TibaMe A.I. 應用工程師青年班第一期
自然語言專題：Blog樂

組員名單：
袁學平、劉康彥、孫韶甫
林靖凱、陳豐王、丁之琳

組長：**吳秉軒**

TibaMe

報告架構

01 專題介紹

02 實作：分類

03 實作：推薦

04 實作：分析與摘要

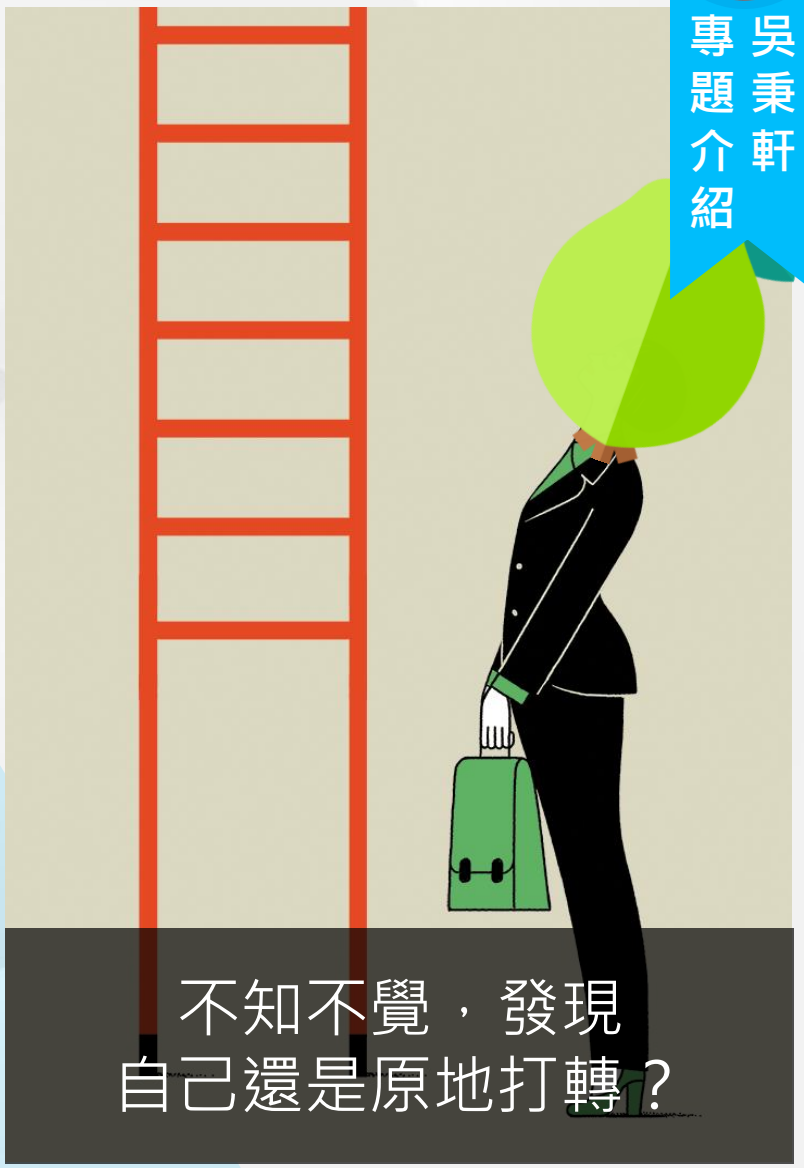
05 應用部署

06 總結

專題介紹

報告人：**吳秉軒**

日劇裡的小故事

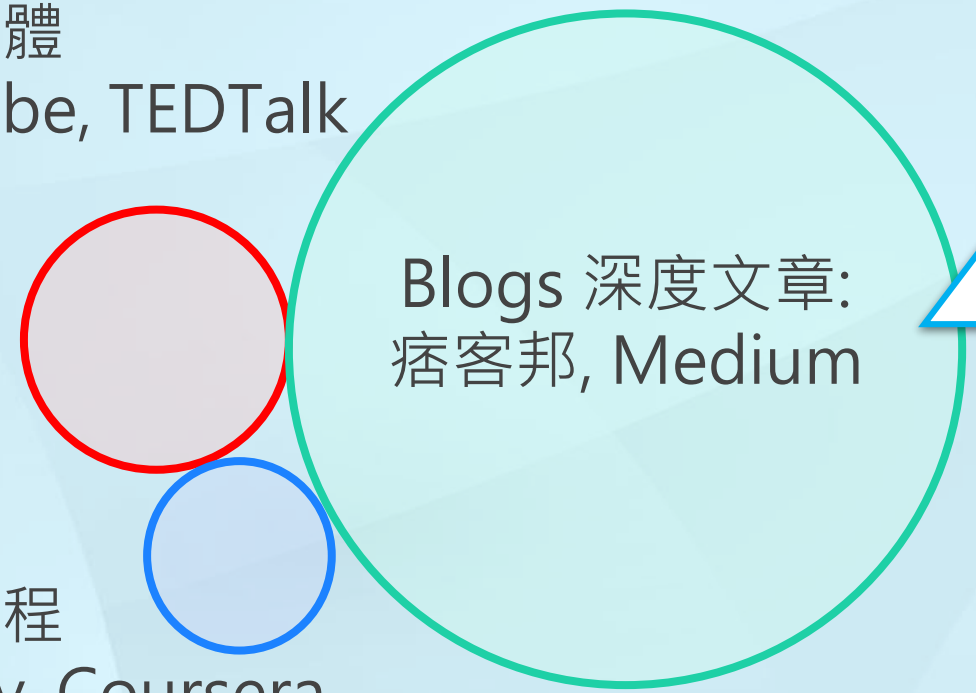


吳秉軒
專題介紹

學習，職場只是開始

- 全球行動應用市場將從 2019 年，21 億美金到 2026 年，53 億美金。
- 終生學習、學習場域為重點。「微學習」趨勢：用少於 10 分鐘獲取知識：

線上媒體
YouTube, TEDTalk



線上課程
Udemy, Coursera

快速、豐富

7,000 萬 +

每個月部落格文章新增數量

1 億 / 250 萬

全球 / 臺灣獨立造訪

來源：statista-com, optinmonster.com, pro.similarweb.com

使用部落格作為學習資源的挑戰

專題介紹 吳秉軒

- 作者面：全球最大的部落格平台，不支援中文推薦。

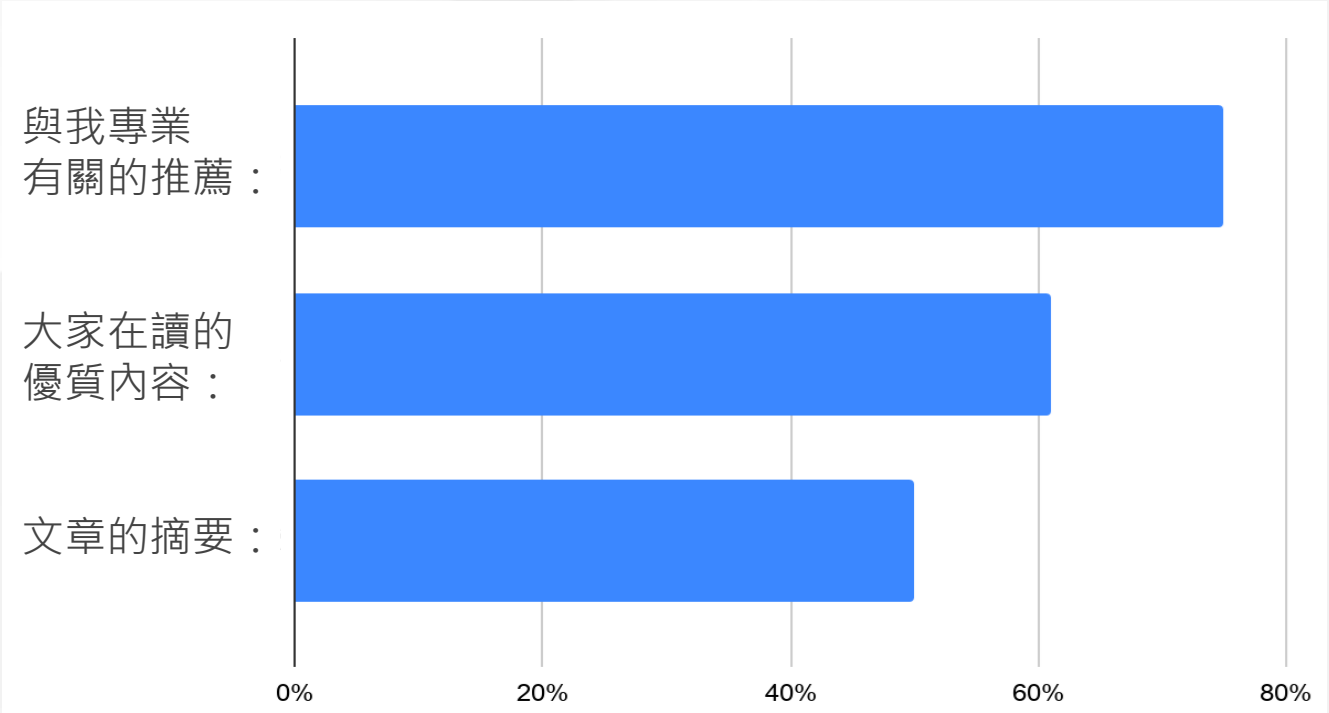
「我的文章沒人看」

再、再堅持一下Q_Q

很現實的，如《讓自己的作品被看見，是作者的責任》這篇所說，寫文章就是想要有人看，不論是想靠這賺錢、賺知名度、或是單純宣揚理念都好，否則大可不用費心使用公開平台，只要用 Bear 或是 Evernote 等工具自己整理即可。

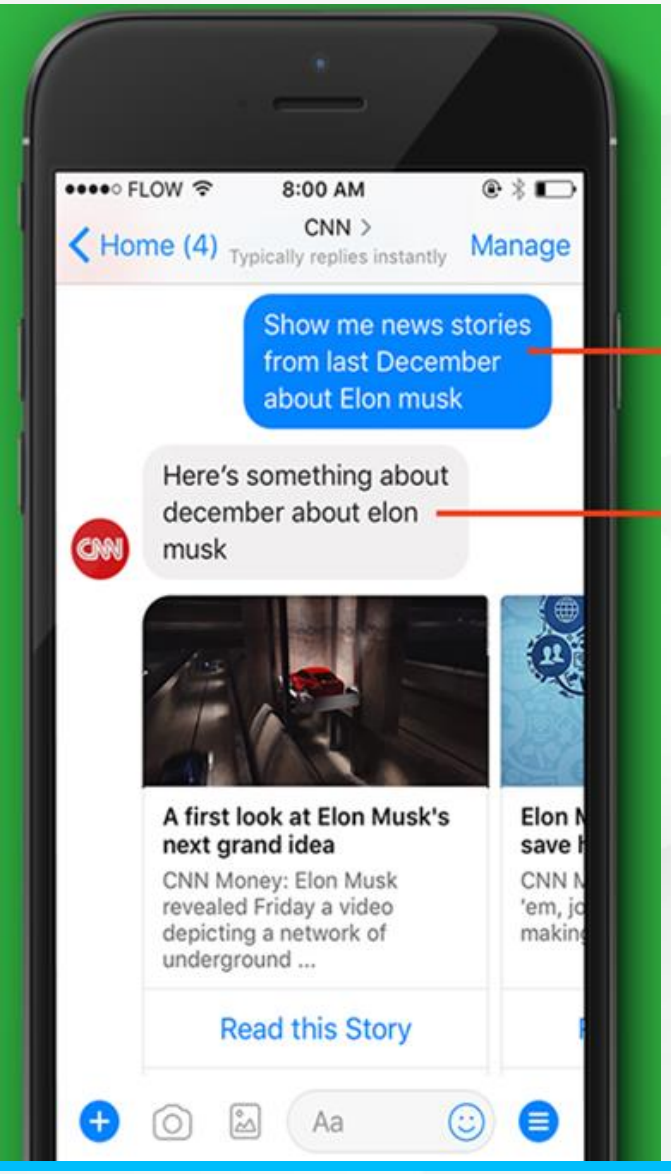
- 讀者面：龐大的資源量和「微學習」無法接軌。

調查 40 位歷屆 TibaMe 成員，最想要的功能：



來源：medium.com

當手機和 Chatbot 遇上部落格



- Medium 臺灣造訪量 42.24% 來自手機。
- 聊天機器人商業應用已成熟，臺灣 LINE 生態系完整。

吳秉軒 專題介紹

來源：pro.similarweb.com, pixnet.net

Blog樂 功能設計

- 以 LINE Bot 做部落格閱讀微學習：

機器學習：文章分類、文章推薦。

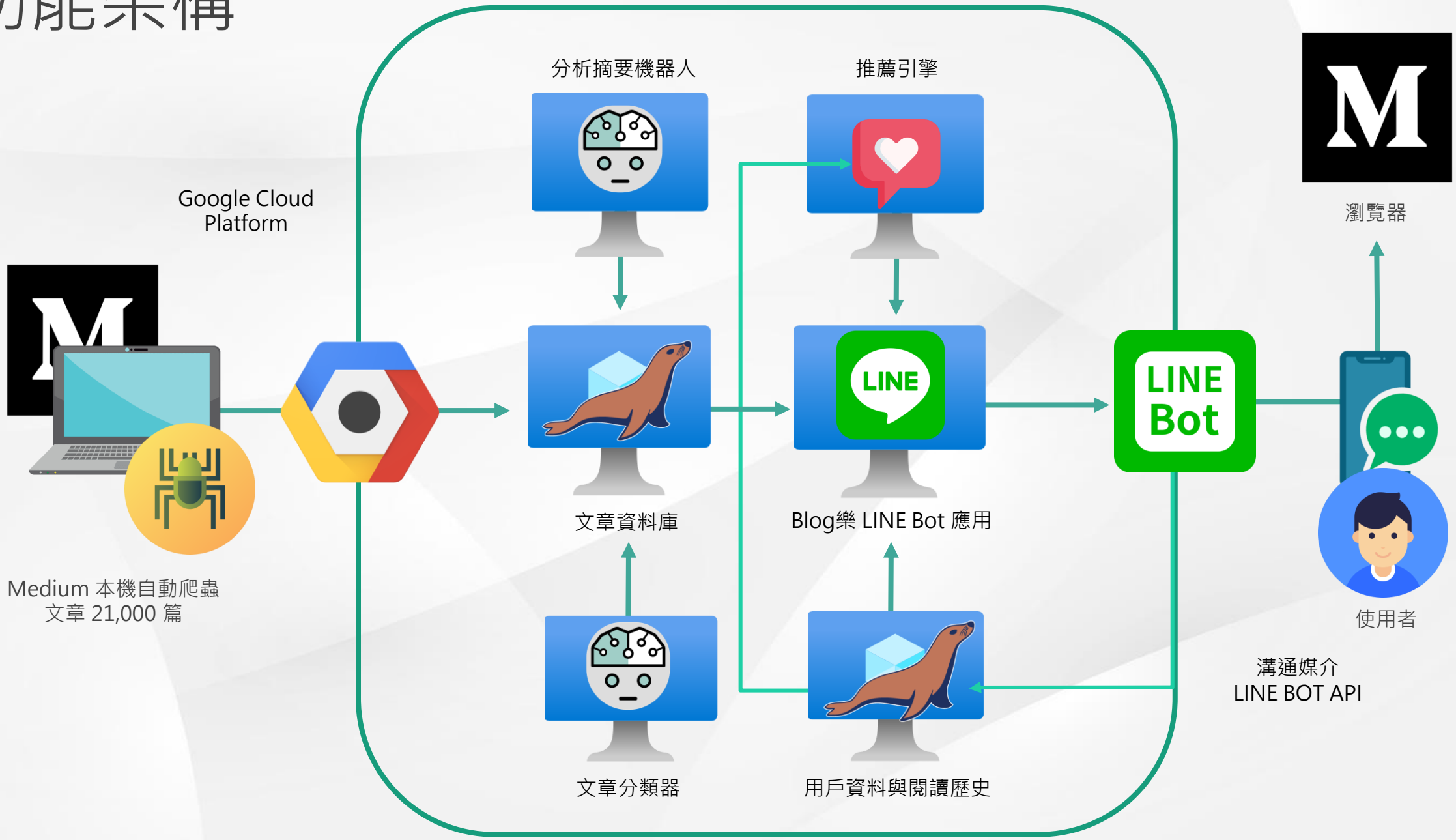
深度學習：文章摘要、文章分析。

- 建立產品黏著度正循環

Blog樂 功能展示

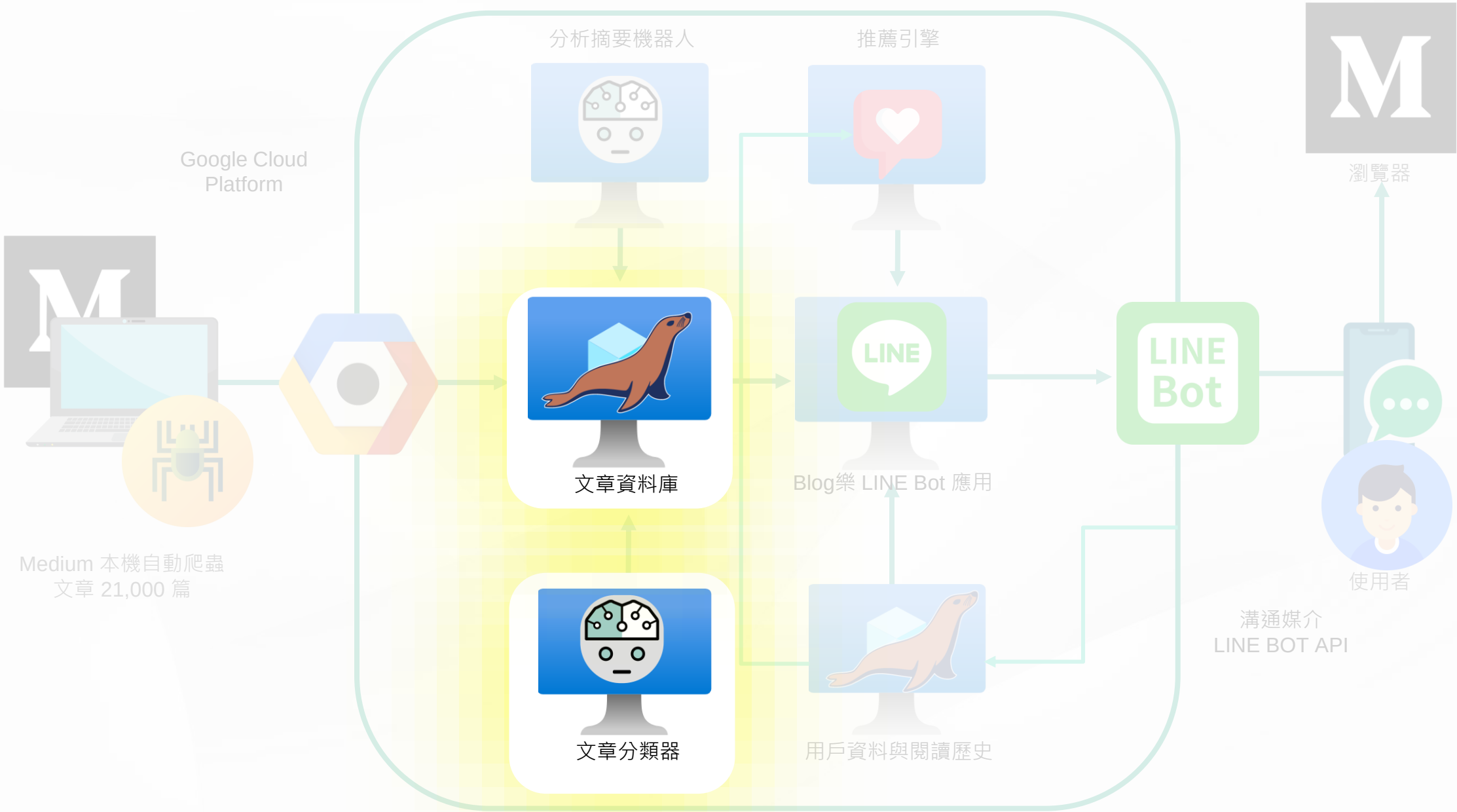
吳秉軒
專題介紹

功能架構



吳秉軒
專題介紹

功能架構：分類

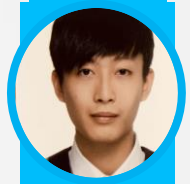



吳秉軒
專題介紹

文章分類

報告人：**劉康彥**

為什麼要做文章分類？



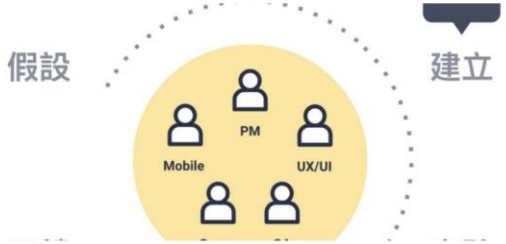
劉康彥

文章分類

- ➔ 不想看到廣告推銷文
- ➔ 快速找到想看的文章

STORIES

Renee S in Whoscall 設計師們的學習筆記
Mar 21, 2019 · 11 min read



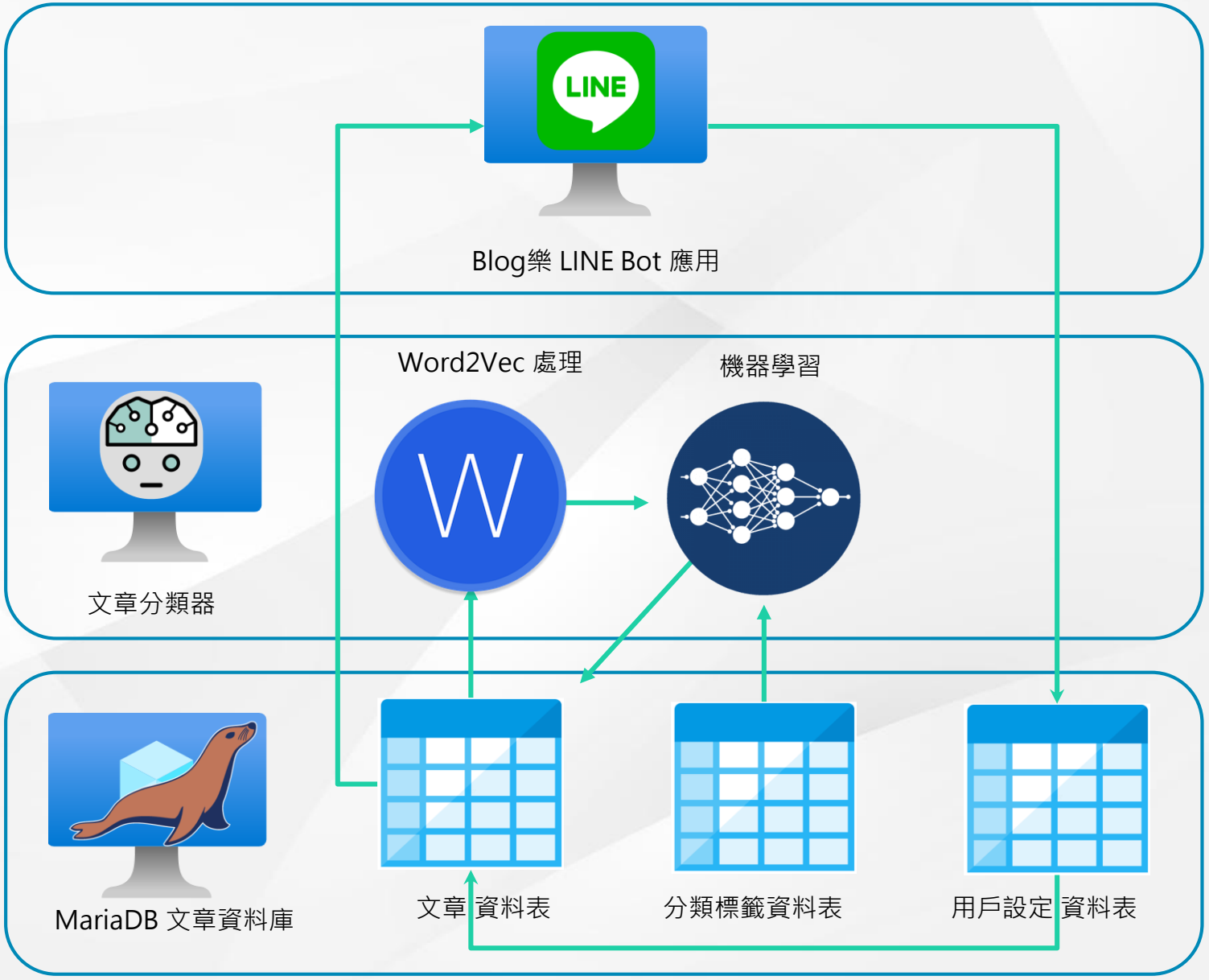
UX 設計師在敏捷團隊的因應之道

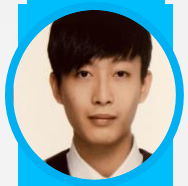
這篇文章適合：正跟著公司跑 Scrum 敏捷開發法，卻一直在流程裡載浮載沉的 UX 設計師

Read more...

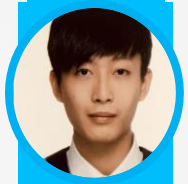
1.8K

功能構想





劉康彥
文章分類



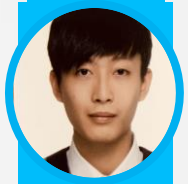
劉康彥
文章分類

資料處理流程

- 瀏覽資料庫補足缺漏值
- 訂下各大分類標記訓練資料
- 文本預處理跟 TF-IDF 特徵提取
- 嘗試使用各種模型並做挑選

註：TF-IDF(term frequency–inverse document frequency)

挑戰與解決方案



劉康彥

文章分類

STORIES

梁啟智

Jun 24, 2019 · 5 min read

如何重建一隊警察

希望你會同意，香港警察需要重建。過去兩星期的局勢發展，可見香港警察的問題已經不是個別的警員或者指揮官的問題，而是整個警隊從一哥起計到PC都出現了嚴重的文化問題。

Read more...

1K

模型預測不佳

能不能有更多選擇？

==========

第 24 篇

如何重建一隊警察 [3]

==========

增加分類

重新標記

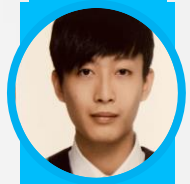
==========

第 24 篇

如何重建一隊警察 [10]

==========

文本分類方式解說與結果



劉康彥
文章分類

二分類

不符合分類要求

多分類

部分文章分類困難

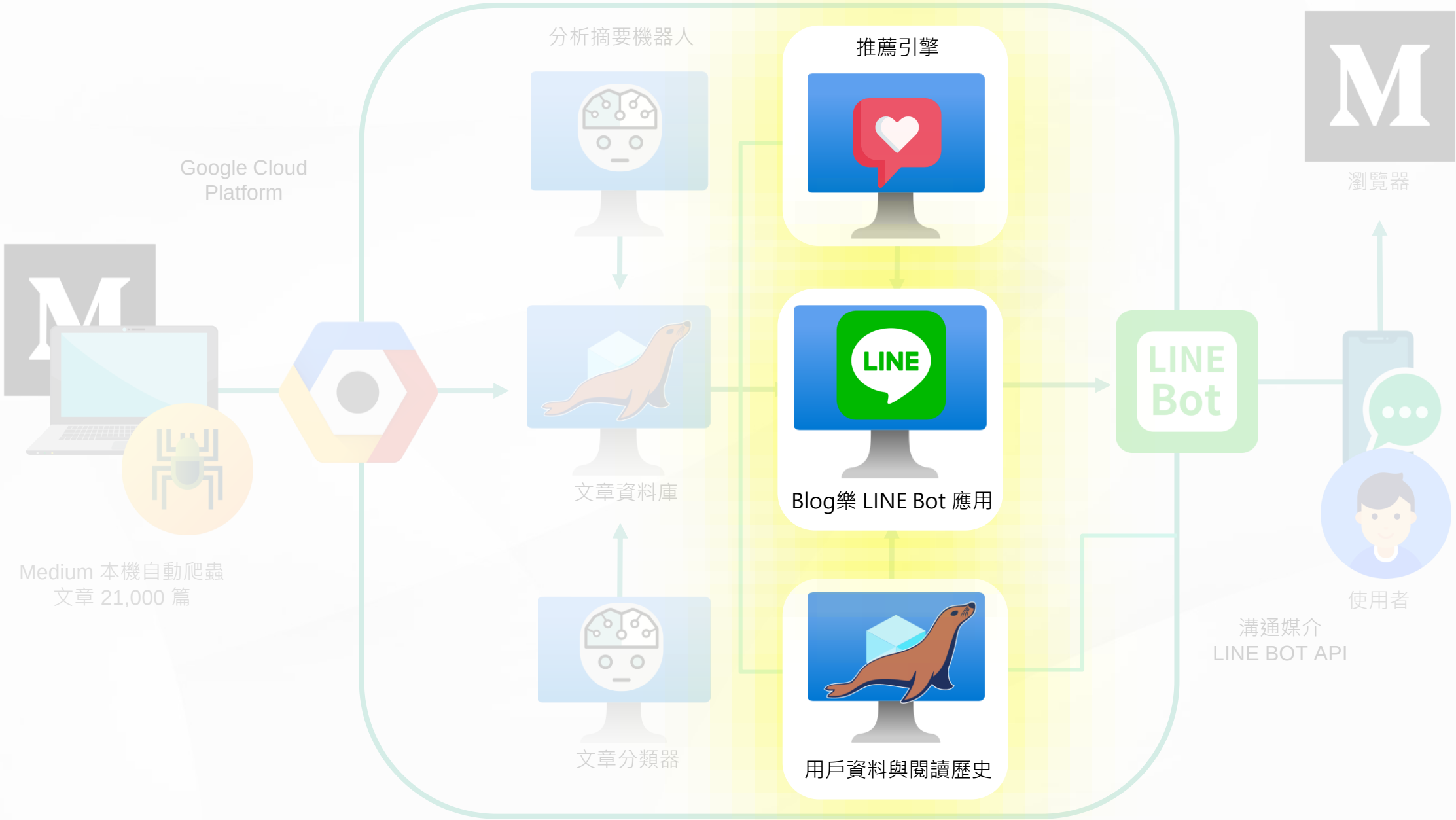
多標籤

更為精準的分類

ID	標題	分類	工程開發	資料科學	使用設計	產品管理	行銷管理	商業投資	職場心得
580	**UX設計師在敏捷團隊的因應之道**	3	0	0	1	1	0	0	1

Top 1 文章分類準確度 75% 以上

功能架構：推薦

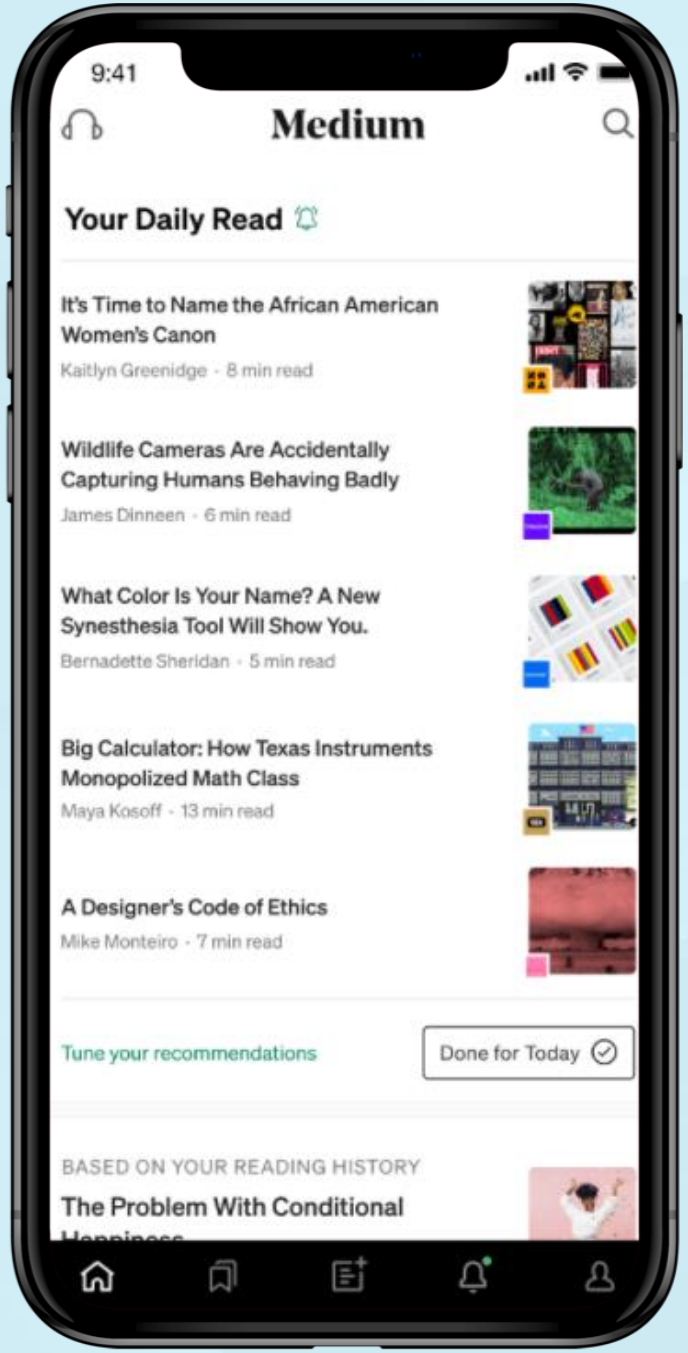


劉彥康

文章分類

文章推薦

報告人：**林靖凱，丁之琳**

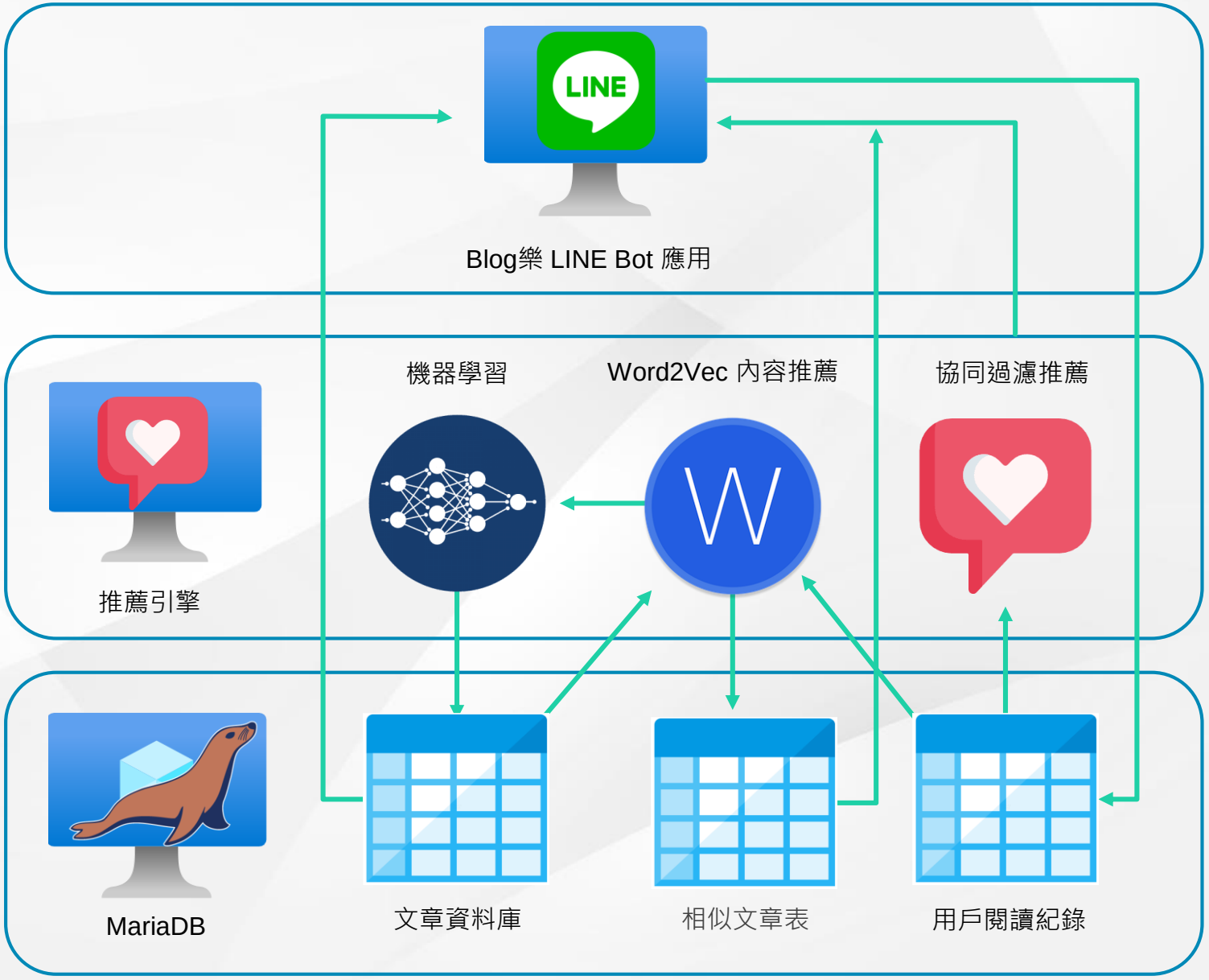


為什麼要推薦系統？

Medium 官方 APP 的呈現

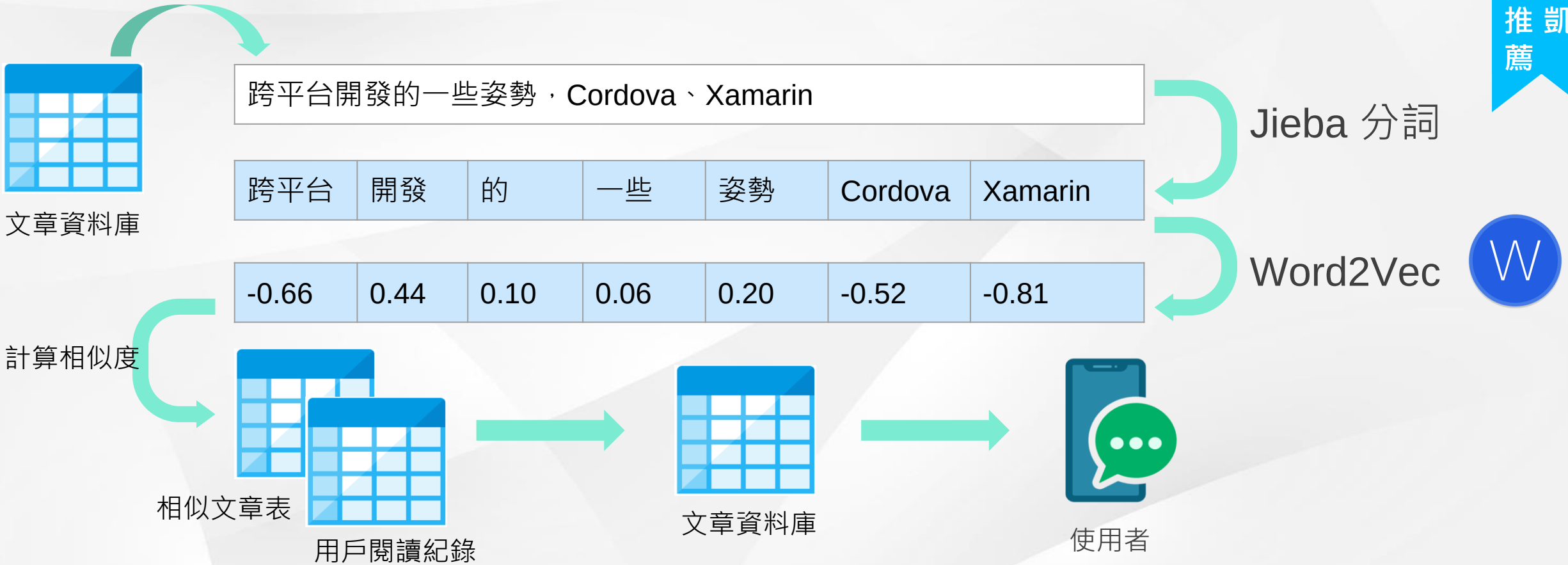
- 沒有中文的文章推薦
- 沒有中文的分類
- 沒有依照讀者過往的閱讀喜好推薦文章

功能構想



文章推薦 林靖凱

推薦文章：過去閱讀記錄推薦 (Content Based)



推薦文章：協同過濾 (Collaborative Filtering)

推薦系統套件：Surprise

用戶閱讀紀錄

	A article	B article	C article	D article
A user	5	4	?	?
B user	?	4	?	3
C user	?	?	1	2

算法：SVD

	A article	B article	C articl	D article
A user	5	4	?	?
B user	5	4	2	3
C user	?	?	1	2

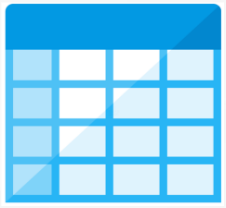
使用者

文章資料庫

挑戰與解決方案

林靖凱
文章推薦

內文相近度推薦，
計算時間較長



相似文章表

增加資料表，將計算的
結果提前存入

冷啟動狀況，使用者
喜好及人數還不夠多

- 熱門分類文章
- 閱讀記錄推薦
- 協同過濾推薦

依照使用者已閱讀
篇數逐步採用開啟
更多推薦方式

結果呈現

文章推薦 林靖凱

w2v 閱讀記錄推薦

```
輸入文章是
資料庫內第 130.0 篇
文章ID: 693
原文標題:[私心推薦] PM / UX / UI 學習資源總整理
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
相似文章排序：
資料庫內第 3596.0 篇
相似度: 0.962354302406311
文章ID: 4708
推薦標題:如何推動設計文化與創新思維
----------------------------------------------
資料庫內第 7706.0 篇
相似度: 0.9610608220100403
文章ID: 9708
推薦標題:騰訊 UIUX 暑期實習分享全記錄
```

surprise 協同過濾推薦

```
IntputFindCmd:
 SELECT `title`  FROM `main` WHERE `id` = 8045
文章ID： 8045
可能的喜好程度： 4.037292825543596
推薦標題： 適合練習照片牆 App 的 flickr API
----------------------------------------------
IntputFindCmd:
 SELECT `title`  FROM `main` WHERE `id` = 4884
文章ID： 4884
可能的喜好程度： 3.798032858903403
推薦標題： 不只是抓資料，API 可以為我們提供各種服務
----------------------------------------------
IntputFindCmd:
 SELECT `title`  FROM `main` WHERE `id` = 1625
文章ID： 1625
可能的喜好程度： 3.716170735378744
推薦標題： 線上支付體系的大戰略思考
```

推薦功能延伸：優質新文預測

丁之琳

優文預測

預測近期具有潛力的優質新文章，並在推薦功能內回推給使用者

以機器學習判斷優質新文推薦

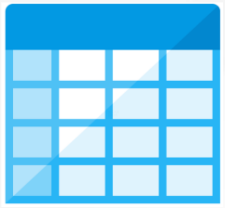
優文預測 丁之琳

- 以固定算式標記訓練集文章熱門程度。
- 不同熱門等級的內文做文本特徵處理。
- 使用機器學習模型預測近期文章熱門程度。

文本預處理與模型選擇

丁優之文琳預測

- 以 TF-IDF 進行文本特徵提取
- 挑選合適模型



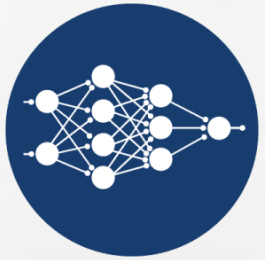
文章資料庫

Jieba 分詞 + TF-IDF

特徵提取

(0, 5416) **0.0379788**
(0, 3489) 0.5089233
……

Naive Bayes



較小的訓練數據表現優秀
訓練速度快

Logistic Regression

模型評估結果

丁之琳

優文預測

- 預測近六個月內的新文在不同文章評比等級的機率，並將高品質文章回推給使用者

文章資料庫

載入預測文章資料

第 342 篇 : PM三點下午茶 #6 專案經理要怎麼跟老闆做好例行專案進度報告呢？

等級:1

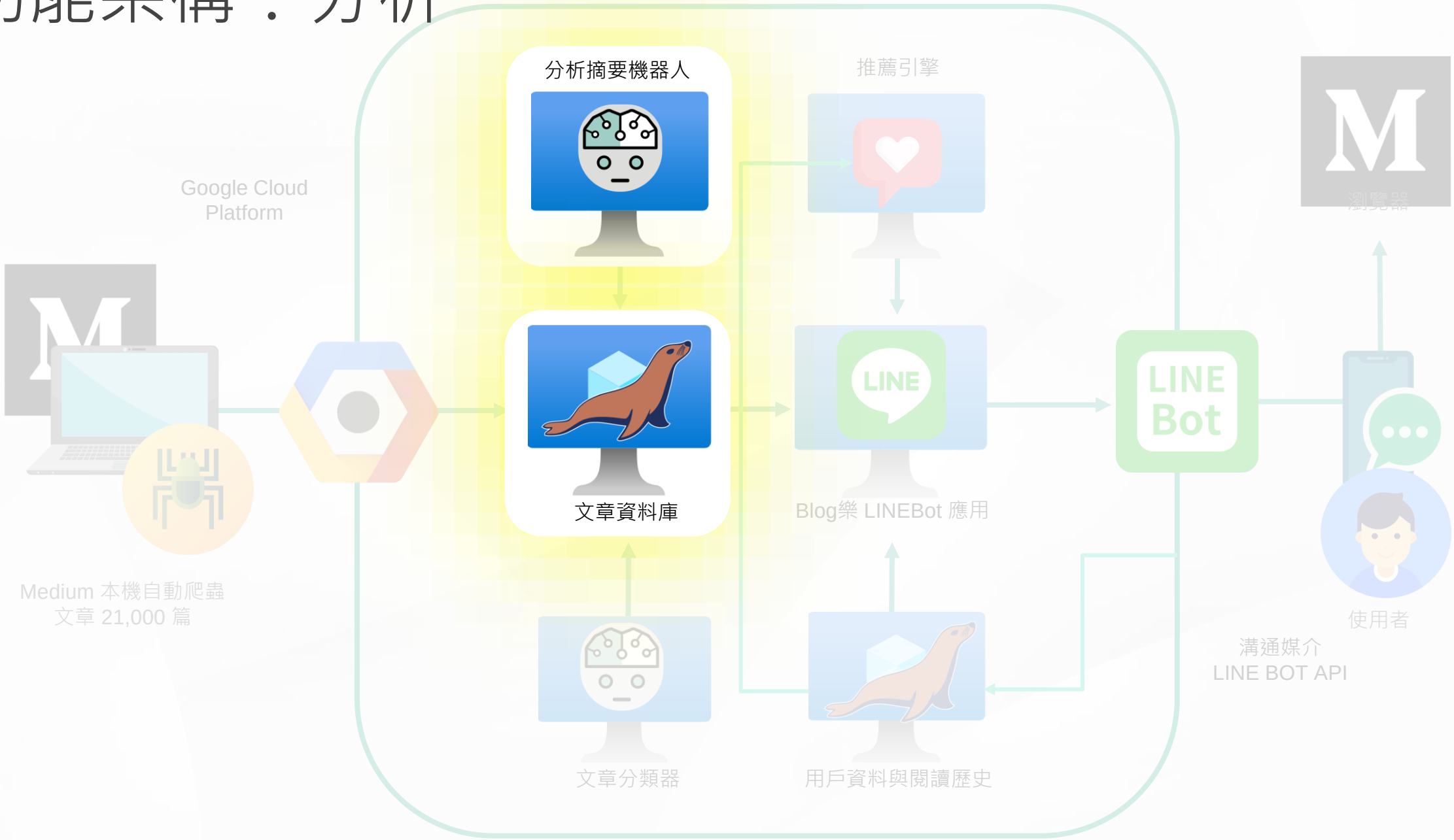
高 : 0.6724

中 : 0.2345

低 : 0.0931

文章資料庫

功能架構：分析



丁之琳
優文預測

文章分析：趨勢分析

報告人：**袁學平**

趨勢分析的需求

使用者

獲取新知 VS 新的靈感

文章作者

功能構想：提供關鍵詞

- 紀錄使用者搜尋關鍵詞：基本做法
- 由 A.I. 分析文章內容抽取關鍵詞：隨時可用

功能架構


袁學平
趨勢分析


排序/選取關鍵詞
分析摘要機器人
趨勢關鍵詞提供
NLP處理
分詞工具/排序模型
新關鍵詞/文章
匯入文章
文章資料庫
LINE Bot
使用者
匯入
用戶資料與閱讀歷史
搜尋和閱讀結果

分詞工具： Jieba / CKIP

袁學平
趨勢分析

探尋顧客內心最真實的意圖-關鍵字搜尋分析（Keyword Search Analytics）

行銷資料科學 Follow
Nov 16, 2018 · 3 min read

以市場區隔與機器學習篩選有價值的顧客

己想要的資訊。企業則希望它的服務或產品，能夠出現在搜尋結果的第一頁，讓消費者有更多的機會予以點擊，而「關鍵字」就是串連「搜尋引擎天秤」兩端的媒介。

「關鍵字」指的是，網路使用者在搜尋引擎裡輸入所欲搜索查詢的字後，經網站比對到的字，即稱為「關鍵字」（keyword）；而「關鍵字搜尋分析」則是網路行銷中很重要的一種分析工具，其有助於企業不斷優化網站設計，以及達成網路行銷的目的。

閱讀更多：翻轉零售— 大數據帶來的零售業革命

從技術上來看，網路使用者在搜尋引擎裡所輸入的字稱為「搜索查詢」（Search Query），而經網頁比對到的字被稱為「關鍵字」（keyword）。

❑ 將中文文本分成詞以**抽取關鍵詞**

Jieba

以 市場 區隔 與 機器 學習 篩選有 價值 的 顧客

CKIP
Chinese Knowledge Information Processing Group

除 Jieba 外再導入 CKIP 中英文分詞結果

模型排序關鍵詞

註：CKIP (Chinese Knowledge Information Processing)

排序模型：BERT

- 目標 NLP 模型 BERT 實作
- 先標籤關鍵詞以供訓練

市場　深度　機器　學習

區隔　Machine　Learning

所有訓練文章

BERT

編碼後文章
作為標籤

0	1	1	1	1	0	1	1	1	1	0	0	…

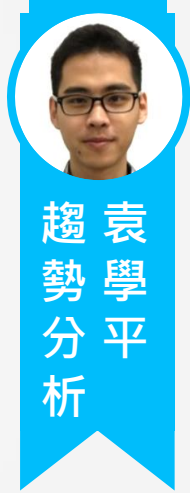
市場區隔　簽約金　機器學習

Machine Learning　深度學習

以 **市場區隔** 與 **機器學習** 篩選...

註：BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding)

分析結果：關鍵詞重要度序列



標籤

近期文章

BERT

重要度編碼

…	0.8	0.8	0.8	0.8	0.1	0.5	0.5	0.5	0.5	…

對應文章篩選關鍵詞

…機器學習或深度學習…

統計所有近期文章後…

工程和資料分析類
排序結果:

1. NativeScript
2. Flutter
3. Android
4. Electron
5. MacOS

功能架構：摘要


分析摘要機器人
推薦引擎
Google Cloud Platform
文章資料庫
Blog樂 LINE Bot 應用
LINE
LINE Bot
Medium 本機自動爬蟲
文章 21,000 篇
文章分類器
用戶資料與閱讀歷史
使用者
溝通媒介
LINE BOT API


袁學平
趨勢分析

文章分析：摘要生成

報告人：**陳豐王**

摘要，有需要嗎？


敏捷式開發(Agile)、瀑布式開發(Waterfall)、敏捷式UX、Lean UX，兜幾？

敏捷式開發 (Agile Methodology)

瀑布式開發(Waterfall)

產品需求 → 設計 → 開發 → 驗證 → 保養更新

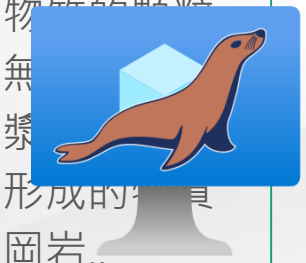
功能構想

陳豐王
文章摘要

內文

地球表層均由堅硬的岩石構成，是人類最常使用的自然資源之一。用肉眼觀察，可以發現各種岩石的外觀、顏色與質地各自不同，依成因可以分為火成岩、沉積岩及變質岩三大類。由岩漿冷卻凝固形成的岩石是火成岩，若岩漿噴發至地表快速冷卻，其物質的顆粒較小，甚至肉眼無如玄武岩；若岩漿處緩慢冷卻，則形成的顆粒較大，如花岡岩..

文章資料庫



摘要生成器

分析摘要機器人

摘要

(條列式：)

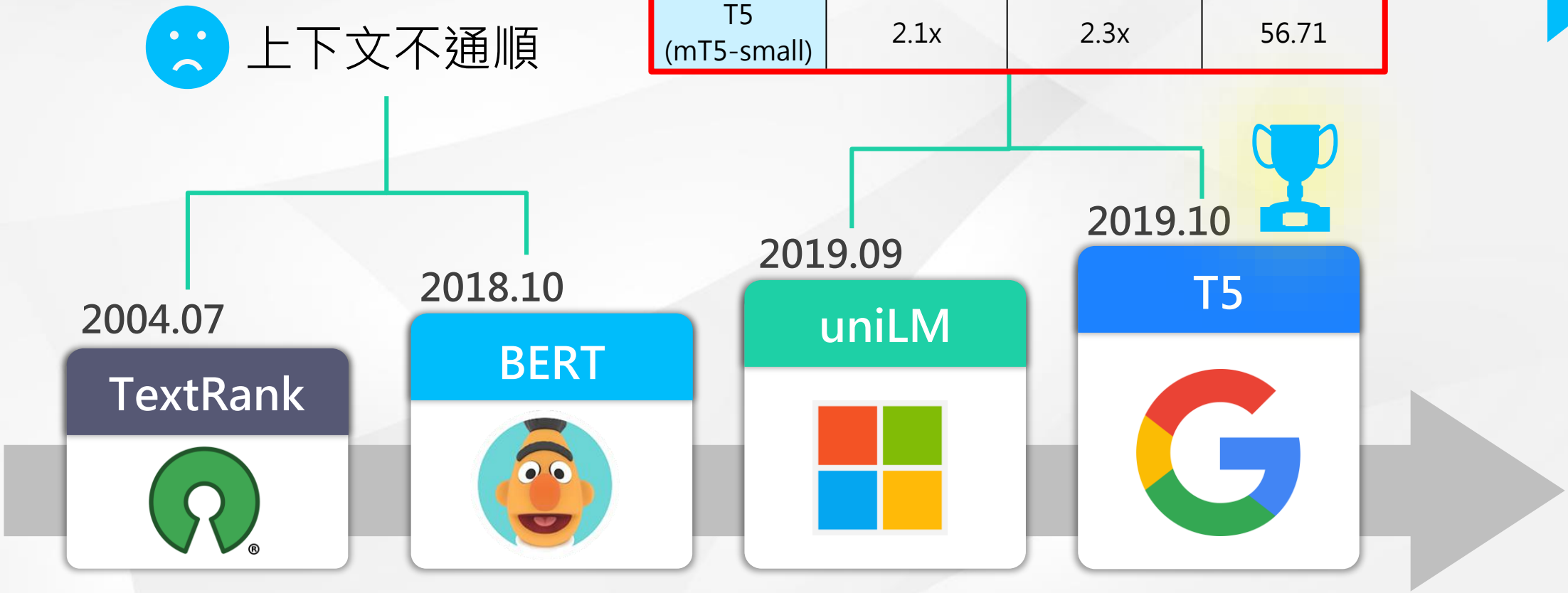
1.地球表層由岩石所構成。
2.依成因可以分為火成岩、沉積岩及變質岩三大類。
3.岩漿冷卻凝固形成的岩石是火成岩，如玄武岩、花岡岩。

(敘述式：)

地球表層係由岩石所構成，依其成因可以分為火成岩、沉積岩及變質岩三大類。火成岩係由岩漿冷卻凝固形成，如玄武岩、花岡岩。

演算法/模型選用

	訓練速度 (1 Epoch)	生成速度 (beam size= 1)	摘要效果 (Rouge-2)
uniLM (BERT-base)	1x	1x	54.91
T5 (mT5-small)	2.1x	2.3x	56.71

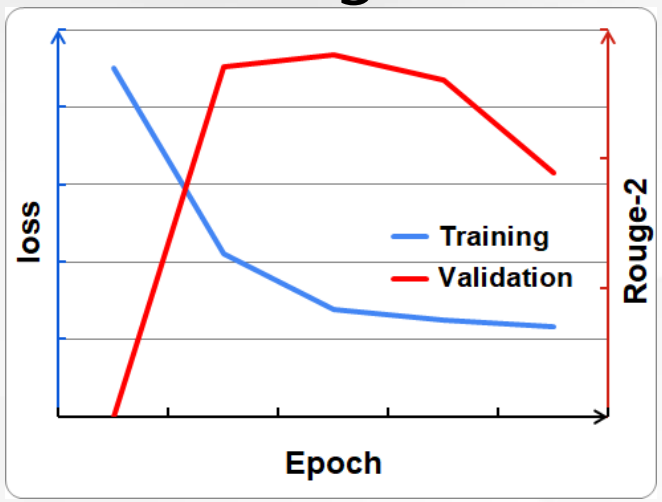


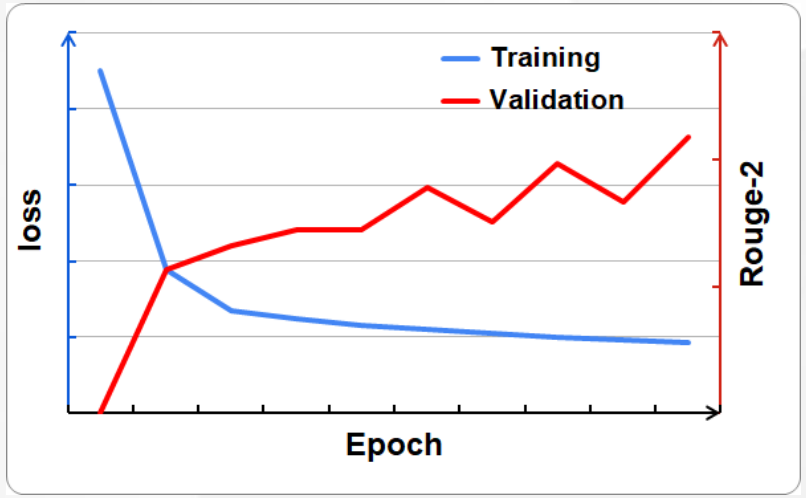
註：Rouge(Recall-Oriented Understudy for Gisting Evaluation)
uniLM(Unified Language Model Pre-training for Natural Language Understanding and Generation)
T5(Text-To-Text Transfer Transformer)

遭遇挑戰與解決方式

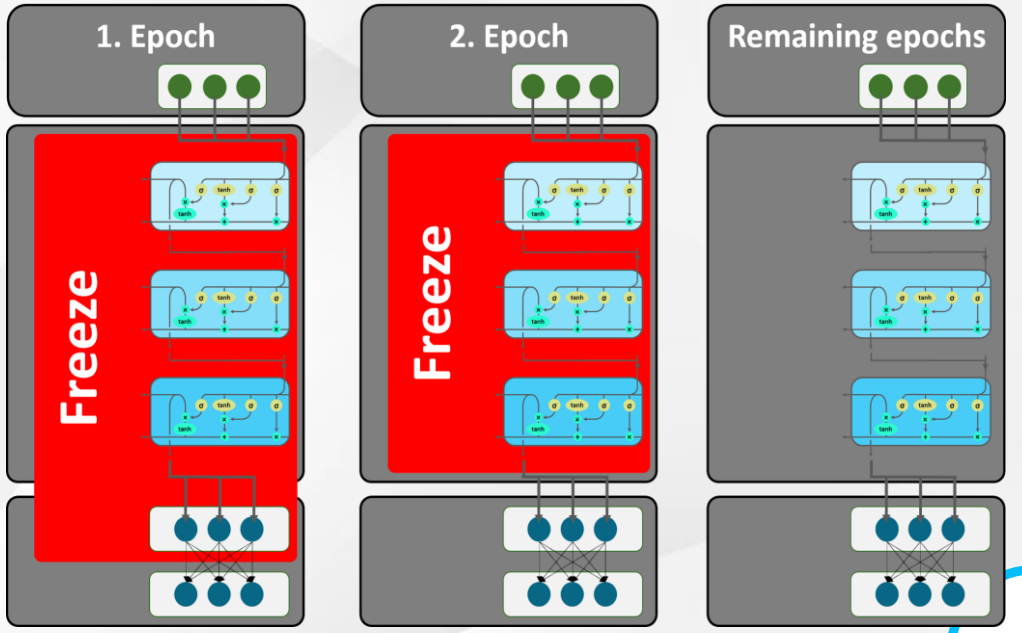
文章摘要 陳豐王

Overfitting

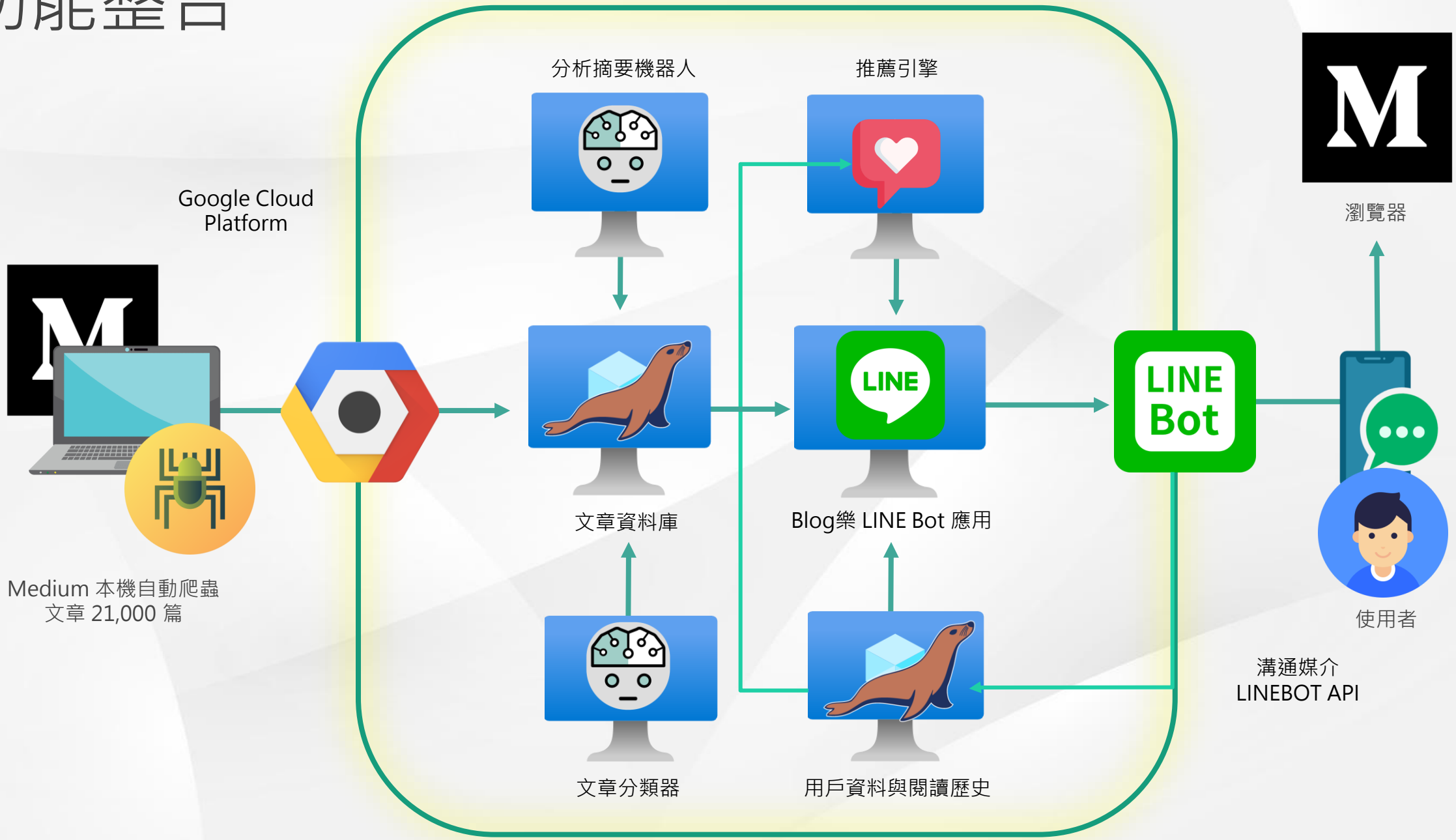




Gradual Unfreezing
逐步解凍



功能整合



陳豐王
文章摘要

應用部署

報告人：**孫韶甫**

上述功能怎麼統合完成？

呈現方式？

網頁？ LINE Bot？

資料庫型式？

MySQL？ MariaDB？
MongoDB？

團隊程式開發？

如何整合？ Git？
Code review？

雲端平台？

Azure？ GCP？ AWS？

Google Cloud Platform

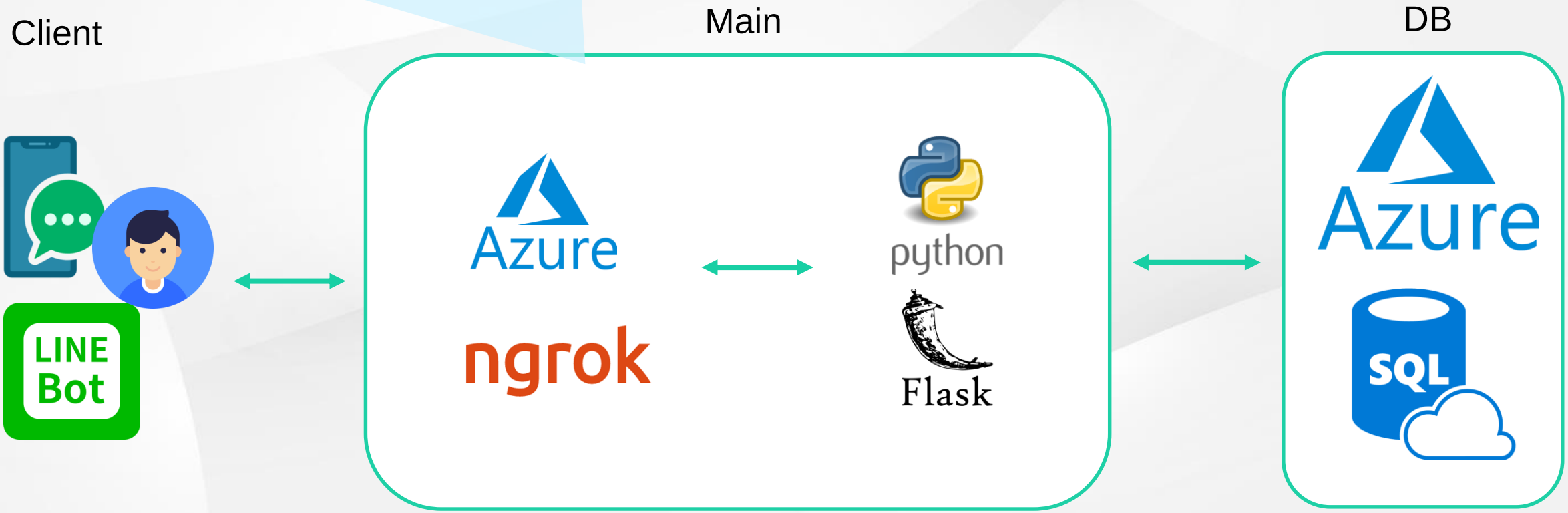
網路架構？

NGINX？ Apache？ uWSGI？
ngrok？ Flask？ Django？...

孫韶甫
應用架構

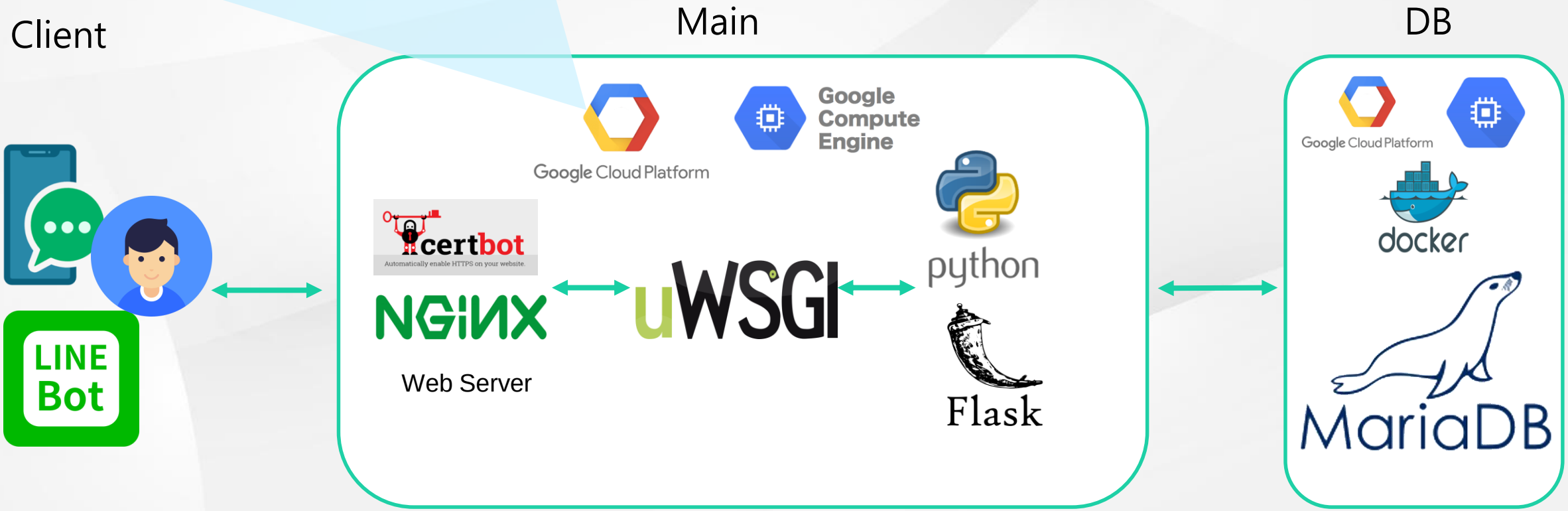
測試階段架構圖

- Azure App Service：長時間沒使用需喚醒。雲端機器收費較昂貴。
- ngrok：每分鐘連線次數有限制。8小時需重新開啟服務(免費版)。
- Azure SQL Server收費：雲端機器費用 + 資料庫容量費用。



部署階段架構圖

- Certbot：免費產生 SSL 憑證。
- NGINX：高效的 Web Server。
- uWSGI：接收 NGINX 動態請求並處理後，發給 Web Application。
- WSGI：Python Web Server GateWay Interface。



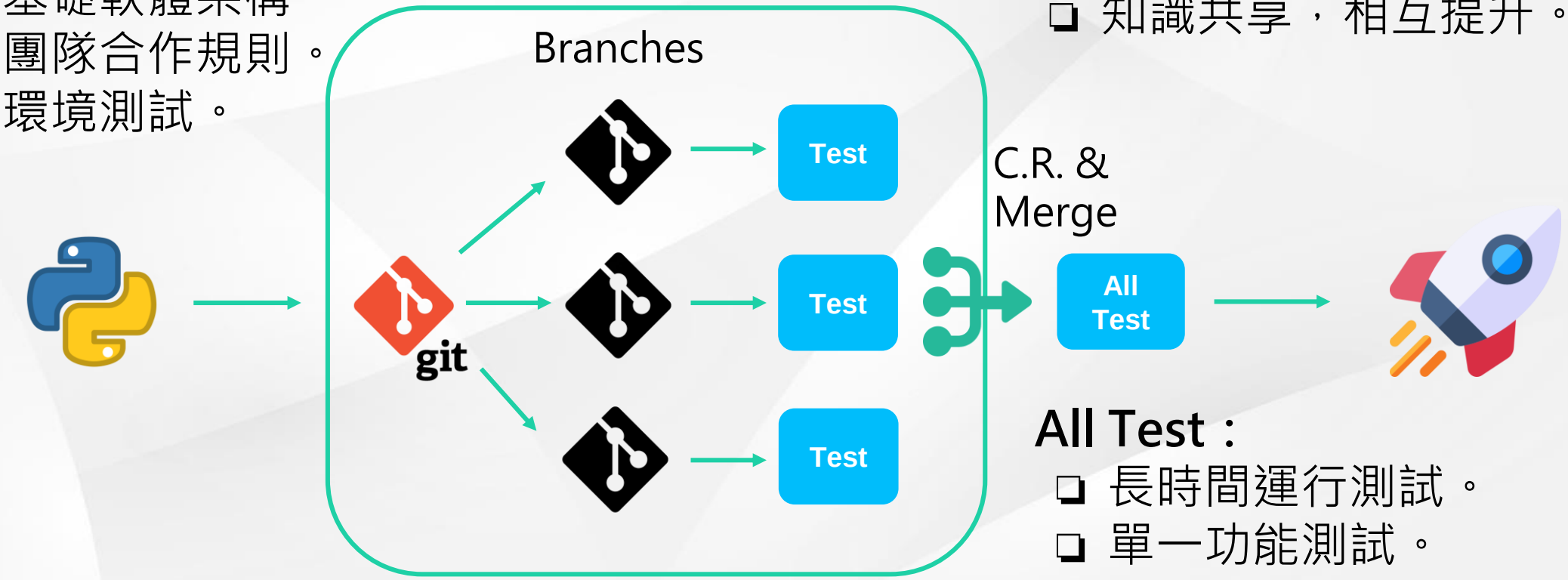
團隊軟體開發

孫韶甫
應用架構

前期準備：

- ❑ 建立功能架構圖。
- ❑ 建立基礎軟體架構。
- ❑ 訂定團隊合作規則。
- ❑ 運行環境測試。

團隊開發 (GitHub flow)



C.R. (Code Review)：

- ❑ 檢閱開發程式碼。
- ❑ 程式碼討論優化。
- ❑ 知識共享，相互提升。

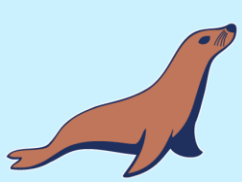
All Test：

- ❑ 長時間運行測試。
- ❑ 單一功能測試。
- ❑ 同時間多人使用測試。

總結

技術總覽

採用服務：




docker



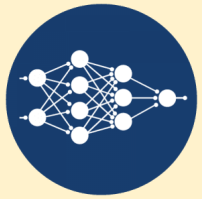
uWSGI NGINX
LINE Bot

使用技術：

- 分詞：
  Jieba、CKIP
- 詞轉向量：
  Gensim、TF-IDF
- 機器學習：
  Naive Bayes、
  Logistic Regression
- 分類系統：
  Label Power Set
- 推薦系統：Scikit-surprise
- 自然語言：
  BERT、UniLM、T5

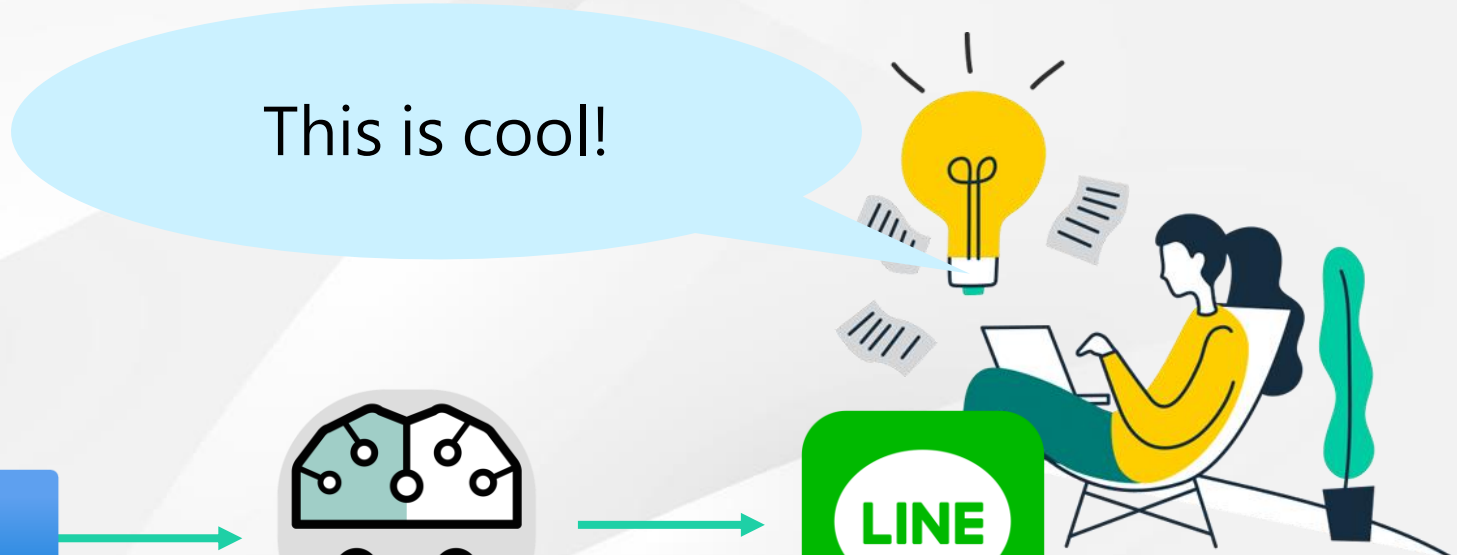
適用領域：

- 大數據資料處理
- 行銷資料分析
- 電商推薦系統
- 數位教學系統
- 數位客服系統

專案總結與未來展望

- 來源擴展與動態執行。
- 探索更多 A.I. 自然語言的功能延伸。



爬蟲上雲

資料庫與服務動態執行

A.I. 延伸功能

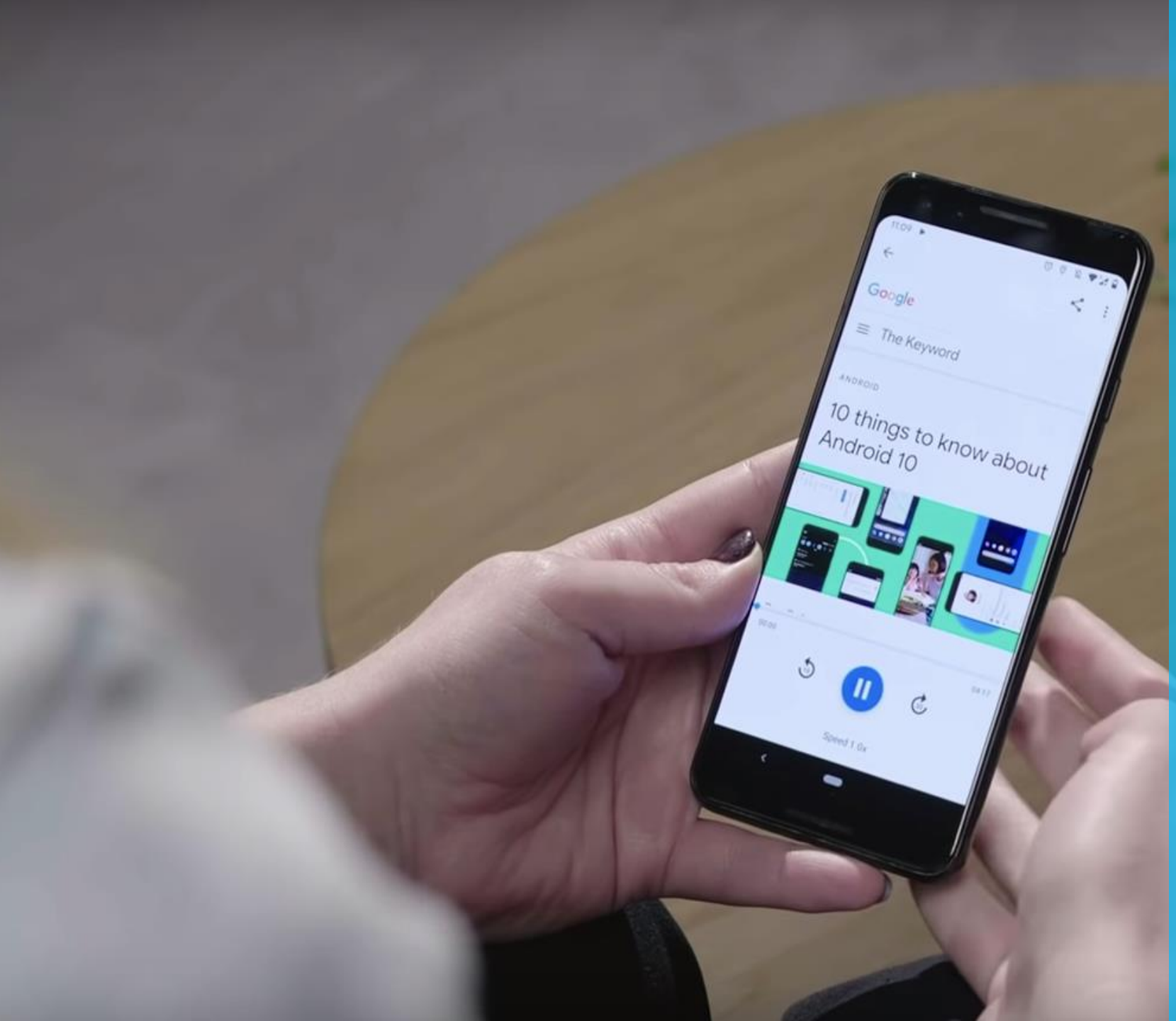
長期服務

團隊分工

專題總結 吳秉軒

	專案管理	資料處理	分類	推薦	分析	摘要	LINE Bot	雲端部署
吳秉軒	★	★		☆	☆		☆	
劉康彥		☆	★	☆			☆	
林靖凱		☆	☆	★			☆	
丁之琳				★	☆	☆	☆	
袁學平			☆		★	☆	☆	
陳豐王		☆			☆	★	☆	
孫韶甫	☆	☆					★	★

★ 負責 ☆ 參與



報告結束

吳秉軒、袁學平、劉康彥、孫韶甫
林靖凱、陳豐王、丁之琳

感謝您的聆聽