2020 Libgirl Corpus Intern

工作進度報告

報告人：李郁萱

時間：2020年05月20日

LIBGIRL

CONTENTS

1 專案概覽

2 工作歷程分享

3 下階段工作計畫

4 實習心得

專案概覽

1 專案概覽

聊天機器人

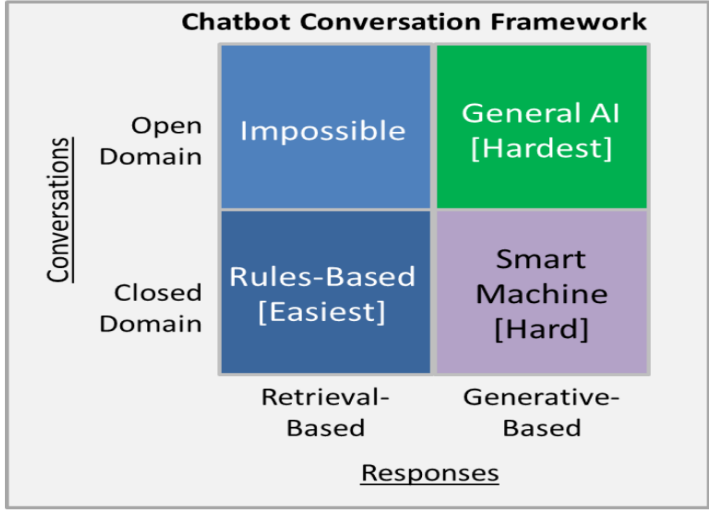
專案方向擬定、技術指導

PCA、評分標準、BERT模型

歧異性分析、語料預處理

郁萱

Domain Knowledge、語料預處理



LIBGIRL

1 專案概覽

專案共同討論

訓練語料選擇與蒐集
英文: EmotionLines, 英文笑話語料集
中文: SRDA 問卷問句集
其他情緒相關語料集

Domain Knowledge
聊天機器人 Domain

語料預處理
JSON檔案讀取
爬蟲
中文Jieba斷詞
Multi-label

知識研究
Fine Tune
問卷評量
K-means演算法

萌太子測試與反饋

LIBGIRL

工作歷程分享

2 工作歷程分享

完成進度

第一階段
問卷研究

信度
效度

SRDA
學術調查
研究資料庫

第二階段
問卷資料庫搜尋

第三階段
問卷爬蟲、清洗

PDF 文字爬蟲
正則表達式

Jieba斷詞
生成問卷字典

第四階段
語料處理

第五階段
問句分群貼標

句向量
K-means分群

LIBGIRL

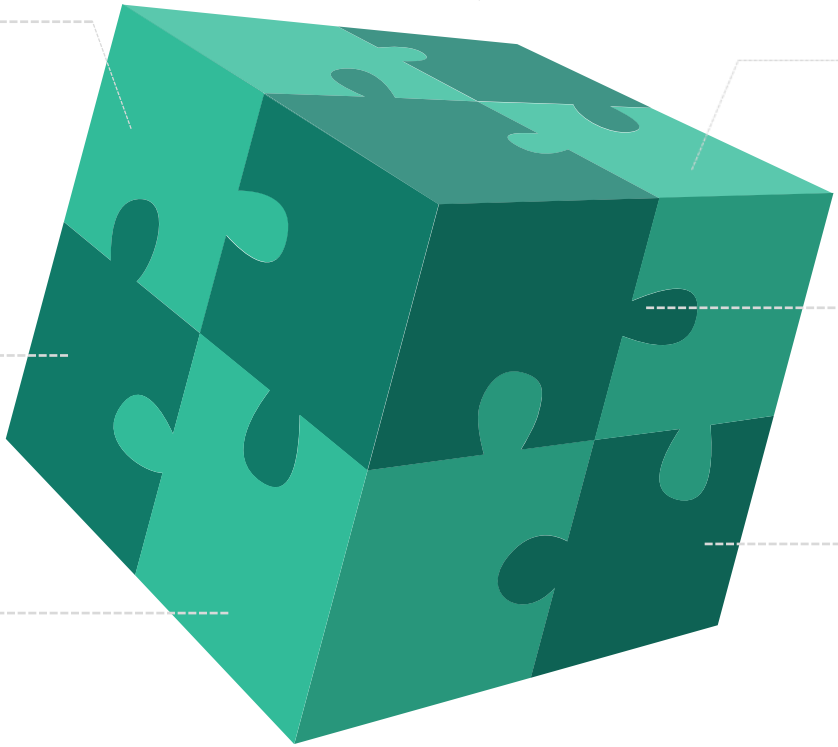
2 工作歷程分享

語料庫介紹

EmotionLines
7種情緒分類、單句

英文笑話語料集
三類笑話、各自有評分分數
一句話或一段對話

SRDA問卷資料庫
各類型問卷、問卷題目
問卷設計說明



仇恨言語識別語料集

Personae語料庫

Twitter上激進分子情緒分析

Death Row

LIBGIRL

2 工作歷程分享

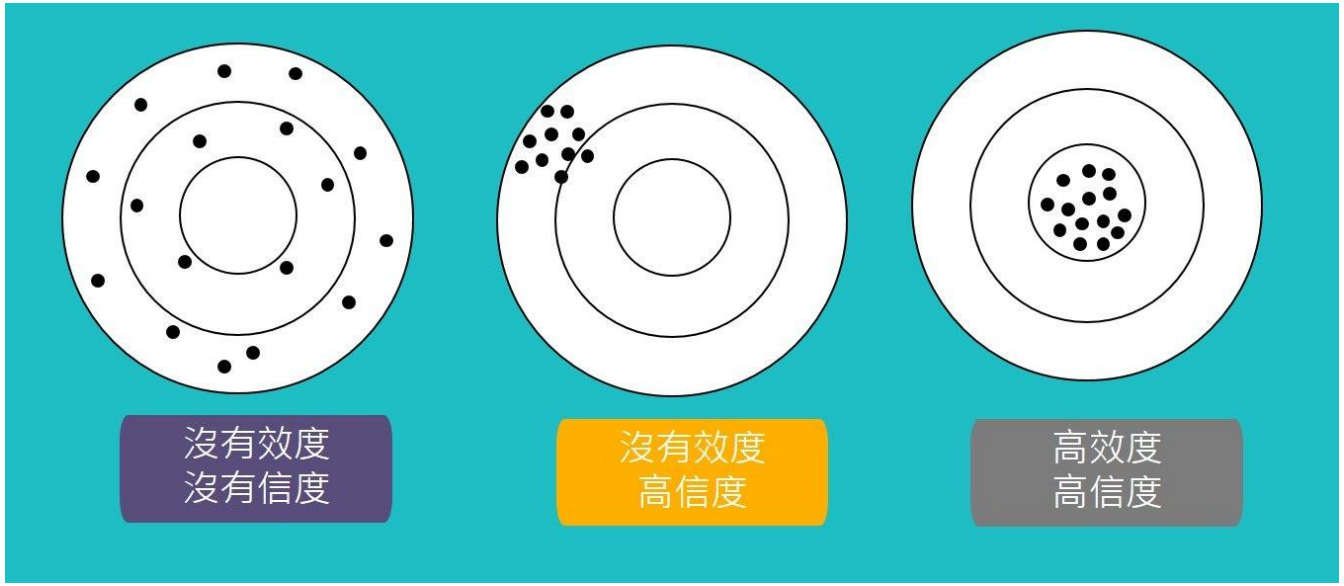
SRDA問卷篩選

目標問卷集

心理
憂鬱
壓力

篩選門檻和結果

符合信效度
共取得20個問卷PDF



圖片來源:永析統計及論文諮詢顧問

LIBGIRL

2 工作歷程分享

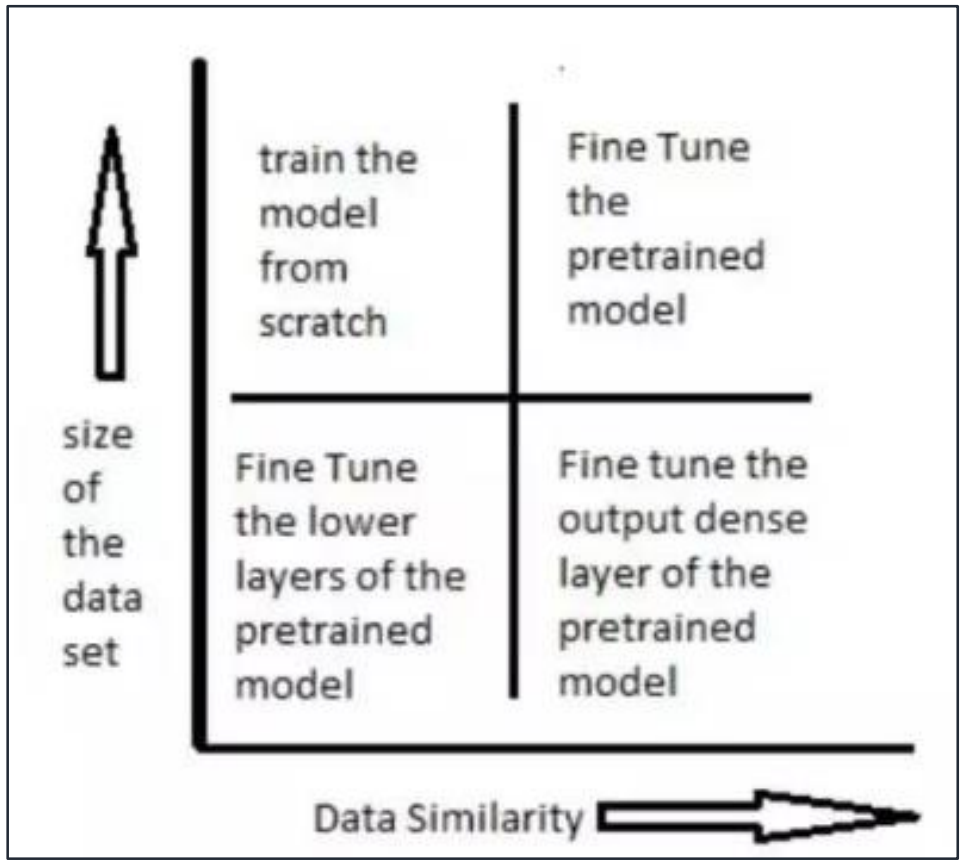
知識研究-Fine Tune

Case1：數據集小，數據相似度高
輸出層修改成需要的結構

Case2：數據集小，數據相似度不高
重新訓練，通過凍結預訓練模型的前k層進行彌補

Case3：數據集大，數據相似度不高
採用預訓練模型不適好方法
預訓練模型權重全都初始化

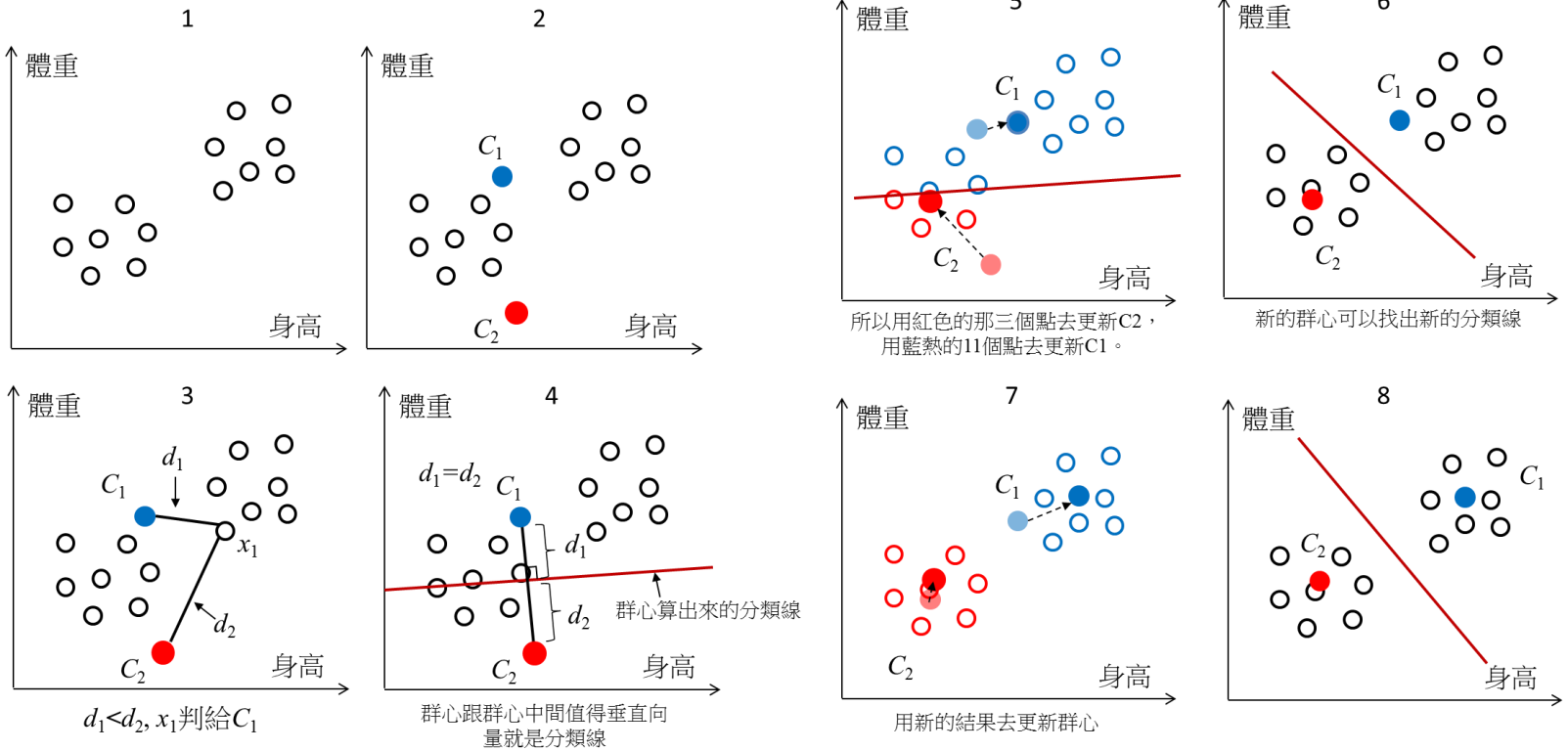
Case4：數據集大，數據相似度高
非常高效率，結構和初始權重不變，在新數據集的基礎上重新訓練



Source: Judy T Raj, Google Cloud Architect

LIBGIRL

2 工作歷程分享 知識研究-K-means演算法



$$\arg\min_{\mu} \sum_{c=1}^{K} \sum_{i=1}^{n_c} \|x_i - \mu_c\|^2 \Big|_{x_i \in S_c}$$

圖片來源:Tommy Huang, Medium

1. 設定分群數量(k值)
隨機給k個群心

2. 各點到所有群心歐式距離
將點判給最近的群心

3. 利用新分類的資料
更新新的群心

4. 重複步驟2~3，直到群心
不再有變動(收斂)

LIBGIRL

2 工作歷程分享

Jieba斷詞

斷詞模式 jieba.cut
精確模式
全模式
搜尋引擎模式
Paddle模式

使用預設詞庫
簡體中文

如何提高斷詞準確性?
使用自定義辭典
改用繁體詞庫
歧異詞辨識 → LSTM 模型做斷詞

其他應用
提取關鍵詞
詞性標註
去停用詞

```
In [61]: jieba_list
Out[61]: ['受訪者',
          '編號',
          '性別',
          '出生',
          '年次',
          '教育',
          '程度',
          '就業',
          '狀態',
          '即將',
          '出生',
          '的',
          '孩子',
          '的',
          '性別',
```

```
In [61]: jieba_list
          '掌控',
          '一切',
          '嗎',
          '你',
          '曾',
          '因個',
          '人無法',
          '控制',
          '的',
          '事情',
```

LIBGIRL

2 工作歷程分享

分群後問卷語料

人工確認
分成多群再次觀察
更改分群方法

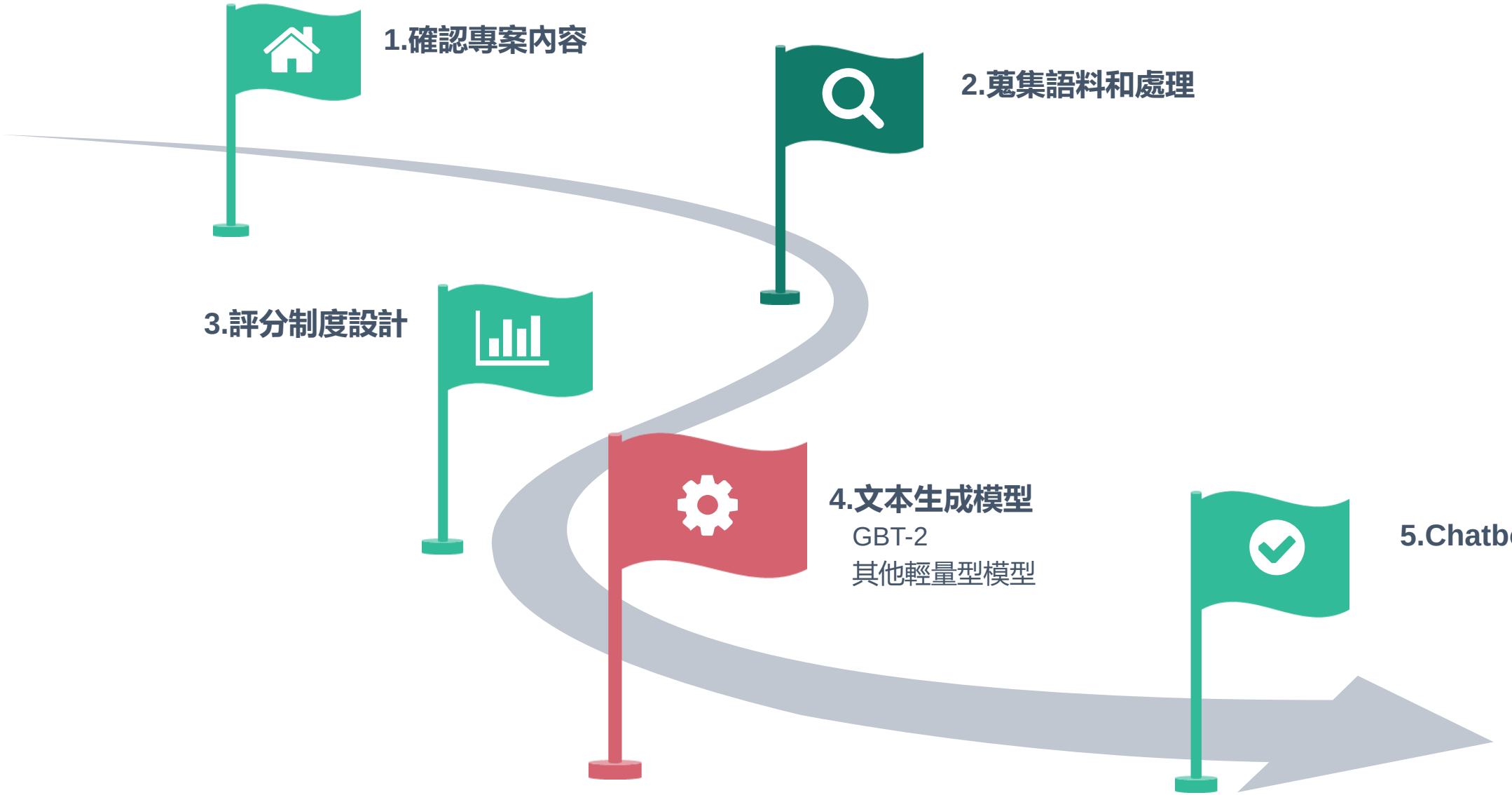
```
for i in range(len(Questions_list)):
    print("問句: {} → 分群 {}".format(Questions_list[i], cluster_labels[i]))
```

```
問句: 受訪者編號 → 分群 1
問句: 性別 → 分群 0
問句: 出生年次 → 分群 0
問句: 教育程度 → 分群 0
問句: 就業狀態 → 分群 0
問句: 即將出生的孩子的性別是 → 分群 1
問句: 即將出生的孩子是您們的第一胎嗎 → 分群 1
問句: 預料之外的事情會讓你煩惱嗎 → 分群 1
問句: 你無法控制生活中的重大事情嗎 → 分群 1
```

LIBGIRL

下階段工作計畫

3 下階段工作計畫



LIBGIRL

3 下階段工作計畫

文本生成模型訓練

GBT-2
其他模型

問句資料集再處理

問句篩選
人工處理
斷詞修正

收集更多語料

提升模型訓練成果

其他任務指派

LIBGIRL

工作心得

4 工作心得

- 開放性討論，共同腦力激盪
- 傑出的團隊領導與實習夥伴
- 持續試誤與快速成長
- 專案規劃清楚明確

LIBGIRL

致謝

感謝所有工作夥伴協助

Thanks for Listening

LIBGIRL