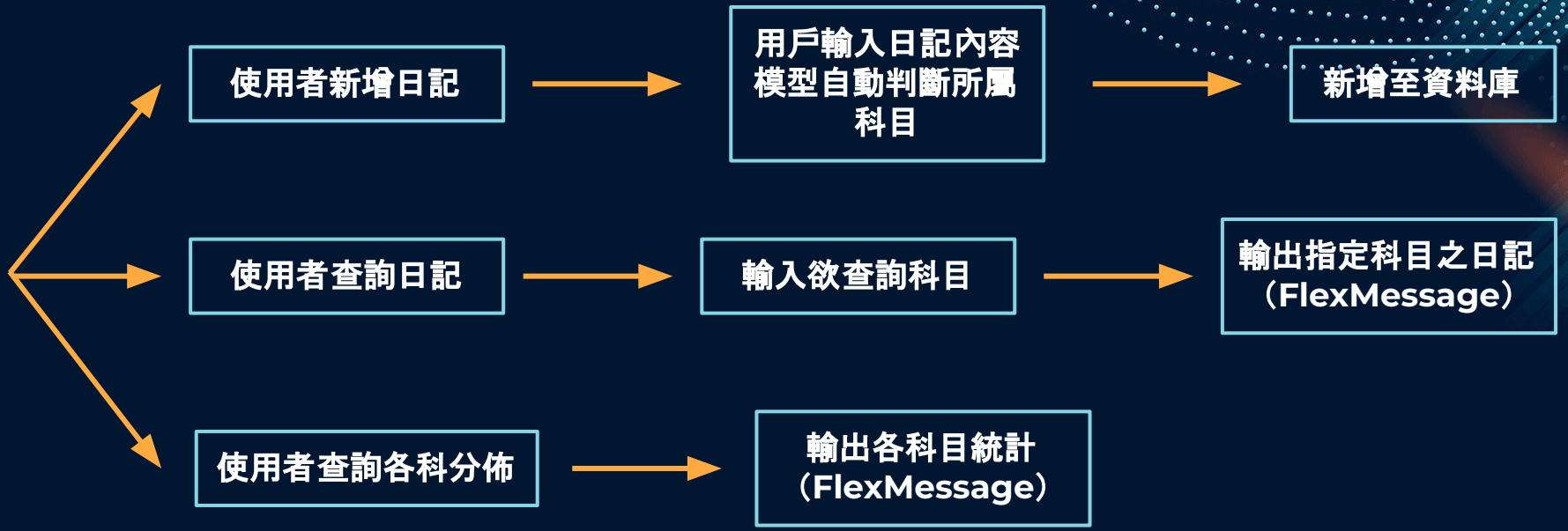
- 將原始各科心得文（rContent、lContent、sContent、wContent）作為分類模型之訓練資料集，同時將每篇心得賦予各自的類別（閱讀、聽力、口說、寫作）
- 利用 jieba 進行精確模式斷詞、後利用 extract.analyze 提取關鍵詞（選擇 20 個）
- 將心得用 CountVectorizer 向量化，類別用 labelencoder 編碼，丟入貝氏分類模型中進行訓練（**accuracy：0.92**）
- 讀取使用者輸入，依照使用者回覆進行指示
- 若為新增：則將日記藉由模型自動判讀所屬科目，儲存至資料庫
- 若為查詢：則將對應科目之心得文全部搜尋出
- 若為各科分佈：輸出截至目前各科所有的心得總數與佔比

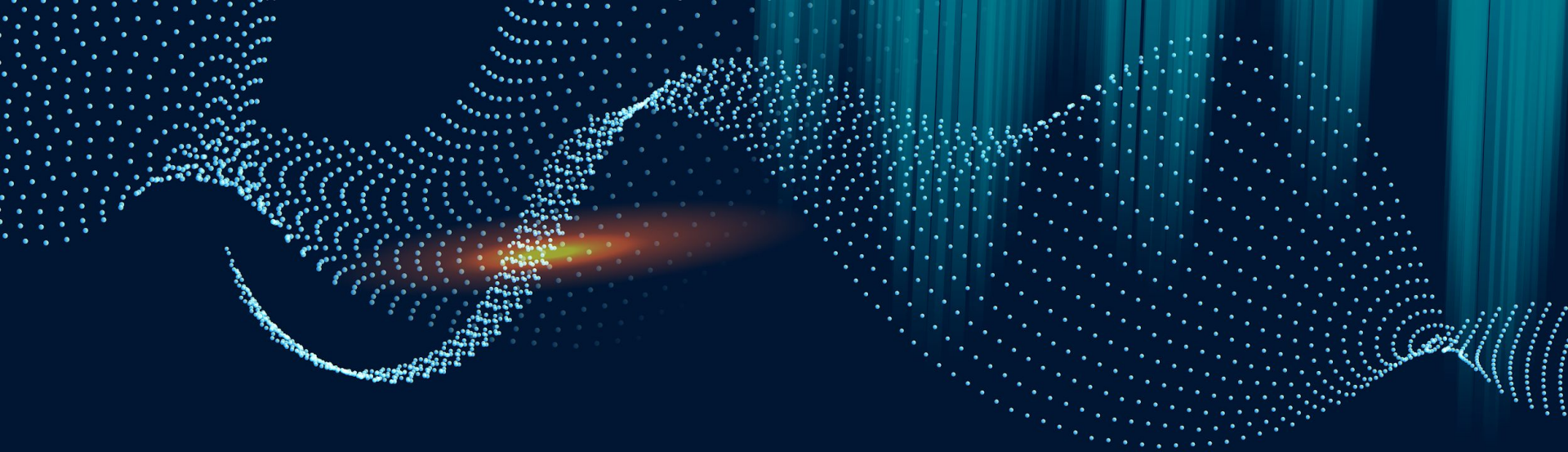
Example（模型訓練）

rContent	lContent	sContent	wContent
閱讀最重要的除了速度和單字（這兩點除了勤練之……	聽力我本身比較強，並沒有聽不懂的問題……	我的口說也算弱，因為本身口才就不好。所以上課……	寫作一直都是我的罩門，因為我不是屬於很懂的……

XXXX	**聽力**
XXXX	**閱讀**
XXXX	**口說**
XXXX	**寫作**
XXXX	**閱讀**

Example（LineBot）





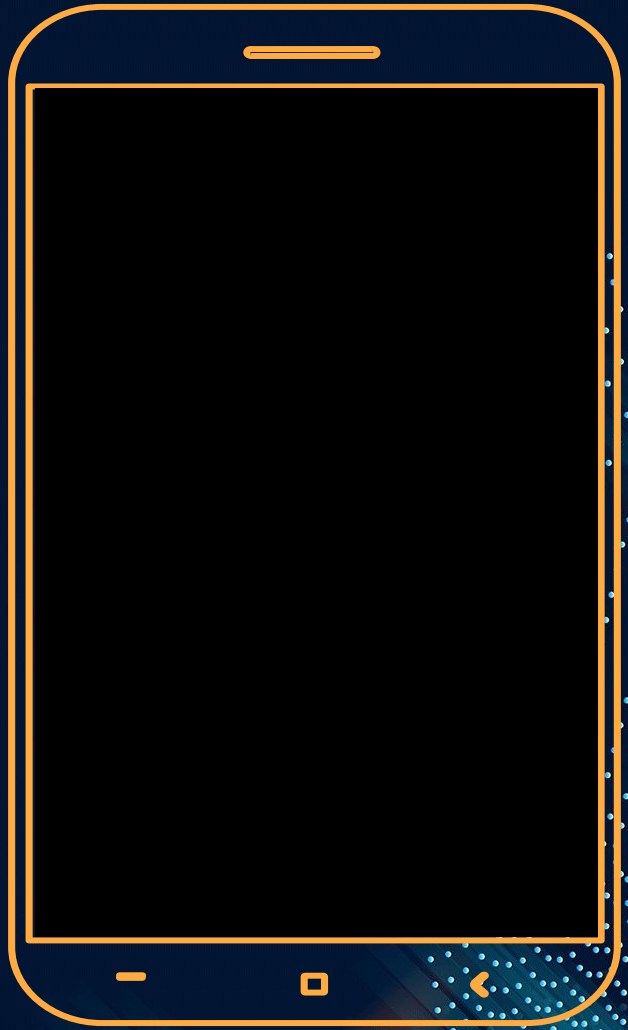
04 實作演示

Demo Time

TOEFL Bot Demo

心得摘要:

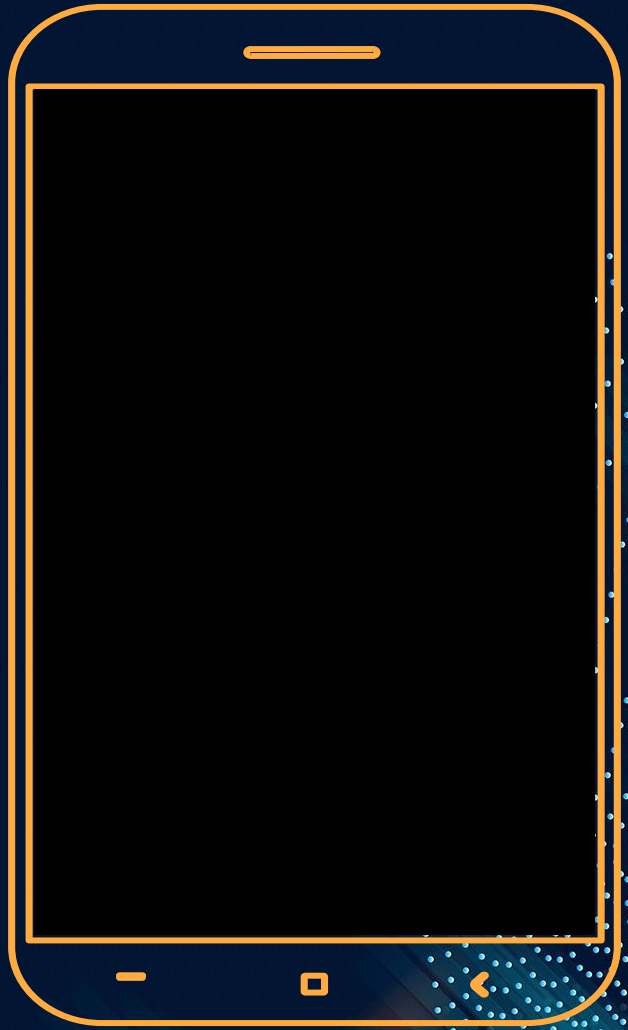
1. 詢問目標分數(示範例外)
2. 自學文 or 補習文
3. 顯示符合條件之隨機 3 篇

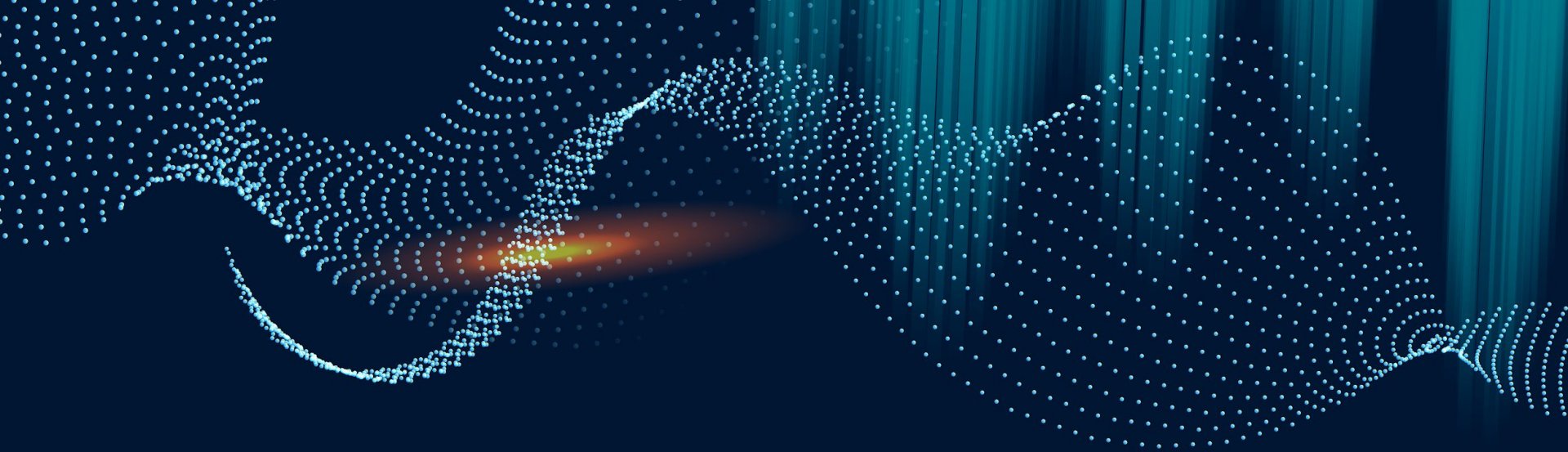


TOEFL Bot Demo

學習歷程:

1. 新增日記(直接輸入文章)
2. 查詢日記(依照科目查詢)
3. 各科分佈(統計報告)





05 方案評估與未來發展

Solution evaluation & Future Plan

方案評估與未來發展

	提升準備效率	回顧自我進度	人機互動
方案評估	快速提供 符合需求之摘要(30 秒)	隨手記錄自身學習歷程 (一機在手，隨時隨地)	讓機器人陪伴你 度過枯燥的準備考試過程
未來發展	替換 **Model** 可改用 SbertSummarizer (三重損失&攣生神經網路)	動態擴充資料集 確保資料同步更新 廣泛收納自學文	擴充功能 每日小考題 筆記本功能 人性化的回覆 加入 Setiment&Importance analysis

Thank you

讓拯救托福機器人成為你海外留學的好幫手