

天主教輔仁大學圖書資訊學系碩士班碩士論文

指導教授：陳舜德 博士

基於自然語言處理技術整合維基百科(Wiki)之
圖書館參考諮詢機器人建置與使用評估



Development and Evaluation of a Library Chatbot
for Reference Service Integrating Natural Language
Processing Based on Wiki

研究生：黃丰嘉 撰

中華民國 110 年 6 月

私立輔仁大學圖書資訊學系碩士班
論文口試委員會審定書

黃丰嘉 先生之碩士學位論文

基於自然語言處理技術整合維基百科 (Wiki) 之
圖書館參考諮詢機器人建置與使用評估

Development and Evaluation of a Library Chatbot for
Reference Service Integrating Natural Language Processing Based on Wiki

經本委員會審議合格，特此證明。

論文口試委員

周遵儒 (召集人)

周遵儒

顏榮泉

顏榮泉

指導老師

陳舜德

陳舜德

系主任

李正吉

李正吉

中華民國 110 年 06 月



致謝

經過一年半的磨練，回首撰寫這本論文的點點滴滴，心中有說不出的感謝與感動。這一路走來，我有幸獲得不少人的幫助、建議以及督促，才得以如期完成這本碩士論文。

首先，我最想感謝的是我的指導教授陳舜德老師，不僅費心為我找尋相關資源以建構自然語言處理領域的整體概念，還耐心指導我的論文寫作、提供我程式測試上的寶貴意見，更在生活輔導上給予毫無保留的關懷照顧。不只如此，老師的勉勵、提點以及直言說出我不足之處，讓我受益匪淺，也激起了我持續突破的動力。平時也會與我分享一些關於當前資訊產業方面的報導、技術...等，不只讓我有機會實際操作軟硬體設備，也對於資訊領域有更進一步的了解、認識，這讓我由衷慶幸自己跟到一位很為學生著想的好老師。

其次，我要感謝的是在提論文計畫書和口試時國立臺北教育大學數資系顏榮泉教授和臺灣師範大學圖傳系周遵儒教授給予本論文品質上的高度肯定，還有點出書面撰寫上架構性不足的問題，以及對於未來投稿上的寶貴建議，使本論文更佳完善，在此致上深深的謝意。

接著，感謝有著深厚革命情感的一群夥伴們——崇瑋、堃陽，這一年多來在無數次提計畫書、進度報告以及每月成果報告中，耐心聆聽我的簡報，並一針見血地針對內容進行提問，給予我簡報表達上的調整建議，以及督促我按部就班的如期完成我的碩士論文。

最後，要感謝圖資系的助教們、我的好友相姘以及我的家人，還有許多曾經幫助過我的人，因為有你們一路的支持、鼓勵與陪伴，才能讓我在寫論文寫到煩躁時有個抒發、喘息的空間，並在調整完心情後重新埋首寫論文。

本研究感謝科技部「基於自然語言處理技術整合維基百科(Wiki)之圖書館藏查詢參考機器人建置與使用評估計畫」支持，計畫編號：MOST 109-2410-H-030-077

摘要

自 2016 年起，聊天機器人 (Chatbot) 的風潮崛起且應用遍地開花，舉凡股市理財投資 (Line @微股力 ScanTrader)、旅遊搜尋推薦 (Facebook Messenger @Skyscanner)、披薩訂購 (Google Dialogflow @達美樂 Domino) ... 等皆能看到聊天機器人的身影。聊天機器人的出現，提供使用者客製化、不受時空限制的服務體驗，在大大節省客服人力資源之餘，也提高處理重複性工作的效率。如今，隨著語意分析 (semantic analysis)、機器學習... 等技術的大幅躍進，聊天機器人不再是簡單的「關鍵字觸發」任務，而是從大量語料中透過對字詞、句型、句法結構的分析來理解使用者的詢問目的 (意圖)，也就是所謂的自然語言理解 (NLU, Natural Language Understanding) 技術。

圖書館作為重要的知識典藏機構，應當思考如何運用新興科技提供讀者一個更加行動化、個人化與智慧化的參考諮詢介面，從而發展出一套便捷、友善、效能的服務模式，是個值得期待的嘗試。因此，本研究以圖書館為應用場景、LINE 為圖書館線上參考諮詢的系統介面，實際建置一個解答館務問題和提供學科知識服務的參考諮詢機器人—FJCU Lib DialogSys。有別於「規則式」的制定問句的問法，本研究的貢獻在於讓讀者與系統之間的對話得以彈性，經由下列三種方式達成：其一，利用「維基百科 (Wikipedia)」更新即時、詞彙描述淺顯多樣以及內容豐富且持續成長... 等特性，協助問句進行查詢擴展 (Query Expansion)、作為斷詞 (Word Segmentation) 程序的擴增語彙集，以及用以回應讀者學科知識諮詢。其二，在問句的處理上，運用「自然語言處理 (NLP, Natural Language Processing)」技術來幫助判斷讀者的詢問意圖，進而獲得切合其資訊需求的答覆。其三，在界定詢問意圖的顆粒度方面，本研究使用基於密度的 DBSCAN 分群演算法藉由參數的調整將具有類似問法的訓練語料歸為同一個意圖，並儲存於 MongoDB 內。而後，若有增加其他意圖的必要性時，也能輕易地擴充訓練語料，有助於後續的維護與利用。

關鍵字：意圖式聊天機器人、自然語言處理、Wikipedia、查詢擴展、特徵擷取、DBSCAN

Abstract

Since 2016, the trend of Chatbot has risen and applications have blossomed everywhere, such as stock market financial investment (Line @微股力 ScanTrader), travel search recommendations (Facebook Messenger @Skyscanner), pizza ordering (Google Dialogflow @達美樂 Domino), etc. The emergence of Chatbot provides users with a customized service experience that is not limited by time and space, which greatly saves customer service human resources and improves the efficiency of handling repetitive tasks. Nowadays, with the rapid advancement of semantic analysis, machine learning, and other technologies, chatbots are no longer "keyword trigger" tasks, but through matching words, sentence patterns, and syntax from a large amount of corpus. The analysis of the structure to understand the user's inquiry purpose (intention), which is called natural language understanding (NLU) technology.

As an important knowledge collection institution, libraries should think about how to use emerging technologies to provide readers with a more mobile, personal, and intelligent reference interface, to develop a set of convenient, friendly, and efficient service models, which is worth looking forward to trying. Therefore, this research uses the library as the application scenario and LINE as the library's online reference consultation system interface and builds a reference consultation robot — FJCU Lib DialogSys, which answers library FAQs questions and provides subject knowledge services. Different from using the "Rule-based" method to set keywords to trigger reply, the contribution of this research is to make the dialogue between the reader and the system flexible, which can be achieved through the following three ways: First, use "Wikipedia" because of its quickly updating, diverse vocabulary description, rich in content and continuous growth, etc., which not only assist in the query expansion of the FAQ questions, as a extend word set for the Word Segmentation preprocessing, but also to respond to readers' subject knowledge inquiries. Second, in the processing of FAQ questions, "NLP (Natural Language Processing)" technology is used to help judge the reader's intent to obtain a response that meets their information needs. Third, in terms of defining the granularity of query intent, this study uses the density-based DBSCAN clustering algorithm to group training corpora with similar question structures into the

same intent through parameter adjustment and store them in MongoDB. Then, if it is necessary to add other intentions, the training corpus can also be easily expanded, which is helpful for subsequent maintenance and utilization.

Keyword: Intent-based Chatbot, Natural Language Processing, Wikipedia, Query Expansion, Feature Extraction, DBSCAN



目錄

第一章 緒論	1
1.1 研究背景與動機.....	1
1.2 研究目的.....	2
1.3 研究範圍與限制.....	3
1.4 研究流程.....	3
第二章 文獻探討	6
2.1 聊天機器人.....	6
2.1.1 定義.....	6
2.1.2 聊天機器人的兩種類型.....	7
2.1.3 聊天機器人的兩種架構方法.....	7
2.1.4 聊天機器人的核心元件.....	9
2.1.5 中文聊天機器人在圖書館的應用.....	14
2.2 維基百科.....	18
2.2.1 Wiki 簡介.....	18
2.2.2 維基百科的性質與特點.....	18
2.2.3 維基百科於自然語言處理的相關研究.....	20
第三章 建構意圖式聊天機器人模型	24
3.1 系統架構.....	24
3.2 開發平台與工具.....	26
3.3 演算法.....	27
3.3.1 特徵詞加權演算法：TF-IDF.....	27
3.3.2 Clustering 演算法：DBSCAN.....	28
3.3.3 詞向量演算法：Word2Vec.....	29
3.4 系統建置流程.....	31
3.4.1 訓練語料蒐集與分析.....	31
3.4.2 訓練語料預處理.....	37
3.4.3 LineBot 系統功能建置.....	45

第四章 系統分析與評估..... 65

4.1 評估方法說明.....	65
4.2 測試語料.....	67
4.3 實驗結果.....	70
4.4 實驗分析.....	73
4.4.1 某些排名(Rank)較後的意圖相關、需求滿足程度反而較好.....	73
4.4.2 排名第一(Rank1)卻出現意圖相關最差、需求滿足最好.....	74
4.4.3 某些評價與訓練語料相關的意圖集群卻沒有輸出其系統回覆.....	75
4.4.4 可能影響系統意圖識別結果的因素.....	76

第五章 結論與建議 78

5.1 結論.....	78
5.1.1 系統建置概述.....	78
5.1.2 使用評估結果.....	79
5.1.3 相較於圖資領域其他中文聊天機器人的優勢.....	79
5.2 未來展望.....	80

附錄 82

附錄一 104 間圖書館的常見問答集蒐集名單.....	82
附錄二 30 筆測試資料的系統意圖辨識結果.....	86

中文參考文獻 122

英文參考文獻 123

圖表目錄

圖表 1 線上參考服務機器人(FJCU LIB DIALOGSYS)系統架構圖	4
圖表 2 依據對話能力與回覆機制分類的聊天機器人類別	8
圖表 3 智慧型聊天機器人的整體架構	10
圖表 4 中國清華大學圖書館「小圖」的系統架構	15
圖表 5 台師大圖書館參考諮詢機器人的系統架構	16
圖表 6 一般的內部鏈結(以[[藝術]]為例)	21
圖表 7 管道鏈結(以[[漢語 中文]]為例)	22
圖表 8 基於意圖辨識結果之回應內容生成流程	25
圖表 9 TF (TERM FREQUENCY) 變體	28
圖表 10 WORD2VEC 的兩種模型架構	30
圖表 11 圖書館常見問答集的語料蒐集範例	31
圖表 12 字符(CHARACTER)層次的 BIGRAM 次數統計長條圖	32
圖表 13 字符(CHARACTER)層次的 TRIGRAM 次數統計長條圖	33
圖表 14 「如何」一詞之共現詞分析的次數統計長條圖	34
圖表 15 字詞(WORD)層次的 BIGRAM 次數統計長條圖	34
圖表 16 「查詢」一詞之共現詞分析的次數統計長條圖	35
圖表 17 「資料庫」一詞之共現詞分析的次數統計長條圖	35
圖表 18 整併相似類別與修正問句分類之範例	36
圖表 19 將引號內的詞納入圖資領域慣用語之範例	37
圖表 20 尚未考量 N 字詞的共現關係之斷詞範例	38
圖表 21 使用 TRANSFORMERS 演算法的重新分類結果，以問句「在架上找不到書，該怎麼辦？」 為例	39
圖表 22 使用 K-MEANS 演算法設定分群數目為 80 的分群結果，以問句「在架上找不到書，該怎 麼辦？」為例	40
圖表 23 使用 DBSCAN 演算法設定半徑 0.8、最小資料數目為 2 的分群結果，以問句「在架上找 不到書，該怎麼辦？」為例	41
圖表 24 使用 K-MEANS 和 DBSCAN 演算法的分群結果，以圖書館開放時間相關的問題句為例	42
圖表 25 合併「具有相同意圖但不同問法」集群的範例，以圖書館開放時間相關的問題句為例	44
圖表 26 系統功能模組	45
圖表 27 維基百科的分類架構，以「零食」為例	46
圖表 28 WORD2VEC 模型的相似度計算結果，以「零食」為例	47
圖表 29 經問題解析與查詢詞彙擴展的輸出結果	48
圖表 30 經意圖辨識與排序的候選意圖集群結果	49
圖表 31 九種問題類型的常見疑問詞問法	50
圖表 32 「項目(具體時間)」問題類型之回覆判斷條件	51

圖表 33 寒暑假或其他情況的「項目(具體時間)」問題類型範例.....	52
圖表 34 平日情況的「項目(具體時間)」問題類型範例.....	52
圖表 35 假日情況的「項目(具體時間)」問題類型範例.....	52
圖表 36 疫情情況的「項目(具體時間)」問題類型範例.....	53
圖表 37 「持續時間」問題類型的預約書借閱範例.....	53
圖表 38 「持續時間」問題類型的論文審核範例.....	54
圖表 39 「項目(規則)」問題類型的贈書限制範例.....	55
圖表 40 「項目(規則)」問題類型的續借問題範例.....	55
圖表 41 「項目(規則)」問題類型的大學生借閱資格範例.....	55
圖表 42 「項目(規則)」問題類型的研究生借閱資格範例.....	56
圖表 43 「項目(規則)」問題類型的教師借閱資格範例.....	56
圖表 44 「地點」問題類型的圖書館開館範例.....	57
圖表 45 「地點」問題類型的研究所考古題查找範例.....	57
圖表 46 「地點」問題類型的預約書領取範例.....	57
圖表 47 「數量」問題類型的借書範例.....	58
圖表 48 「原因」問題類型的找不到書籍範例.....	59
圖表 49 「方法」問題類型的找不到書籍範例.....	60
圖表 50 「方法」問題類型的空間申請範例.....	60
圖表 51 「方法」問題類型的證件遺失處理範例.....	60
圖表 52 「是否」問題類型的期刊外借範例.....	61
圖表 53 「是否」問題類型的圖書館飲食範例.....	61
圖表 54 「查詢館藏」意圖集群的訓練語料範例.....	62
圖表 55 「是否」問題類型的查詢館藏範例.....	62
圖表 56 「定義」問題類型的資料庫類型範例.....	63
圖表 57 「查詢維基百科」意圖集群的訓練語料範例.....	63
圖表 58 「定義」問題類型的查詢核酸檢測範例.....	64
圖表 59 「定義」問題類型的查詢 MEME 範例.....	64
圖表 60 測試語料.....	68
圖表 61 測試語料的意圖集群排序結果.....	69
圖表 62 測試語料與「訓練語料」的相關程度評分.....	71
圖表 63 「系統回覆」對於滿足讀者資訊需求的程度.....	72
圖表 64 「系統回覆」對於滿足讀者資訊需求的程度，以測試語料編號 1、8、12 為例.....	73
圖表 65 測試語料與「訓練語料」的相關程度評分，以測試語料編號 1、8、12 為例.....	73
圖表 66 「系統回覆」對於滿足讀者資訊需求的程度，以測試語料編號 22、24、26、27 為例.....	74
圖表 67 測試語料與「訓練語料」的相關程度評分，以測試語料編號 22、24、26、27 為例.....	75
圖表 68 測試語料與「訓練語料」的相關程度評分，以測試語料編號 2、6、13、14、15 為例.....	76
圖表 69 意圖集群編號 38 的訓練語料.....	77



第一章 緒論

本研究探討圖書館參考服務機器人(FJCU Lib DialogSys)的建置與實際應用，旨在基於維基百科(Wikipedia)的基礎上，借助自然語言處理技術來改善問句的預處理與查詢擴展，並且作為解答名詞術語的知識庫使用。

1.1 研究背景與動機

隨著人工智慧的發展，聊天機器人(Chatbot)的相關技術漸趨成熟且應用廣泛。近年來，各大通訊軟體相繼釋出 API，方便開發者將自家產品服務串接至大眾常用的通訊軟體上。不只如此，IBM Watson Assistant、Amazon Lex、Google Dialogflow 等對話平台的出現，讓一般人也能輕易建置，促使 Chatbot 逐漸成為與使用者溝通互動的重要工具。

對於圖書館而言，線上參考諮詢系統很早便被提出並且實際應用，而其本身在降低參考館員的工作量上有一定的助益。早期所建置的線上參考諮詢系統，類似於專家知識系統，多以關鍵語彙對問答資料庫(Q&A Database)進行檢索查詢，再將預存於問答資料庫中對應的答案回覆給讀者。此種模式技術上較為簡單，卻有著極為致命的缺點—僅能處理制式的問題。而如今，伴隨自然語言處理、語音辨識的發展漸趨成熟，以及深度學習技術的突破，促使如同真人般對談的聊天機器人應運而生。

近年來，面臨讀者實際到館人數逐年下降、線上參考諮詢需求逐年提高以及營運總經費與服務人力日益緊縮的大學圖書館，如此現況與困境點出了聊天機器人應用於大學圖書館參考諮詢服務的價值。因應科技輔助之普及，國外的大學圖書館陸續推出聊天機器人的參考諮詢服務應用。例如：2010 年中國清華大學以 A.L.I.C.E. 開源軟體為基礎、XML 為語料庫儲存格式，搭配搜尋引擎技術來建構「小圖」(姚飛、紀磊、張成昱、陳武，2011)；2014 年美國加利福尼亞大學爾灣分校(University of California, Irvine)使用開源軟體 Program-O 開發網頁版聊天機器人「ANTswers」，其藉由 PHP 編寫的 AIML 直譯器透過模式匹配來回覆讀者(Kane, 2016)；2019 年美國奧克拉荷馬大學(University of Oklahoma)則利用 Ivy.ai 對話平台構建網頁版聊

天機器人「Bizzy」(Young, 2019)...等，皆獲得廣大且熱烈的迴響。反觀台灣，近年在圖資領域上的聊天機器人研究才逐漸顯露，像是陳宜琳(2019)自行開發一個使用「模式匹配」方式建構的台師大圖書館參考諮詢機器人、范蔚敏(2020)則在 Google DialogFlow 對話平台的基礎上建置一個台灣大學圖書館虛擬參考諮詢服務機器人。綜上所述，台灣的圖書館在聊天機器人的研究與應用上，仍有極大的推展、探索以及改善空間。

考察國內外的文獻，不難看出將維基百科(Wikipedia)應用於聊天機器人上，已不是全新的想法(Hussain & Athula, 2018; Augello, Pilato, Vassallo & Gaglio, 2009)。然而，在圖資領域中，將維基百科同時作為斷詞程序擴增語彙集、問句查詢擴展以及學科知識庫使用的參考服務機器人之科技較為欠缺，更遑論應用於中文的聊天機器人上了。事實上，維基百科(Wikipedia)的運用，可以不只是作為全球網路上最大且最受大眾歡迎的參考工具書而已，它的各項優勢足以支撐作為本研究在訓練語料蒐集上之相當不錯的材料。舉例來說，由於維基百科(Wikipedia)集合全球群眾智慧進行協作、編輯，因此擁有多國語言、即時更新且持續成長的特性，以及最具規模、資源包羅萬象、敘述風格多樣的優勢。不僅如此，眾人協作的特性，也使得維基百科條目的建立上其詞彙與敘述會更加地符合自然語言，不會流於罕見而艱深的專業術語，有助於作為參考諮詢服務背後的知識庫，方便讀者以自然語言查詢。

1.2 研究目的

根據背景與動機，本研究擬以輔仁大學(輔大)圖書館為應用場域，實際建置一個基於維基百科(Wikipedia)用來解答館務問題和提供學科知識服務的參考諮詢機器人(FJCU Lib DialogSys)，並調查其使用評價，藉以達到輔助輔大圖書館進行即時參考服務的目標。

本研究主要之研究目的有三項：其一，探討維基百科(Wikipedia)於聊天機器人建置的應用與價值。其二，實際建置一個解答館務問題和提供學科知識服務的圖書館參考服務機器人。其三，藉由本研究提出的方法，讓受試者分別針對聊天機器人的原始訓練語料樣本以及系統回覆結果進行評價，透過「原始訓練語料樣本」與讀

者問題之間的相關程度來了解系統能否正確辨識相關的意圖，以及通過評價「系統回覆結果」對於讀者的資訊需求滿足程度來得知系統答覆是否切中讀者需求。

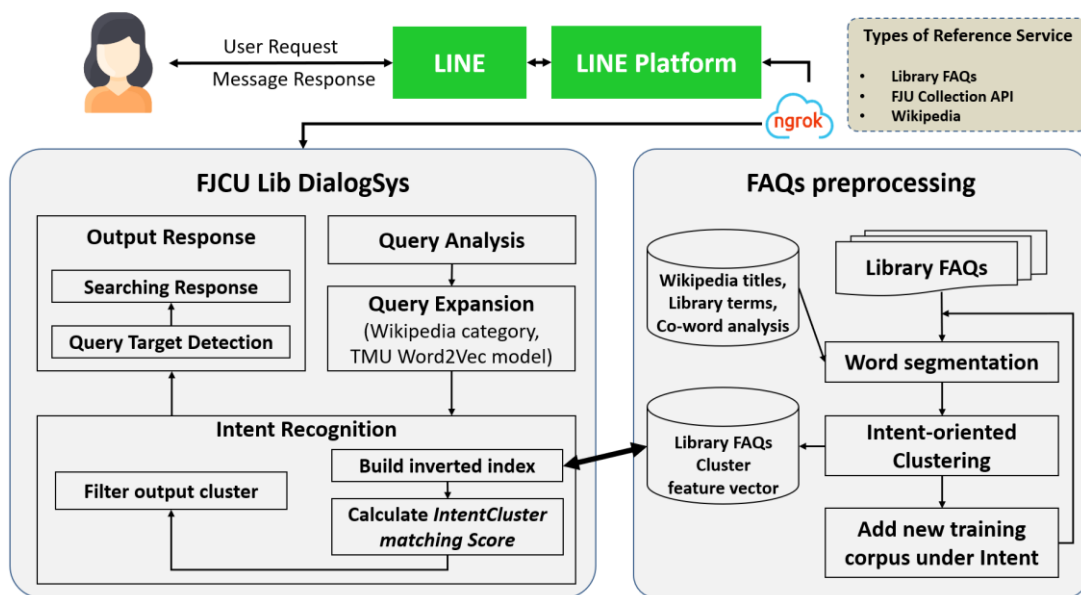
1.3 研究範圍與限制

本研究之範圍與限制，敘述如下：

1. 本研究適用於大專校院圖書館進行參考諮詢服務，其他類型的圖書館可酌予參考。
2. 本研究之圖書館參考諮詢機器人的語料庫資料來源，取自中文維基百科的條目紀錄（截至 2021年 3月 1日）。
3. 本研究之圖書館參考諮詢機器人的答題範圍，著重在日常的常見問題以及事實型的參考問題，不含指引型、研究型或建議型參考問題。
4. 本研究之圖書館參考諮詢機器人的訊息接收與回覆僅包含中文文字訊息，貼圖、音訊並不在範圍內。
5. 本研究之圖書館參考諮詢機器人的使用評估測試對象為輔仁大學之學生及教職員。

1.4 研究流程

本研究的實施分為五大階段：第一階段圖書館線上常見問答集之蒐集；第二階段針對問答訓練語料進行分析和預處理，並依據詢問意圖(Intent)分群；第三階段聊天機器人系統的功能建置；第四階段聊天機器人系統與 LINE 介面整合測試；第五階段實驗設計並蒐集受試者對本圖書館參考諮詢機器人的意圖辨識相關程度、資訊需求滿足程度評價。



圖表 1 線上參考服務機器人(FJCU Lib DialogSys)系統架構圖
(研究者自行整理)

首先，本研究使用 Python 程式語言進行網路爬蟲工作，蒐集全台的大學圖書館、公共圖書館及國家圖書館的線上常見問答集。接著，為了建置一個基於詢問意圖(Intent)作為回應基礎的系統，我們先是對問答語料進行 N-gram、字詞頻率以及共現詞分析，以便對蒐得之語料有個初步的認識。接著，為使系統正確理解讀者的詢問意圖(Intent)，「中文斷詞的準確度」尤其重要，因此在預處理階段透過「維基百科條目」、「圖資領域慣用語」和「高出現頻率的共現 N 字詞」的加入來幫助提升 CkipTagger 中文斷詞的準確性。另外，更應用 TF-IDF 和 DBSCAN 演算法，在 Intent-oriented Clustering 階段為問答訓練語料中具有類似問法結構的問題進行分群(即歸類為同一種意圖集群)，其結果將作為本系統的訓練語料庫，並於後續聊天機器人系統功能建置的 Intent Recognition 階段中使用。

在聊天機器人系統的功能建置上，首先我們把接收到的讀者問題進行 Query Analysis 的預處理動作。當讀者問題中出現訓練語料中所沒有的詞彙時，透過臺北醫學大學所訓練的 Word2Vec 模型以及維基百科的分類架構進行相關或相似字詞的查詢擴展(Query Expansion)。接著，在 Intent Recognition 階段，先是透過反向索引(Inverted index)的建置來為後續的意圖集群匹配加速運算，再將擴展後的

字詞連同中文斷詞結果用於計算與意圖集群之間的相似度分數(IntentCluster Matching Score)，最後篩選出評分在平均以上的意圖集群作為系統輸出回覆的基礎。緊接著是規劃系統的回覆內容，依據這些相關意圖集群的篩選結果，將進一步透過自然語言處理技術擷取查詢目標(Query Target Detection)，並視讀者查詢意圖決定該從何種資源或工具中獲取解答(Searching Response)，像是檢索訓練語料庫來回答館務常見問題、查詢輔大館藏 API 來回應館藏查詢問題、透過 wptools 向維基百科搜尋學科知識問題。一旦本系統所需元件(問句查詢擴展、訓練語料庫、學科知識庫)建置完成，便將之與 LINE 介面整合，同時利用 Ngrok 將本機與外網連接(即架設伺服器)，讓讀者能夠透過 LINE 介面使用本研究所建置的圖書館參考服務機器人—FJCU Lib DialogSys。

關於研究成果的評估，本研究將提出一項評估方法，讓讀者針對本聊天機器人的原始訓練語料樣本以及系統回覆結果進行評估，藉此了解讀者對於意圖辨識的相關程度以及資訊需求的滿足程度，從而得知該聊天機器人能否正確辨識相關的意圖並且其答覆是否切中讀者需求。



第二章 文獻探討

本研究擬建置一個以自然語言為對話基礎，以維基百科為斷詞擴增語彙集、問句查詢擴展及學科知識庫的工具，實現館藏查詢和數位參考服務的圖書館聊天機器人。

實際建置前，我們藉由文獻探討聊天機器人、維基百科的相關理論，並於本章分為兩節做說明，第一節先簡述聊天機器人的定義、兩種類型，並說明建置一個聊天機器人兩種主要的架構方法與核心元件，接著簡述曾建置且中文對話的聊天機器人在圖書館上的應用，以及說明建置中文聊天機器人過程中所遭遇的挑戰。第二節則簡介本研究中作為語料庫、知識庫和查詢擴展的維基百科，闡述其性質與特點，以及其於自然語言處理方面的相關研究。

2.1 聊天機器人

2.1.1 定義

「聊天機器人 (Chatbot)」一詞，又可稱作 conversational agent 或是 dialogue system。根據牛津辭典的解釋，聊天機器人 (Chatbot) 是一個電腦程式，被設計用於模擬與人類對話，特別是應用於網際網路上(Oxford Dictionary, 2019)。

1950 年，人工智慧領域的先驅 Alan Turing 拋出「機器能否思考？(Can machines think?)」的命題，並提出機器是否具有智慧的「圖靈測試 (Turing Test)」此一判斷準則(Turing, 1950)。圖靈測試的概念是：當一個人工智慧系統被誤認為「人類」，亦即該機器在某種程度上重現人類智慧，那麼我們或許就可以說它「具有智慧」。隨後，Joseph Weizenbaum 於 1966 年打造了歷史上第一個聊天機器人—ELIZA (Weizenbaum, 1966)。

今日，聊天機器人被廣泛應用於顧客服務、智慧導購、虛擬助理...等方面。整體而言，當前的發展方向是提高對於使用者任意形式對話的理解能力，甚至掌握使用者的情緒，並能因應使用者的意圖與需求，做出正確的回應並協助使用者完成特定工作。

2.1.2 聊天機器人的兩種類型

以聊天機器人的類型(Type)來說，根據 Al-Zubaide 和 Issa (2011)的分類，主要可分為 Rule-Based chatbot 和 AI-Based chatbot 兩種。

基於規則 (Rule-Based) 的聊天機器人是藉由事先編寫的條件或規則來作為判斷該執行何種回覆動作的依據；而基於人工智慧 (AI-Based) 的聊天機器人，亦可稱為基於意圖 (Intent-Based) 的聊天機器人，則是透過大量對話語料的訓練，以及機器學習、自然語言處理(NLP) 技術的輔助來達到如同人類自我學習並獲得有效訊息的能力，從而能夠自動判斷該回覆何種訊息(Rosruen & Samanchuen, 2018)。

本研究所建置的聊天機器人屬於「基於人工智慧 (AI-Based)」，亦即「基於意圖 (Intent-Based)」的聊天機器人。

2.1.3 聊天機器人的兩種架構方法

聊天機器人發展至今已有相當歷史，Jurafsky & Martin (2019)整理過去聊天機器人所使用之系統架構主要可分為兩種：一種為基於規則(Rule-based)，另一種為基於語料庫(Corpus-based)。而基於語料庫(Corpus-based)聊天機器人，可再細分成資訊檢索型(IR-based Chatbot)，以及編碼解碼型(Encoder decoder Chatbot)。

早期最重要且最具影響力的聊天機器人— ELIZA (Weizenbaum, 1983)，以及目前已知的第一個通過圖靈測試的聊天機器人— PARRY (Colby et al., 1972)，都是屬於「基於規則 (Rule-based)」這類系統架構。基於規則的架構是將使用者的查詢 (query) 去匹配已預先人工定義的規則 (rule)，若有匹配到則產生相應的答案，反之回覆該查詢的解答不存在，以此方式模擬不斷深入話題的對話模式（見圖表 2）。

這種「基於規則 (Rule-based)」的架構方法，雖然實作上非常簡單，只需撰寫規則便能讓 Chatbot 按特定規則執行判斷並回覆準確的結果。然而，若面對較為複雜的任務時，相對地就必須撰寫更多的規則來達到目標，實在耗時。

```

function ELIZA GENERATOR(user sentence) returns response

Find the word w in sentence that has the highest keyword rank
if w exists
    Choose the highest ranked rule r for w that matches sentence
    response ← Apply the transform in r to sentence
    if w = 'my'
        future ← Apply a transformation from the 'memory' rule list to sentence
        Push future onto memory stack
    else (no keyword applies)
        either
            response ← Apply the transform for the NONE keyword to sentence
        or
            response ← Pop the top response from the memory stack
return(response)

```

圖表 2 依據對話能力與回覆機制分類的聊天機器人類別

資料來源：Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing*. Retrieved from <https://web.stanford.edu/~jurafsky/slp3/26.pdf>

另一種，「基於語料庫 (Corpus-based)」是目前最新且最廣泛使用的架構方法，其採用數據驅動的方式，從人與人、人與機器之間的對話語料或是敘述性文檔中透過機器學習方法挖掘出規則，以取代人工耗時的制定，因此亦可稱為「Data-driven chatbot」(Mnasri, 2019)。此一架構方法之下，又可再分為藉由演算法的相似度比對來生成回應的資訊檢索型聊天機器人 (IR-based Chatbot)，以及藉由機器學習將原始輸入直接轉化為目標回應輸出的編碼解碼型聊天機器人 (Encoder decoder Chatbot)。

資訊檢索型聊天機器人 (IR-based Chatbot) 的運作原理與搜尋引擎類似，意即 Chatbot 的用戶輸入可看作相當於搜尋引擎的 query，而系統輸出則相當於搜尋引擎的檢索結果。實作上，使用者輸入會與語料庫文本進行相似度的匹配，以便從語料庫中挑選與使用者輸入最為相似的語句來作為輸出回應，此種方式某種程度上已能反映語料的訓練成果 (Jafarpour et al., 2010; Leuski & Traum, 2011)。在找尋最匹配的 Question-Reply pair 上，餘弦相似度 (Cosine Similarity) 已被廣泛用於 word-level 的向量空間模型 (Banchs and Li, 2012)，並與 TF-IDF (Gandhe & Traum,

2013; Charras et al., 2016) 單詞加權演算法搭配使用。此種相似度匹配的想法，來自良好的回應通常會與前一回合共享單詞或語義，因此若兩者的相似度越高，則表示機器人的回覆越不容易顯得答非所問。

此種基於資訊檢索方式建構的聊天機器人，以檢索資料庫中相應的回覆來回答與主題相關的問題，其優勢在於所回覆的內容不會出現文法錯誤，然而缺點是無法回答超出資料庫範圍的問題。

編碼解碼型聊天機器人 (Encoder decoder Chatbot) 則是把使用者的問題輸入與系統的回覆輸出當作是「預測型」的問題來處理，意即當使用者輸入某一問題，則系統必須預測該輸出何種回覆。最常使用的技術為深度學習的 seq2seq 模型 (Li et al., 2015)。從技術上來說，seq2seq 模型由編碼器 (Encoder) 和解碼器 (Decoder) 組成，編碼器讀取輸入序列並將其轉換為儲存了輸入序列含義的上下文向量 (context vector)，而解碼器則是直接接收並解碼這些上下文向量後再輸出。在訓練學習階段，seq2seq 模型會學習將上下文向量 (context vector) 的隱藏狀態映射到真實的輸出序列，而在推斷預測階段，解碼器將輸出關於學習目標的預測序列 (Mnasri, 2019)。

此種基於 Seq2seq 模型訓練的聊天機器人，Vinyals & Le (2015) 表明雖然系統簡單、能夠自動且智慧地生成相應的回覆，但仍有相當多的限制與挑戰(如：存在語法錯誤、不容易訓練的缺陷)，目前尚且無法通過圖靈測試。

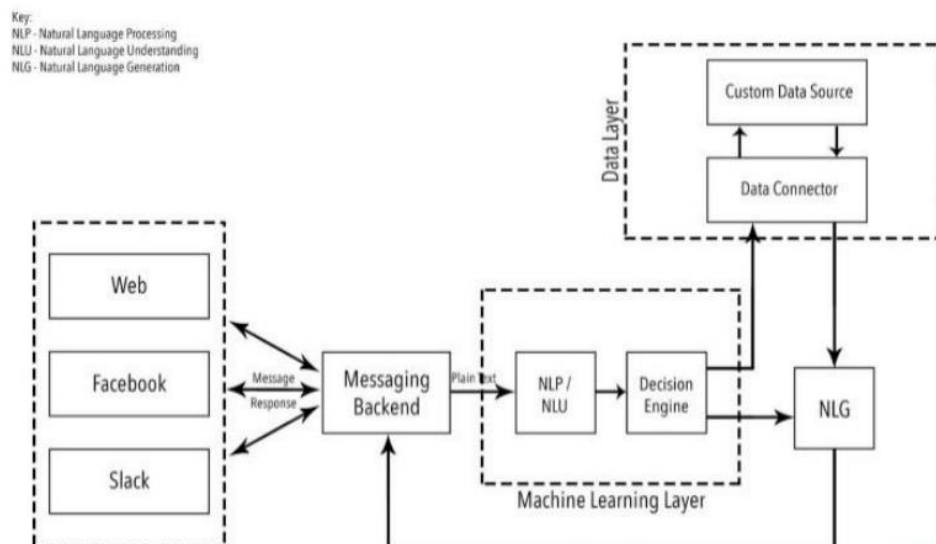
故而，本研究採用「基於語料庫 (Corpus-based)」的資訊檢索型聊天機器人 (IR-based Chatbot) 來作為圖書館參考諮詢機器人的建置框架。

2.1.4 聊天機器人的核心元件

一般而言，若要建置一個智慧型的聊天機器人，其整體架構如圖表 3 所示。大致上，依據使用者訊息傳遞的處理順序可能會需要 ASR (自動語音辨識)、NLU (自然語言理解) 和 NLG (自然語言生成) 這三項核心元件。其中，NLU 和 NLG 皆建立在自然語言處理 (Natural Language Processing, NLP) 技術的基礎上。另外，假使要讓一個聊天機器人系統更具智慧化，能夠很大程度地模擬人類回答問

題，必須讓它具備足夠的知識，而這仰賴「外部知識庫（相當於圖表 3 中 Data Layer 的區塊）」的導入才有可能實現。

自動語音辨識 (Automatic Speech Recognition, ASR) 負責將使用者的語音訊號轉換為文字輸入，使得使用者無須操作鍵盤輸入，即可立即對聊天機器人透過語音下指令。接著是自然語言理解 (Natural Language Understanding, NLU)，它負責處理非結構化的文字資料，並將之轉換為系統可以理解並處理的形式。相較於 NLG，NLU 所要處理的部分更加地複雜且費力，這是因為自然語言具有相當豐富的結構與表達方式的緣故。最後，自然語言生成 (Natural Language Generation, NLG) 則是負責規劃輸出給使用者的回覆訊息。



圖表 3 智慧型聊天機器人的整體架構

資料來源：Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020, March). A smart chatbot architecture based NLP and machine learning for health care assistance. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security* (pp. 1-6).

談及 NLU 和 NLG 兩項核心元件皆以自然語言處理 (Natural Language Processing, NLP) 技術為基礎，得知在聊天機器人的建構上 NLP 尤為重要，故特別為 NLP 稍加說明。

2.1.4.1 自然語言處理

自然語言處理 (Natural Language Processing, NLP) 是對人類語言文字使用斷詞、過濾停用詞、詞幹提取和同義詞搜索...等自動化處理的技術，使電腦從自然語言中提取有意義訊息的過程 (Warschauer & Healey, 1998)。其終極目標為實現以自然語言為媒介，讓電腦與人類能夠順暢的溝通，並完成各項指定的任務 (曾元顯，2012)。

自然語言處理的運作機制又可分為自然語言理解 (Natural Language Understanding, NLU) 和自然語言生成 (Natural Language Generation, NLG)。

(一) 自然語言理解 (NLU)

在聊天機器人系統中，自然語言理解模組作為理解使用者輸入之意圖的解析器，負責「處理使用者的輸入」，提取從使用者輸入中獲得的有意義訊息，或是需要進一步查詢的細節知識，並以此協助後續判斷使用者的詢問意圖，以及採取相應的動作與訊息回應。

而在進行後續動作之前，首先必須執行自然語言的預處理工作，步驟依序為斷詞、過濾停用詞、詞幹提取和同義詞搜索，詳述如下：(Al-Zubaide & Issa, 2011)

1. 斷詞 (Tokenization)

「斷詞」是自然語言處理中相當重要的第一步，它指的是以有意義的字詞為最小單位將使用者的輸入進行拆分。而拆分的依據會隨每種語言的特性而有所不同，例如：英文能夠直接使用空格、標點符號等分隔符號進行字詞拆分；中文則需要借助詞典中收錄的詞，才能降低字詞產生錯誤切分的機會。

2. 過濾停用詞 (Stopper Filtering)

「過濾停用詞」是指去除經由斷詞後所得的常用功能詞，包含限定詞、介係詞、連接詞...等，或是出現次數頻繁但本身並無太大意義的字詞，例如：關於 (About)、他們 (Them)...等。

3. 詞幹提取 (Word Stemming)

「詞幹提取」通常用於具有詞性或時態變化的語言中，像是英文、法文，此一步驟會透過常見的 Porter Stemmer 詞幹提取演算法將同一字詞的各種型態變化視為是同一類，有助於字詞的查詢擴展，亦可被認為是一種搜尋上的容錯技術。

4. 同義詞搜索 (Finding Synonyms/alternatives)

「同義詞搜索」有助於字詞的查詢擴展。由於使用者的表述可能不夠明確，或是其查詢詞與知識庫的描述不一致，而導致系統無法找到相符的匹配。因此，藉由搜尋其同義詞，能夠達到提高召回率 (Recall) 的目的。

除了進行上述的自然語言預處理工作，在理解每個單詞的特性與結構上還需要注意四種自然語言的歧異 (Gill, 2019; Riaz et al., 2019)：

1. 詞彙歧義 (Lexical Ambiguity)

指當一個單詞具有多重含義時所造成的歧義。例如：一個單詞具有相同拼寫及發音卻有不同含義之「同形異義的歧義 (Homonymy Ambiguity)」、具有相同發音但拼寫及含義皆不同的多個單詞之「同音異義的歧義 (Homophony Ambiguity)」(Klepousniotou, 2002)。

2. 句法歧義 (Syntactic Ambiguity)

指當一個語句能夠被解析成多個具有不同含義的語法結構時所造成的歧義，亦即因句法解析的不同而產生多重解釋的歧義。

3. 語義歧義 (Semantic Ambiguity)

指的是一個句子在相同的句法結構下產生不同解讀的歧義，這可能是因為語氣或上下文的緣故而造成歧異的現象。

4. 對應歧義 (Anaphoric Ambiguity)

在文章中經常出現使用短語或單詞來代替前面已出現過的事物，而對應歧義指的便是在後面句子中所代指之事物的不同而造成不同含義的歧異。

(二) 自然語言生成 (NLG)

在聊天機器人系統中，自然語言生成模組負責「自動生成使用者可讀的輸出」，經過 Decision Engine 的意圖判斷後，選擇接下來所要執行的動作，而在接收來自外部知識庫傳回的結構化數據 (即解答) 後，進一步生成最終給使用者的可讀文字訊息。

NLG 的過程共分為文本規劃 (Text Planning)、語句規劃 (Sentence Planning) 以及語言實現 (Linguistic Realization) 三個階段 (Gill, 2019; Rambow et al., 2001; Reiter & Dale, 1997)。首先，「文本規劃 (Text Planning)」是指決定所欲生成的文本應該置入哪些從外部知識庫回傳的結構化數據，以完成最後輸出文本的基礎內容規劃。接著，「語句規劃 (Sentence Planning)」指的是將多個結構化資訊進行合併，並組織成一個句子，來表達所要呈現的完整訊息。最後，「語言實現 (Linguistic Realization)」：則是對第二階段生成的語句進行文法修正，從而產生語法通順的最終結果。

關於 NLG 生成可讀文字訊息的方法共有三種，詳述如下：(Chen et al., 2017)

1. 神經生成模型 (Neural Generative Models)

如今，大量的對話語料多半從社群網站中取得，並將之作為模型訓練的材料。神經生成模型使用基於深度學習的神經機器翻譯 (Neural Machine Translation) 技術，藉由對話語料的訓練和使用者的輸入直接重寫出切合其需求的回答。神經生成模型以序列到序列 (Sequence-to-Sequence Models, Seq2Seq) 的編碼-解碼生成架構 (Encoder-Decoder structure) 為基礎，其核心概念為利用遞歸神經網路 (Recurrent Neural Network, RNN) 來學習一個輸入序列的所有資訊，將之編碼並濃縮至一個向量中，再利用另一個遞歸神經網路來將此資訊解碼以生成輸出序列 (Hsieh et al., 2016)。此種生成回應的方式不依賴知識庫中預先定義的回應，能夠很大程度地模擬人類的對話交流。然而，一個好的模型並不好訓練，而且生成的回應容易有語法錯誤的問題。

2. 檢索式方法 (Retrieval-based Methods)

檢索式方法是一種訊息-回應匹配 (Message-Response Matching) 的方式，從已預先定義回應的知識庫中利用比對演算法來選擇合適的答案作為輸出，此種生成回應方式較為簡單，且不會產生語法錯誤以及任何新文本。而所需注意的是「比對演算法 (Matching algorithm)」要能克服輸入訊息與輸出回應之間的語意間隙 (semantic gap) (Hu et al., 2014)。另外，對於知識庫中尚未定義的問題或是需要根據對話上下文進行反應的情況，可能就無法處理。

3. 混合式方法 (Hybrid Methods)

「混合式方法」結合了神經生成模型 (Neural Generative Models) 與檢索式方法 (Retrieval-based Methods) 兩種回應生成方式，對於聊天機器人的表現具有顯著提升的效果 (Qiu et al., 2017; Song et al., 2016)。此種方法改善「檢索式方法」容易給予精確但過於生硬的答覆，以及解決「神經生成模型」給予流暢卻無意義資訊的回應缺點。

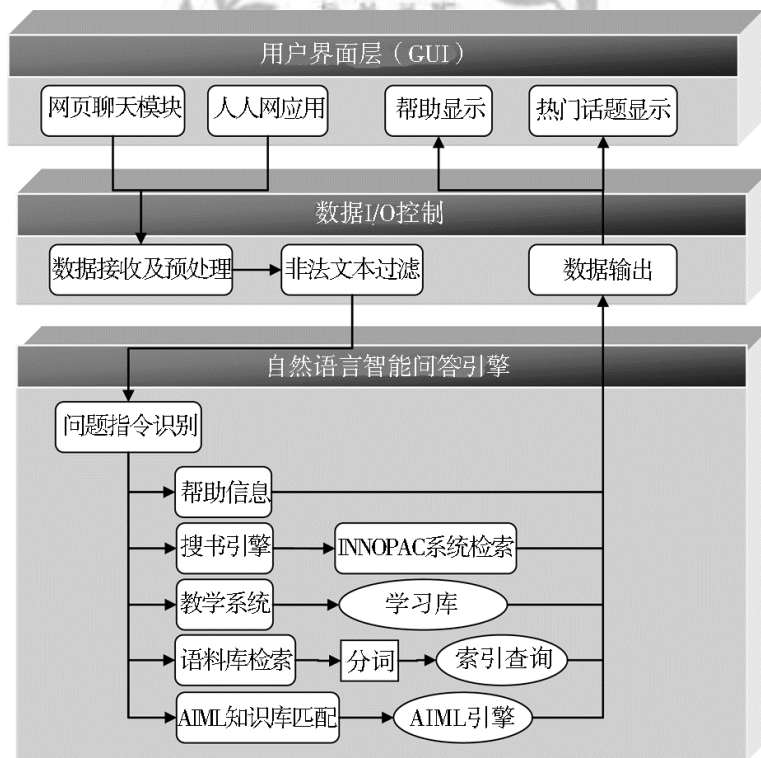
總而言之，若想要達到彈性地根據讀者詢問意圖進行回應，則必須借助自然語言處理技術的配合，才能讓一個聊天機器人如同人類般對話，而不是一個只會制式關鍵字匹配的單純程式。

2.1.5 中文聊天機器人在圖書館的應用

因應科技輔助的普及，使得圖書館開始導入聊天機器人來解決讀者的常見問題。近幾年，國內外的大學圖書館陸續推出聊天機器人的參考諮詢服務應用，其中不乏基於中文的圖書館參考諮詢服務機器人。

在圖資領域上，已知基於中文的聊天機器人包含中國清華大學圖書館的「小圖」(姚飛、紀磊、張成昱、陳武，2011)、台師大圖書館參考諮詢機器人(陳宜琳，2019) 以及台灣大學圖書館虛擬參考諮詢服務機器人(范蔚敏，2020)，以下簡述各聊天機器人的建置架構：

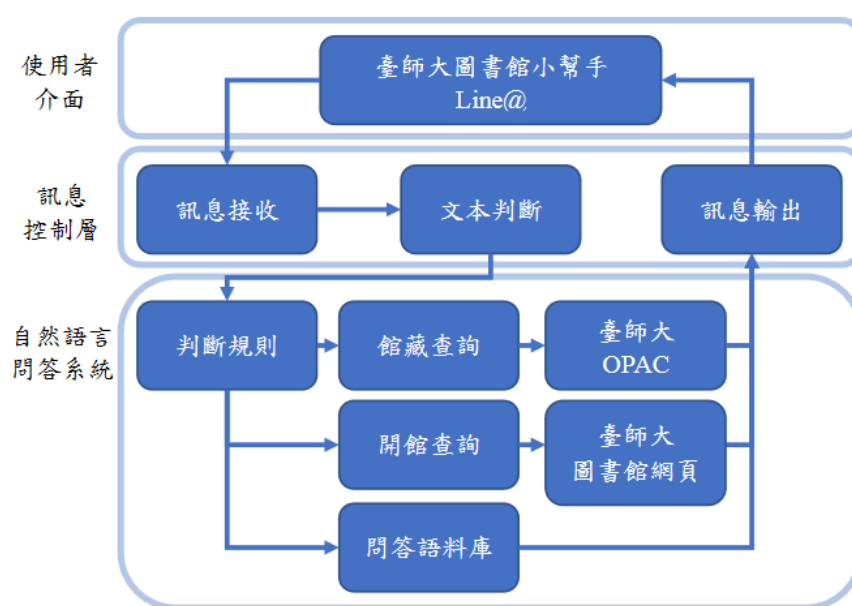
2010 年設計並以開源軟體 A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) 為基礎建置的「小圖」，提供參考諮詢、圖書檢索、自我學習...等多種服務，而其所使用的關鍵技術包含了自然語言處理(如：中文分詞)、搜尋引擎技術(如：設置語料庫、建立倒排索引、以 TF-IDF 演算法進行相似度匹配)、教學系統以及知識推理機制。當使用者輸入其問題指令後，小圖會先進行問題指令的識別，並導引到相應的服務功能，再行處理和檢索。例如：當使用者輸入「Q: 問題 A:答案」的句型時，系統將判斷使用者想要教育小圖，並將教學指令添加到學習庫中；當使用者想要搜尋圖書時，則會導引至 INNOPAC 系統檢索；當使用者想要詢問一些常見問題時，系統會先進行中文分詞，並以搜尋引擎的檢索方式查詢；而當無法在語料庫中找到適當答案時，小圖會依靠模式匹配和推理機制從儲存了 4 萬多條知識分類的 AIML 知識庫中取得推理後的答案(見圖表 4)。



圖表 4 中國清華大學圖書館「小圖」的系統架構

資料來源：姚飛、紀磊、張成昱、陳武（2011）。實時虛擬參考諮詢服務新嘗試—清華大學圖書館智能聊天機器人。現代圖書情報技術，27(4)，77-81。

2019年由陳宜琳建置的台師大圖書館參考諮詢機器人，提供了館藏查詢、開館查詢與問答語料庫回覆三大功能，使用的語料來自圖書館參考紀錄與常見問答集。在系統架構上，主要分為三層：使用者介面、訊息控制層以及自然語言問答系統。當文字訊息送至自然語言問答系統時，經由人工撰寫的「模式匹配」規則判斷後，依據判斷結果送入相應的機器人功能模組中，接著再與問答語料庫問題相似度比對、經排序後返回最適答案。



圖表 5 台師大圖書館參考諮詢機器人的系統架構

資料來源：陳宜琳（2019）。國立臺灣師範大學圖書館參考諮詢機器人建置與評估（碩士論文）。取自華藝線上圖書館系統。doi:10.6345/NTNU201900975。

2020年，范蔚敏使用 Google DialogFlow 對話平台建置台灣大學圖書館虛擬參考諮詢服務機器人。此一機器人的知識庫來源選用大學圖書館參考服務部落格的文章內容，透過擷取文章內容關鍵字與資源連結來設計聊天機器人對話內容，以及建立同義字字典 (Entities) 與問答對話組 (Intents)。使用上，經過語料訓練後的系統會自動判斷問題的 Intents，並將預先設計的對話內容輸出給使用者。

2.1.5.1 建置過程面臨的挑戰

中文聊天機器人的建置，大多面臨以下問題 (姚飛、紀磊、張成昱、陳武，2011；陳宜琳，2019；范蔚敏，2020)：

1. 語言特性上的差異，使得漢語相較於印歐語系的語言來說，語序更加自由且單詞之間無分隔、虛詞運用較多，無形中增加自然語言處理的困難度。
2. 相較用於處理英文的工具與資源，中文相關的自然語言處理工具和資源十分缺乏。
3. 整理並建置知識庫/語料庫或設計對話內容需要耗費大量人力進行，因此如何有效率地建立是個重要議題。
4. 當資料量大幅增加時，如何兼具擴充性以及運作效能。
5. 關於資源指引查詢的問題，無法告訴使用者直接的答案。例如：「圖書館有無訂天下雜誌？」，機器人只能回答使用者如何利用圖書館的查詢系統進行查訊，無法直覺回覆「有」或「沒有」。
6. 以關鍵詞作為規則判斷所建立的簡易分類器，無法聚焦問題的意圖，將使得判斷結果失焦。也許可以多增加判斷規則，或是以答案類型(answer type)做分類，發展自動分類模組，讓問題比對更加精確。

2.2 維基百科

為使聊天機器人能夠達到如同人類對話般的互動效果，並回答學科知識之問題，本研究欲借助自然語言處理技術於維基百科 (Wikipedia) 上的應用，進行機器人系統的對話語料訓練以及知識庫建置。因此，此一小節將對維基百科進行文獻探討，從 Web 2.0 的交互式網站 Wiki 開始簡介，接著說明維基百科及其特點，最終探討維基百科於自然語言處理的相關研究。

2.2.1 Wiki 簡介

Web 2.0 的浪潮，促使網際網路相關應用的推出，其中一項便是「Wiki」。Wiki 是一種交互式網站，允許任何使用者隨意增添、編輯、移動網頁內容。不僅如此，由於這種網站在網頁生成上相較於靜態網頁更為簡單、增修及發佈頁面的代價小且允許大量網頁相互參照串連，因此在公共服務領域上獲得廣泛的運用 (郭鴻昌，2008)。

其中一項最成功的 Wiki 應用，當屬「維基百科 (Wikipedia)」，其源自於 2001 年 1 月正式展開的網路百科全書計畫。Wikipedia 一詞是由交互式網站系統的「Wiki」和具有百科全書之意的「encyclopedia」所組成的混成詞，中文譯作「維基百科」("Wikipedia", 2019)，也是目前全球網路上最大且最受大眾歡迎的參考工具書。

2.2.2 維基百科的性質與特點

作為一個多語言、內容自由、集合群眾智慧的網路百科全書協作計畫 — 維基百科 (Wikipedia)，其特點整理如下 (Giles, 2005；陸曉曦，2009；郭鴻昌，2008)：

1. 自由開放：社群成員可以不受限制地使用、編輯以及公開發佈。正因為如此，給予協作者們充分的交流、討論，促成條目內容的合理化，並使其內容描述與詞彙選擇更符合自然語言且易於普羅大眾理解，不流於罕見而艱深的專業術語。

2. 協同創作：為社群提供協作式寫作的交流工具，能夠集合群眾智慧進行網頁內容的創建、編輯。
3. 多國語言：自 2001 年以來，截至 2021 年 5 月，維基百科的文章已被超過 301 種語言所創建。其中，中文維基百科已存在至少 1,199,520 篇條目。
4. 準確性：作為一個人人都能參與修改條目的協作平台，維基百科的權威性和準確性受到質疑與挑戰。然而，根據 Giles 在 2005 年對維基百科 (Wikipedia) 與大英百科全書 (Encyclopedia Britannica) 進行內容準確性的研究調查，從隨機挑選的 42 個條目中，可知維基百科平均有 4 個錯誤，大英百科全書則有 3 個，其結果指出兩者的準確性差異不大。
5. 維護快捷：快速創建、存取和即時更新超文本頁面的優勢，彌補了維基百科在準確性的瑕疵。
6. 可增長性：藉由為新增的條目創建頁面連結，得以使系統持續成長，維持資訊的知識性與新穎性。
7. 紀錄修訂歷史：維基百科保留頁面內容的每一次修訂紀錄，以及系統內頁面的變動歷史。
8. 強關聯性的知識鏈結網路：「條目」是維基百科的最基本組成單位，並以節點 (node) 及鏈結 (link) 作為概念的主軸，透過協作者將相關或相似的主題通過鏈結 (link) 的方式來相互引用與描述，使得知識的組織自然形成且漸趨有序化。因此，在知識的鏈結上，除了呈現非線性的網狀結構外，還具有強烈的關聯性 (黃菁、楊帆，2009)。
9. 統一的結構：集合群眾協作而成的維基百科，為了有效創建、快速查找以及方便使用資料庫條目，因此對內容及格式進行規範，並使用本身提供的格式化語法 — Wiki 標記式語言 (Wiki markup language)，以便協作者專注於撰寫條目內容。

10. 涵蓋知識範疇廣：維基百科中所創建的條目不限領域，只要條目具有完備的定義、內容，以及可靠的參考文獻，便能收錄。因此，其收錄的主題條目多元、具規模且包羅萬象。

2.2.3 維基百科於自然語言處理的相關研究

為了讓電腦能夠處理自然語言，首先最基本的便是使其學會「斷詞」和「理解詞的意思」，而後才能進一步「解析句子」，完成以句子為基本單位之語法及語義的自動解析，達到擁有理解或運用人類語言的能力。

在自然語言處理領域上，「詞義辨識 (Word Sense Disambiguation, WSD)」是一個相當重要、長期存在且尚未解決的問題。由於自然語言的歧義性，不可能完全透過語言學家人工訂定的規則來涵蓋所有語言的特性。因此，現今的作法是藉由建立機器學習 (Machine Learning) 的演算法模型，來讓電腦學會從訓練的資料中，自動歸納出語言的特定模式和趨勢，並藉此來解決一個歧義詞在不同語境中分別對應的語義，亦即根據不同的上下文辨識該歧義詞正確的意思。

「詞義辨識 (Word Sense Disambiguation, WSD)」的方法一般可分為三種 (Fogarolli, 2009)。第一種是基於知識庫的詞義消歧方法 (Knowledge-Based Disambiguation Approach)，使用辭典等外部語言資源作為外部知識庫，並對產生歧義的上下文進行分析。第二種是監督式詞義消歧方法 (Supervised Disambiguation Approach)，仰賴從人工標注的訓練語料中進行特徵提取，再把萃取出來的特徵用於分類器的分類上。最後則是非監督式詞義消歧方法 (Unsupervised Disambiguation Approach)，此方法直接從原始、未標注的語料中找尋目標詞彙的上下文特徵，像是共現的情況或特徵資訊，來分類目標詞彙之詞義 (Chen et al., 2005)。

根據 Kamps and Koolen (2008) 關於維基百科之鏈結結構的研究，發現鏈結 (Link) 可以是相關性的指標，除了能幫助定義語義關係，還有助於提高資訊檢索的效能。Fogarolli (2009) 的研究中，維基百科不只作為知識的來源，當頁面之間存在雙向的「強鏈結 (Strong Links)」結構時，表示這兩個頁面具有強烈且相互參

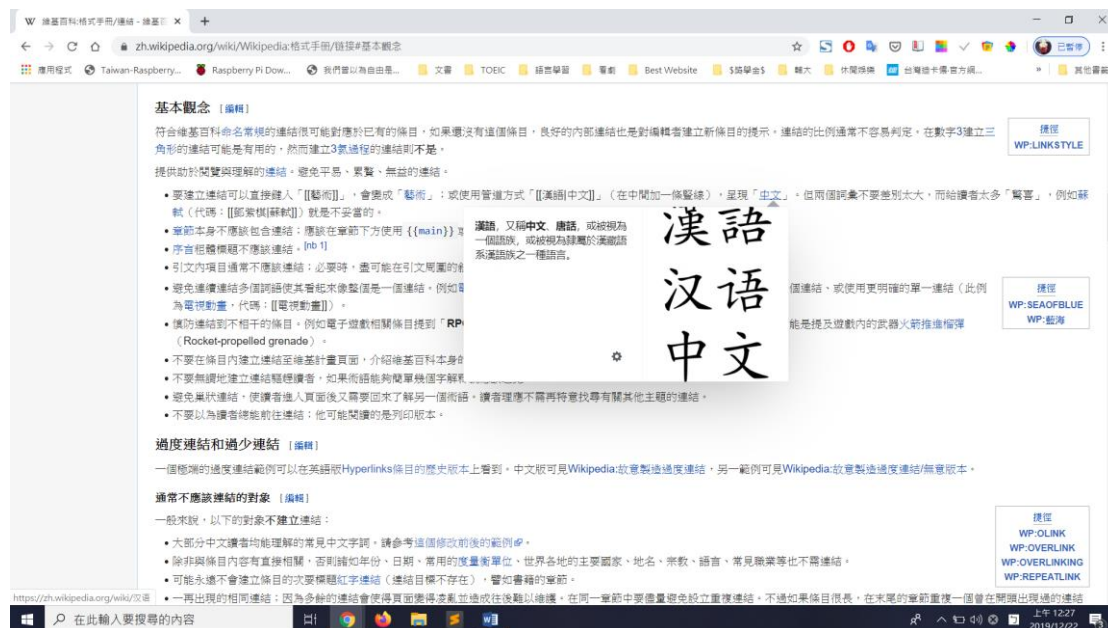
照的關聯性，經研究證實有利於找出兩者之間的重要關係，確定語義上的相關性，甚至能識別主題概念。舉例來說，維基百科的「宗教」條目內具有指向「神祇」的鏈結，而「神祇」條目裡亦有指回起始頁面的「宗教」鏈結，因此可以推斷兩者概念之間具有強相關性。

Mihalcea (2007) 的研究中，使用維基百科條目的內部鏈結作為詞義辨識之詞義標注的來源，而內部鏈結可分為一般和管道 (piped links) 兩種方式呈現，舉例如「[[藝術]]」以及「[[漢語|中文]]」，如圖表 6 和圖表 7 所示。透過使用內部鏈結將相應的維基百科條目作連接，可看作將相應的概念作有意義的解釋，藉此建構帶有詞義標記的訓練語料，不只解決語言歧義的問題，更突破生成帶有語義標注資料之質與量的瓶頸，同時也能為多種語言生成帶有語義標注資料。而在此研究的詞義消歧實驗中，更是證明基於維基百科的詞義標注是可靠的，其有助於建構準確的詞義分類器。



圖表 6 一般的內部鏈結(以[[藝術]]為例)

資料來源：維基百科 (2019)。維基百科：格式手冊/連結。取自 <https://zh.wikipedia.org/wiki/Wikipedia:%E6%A0%BC%E5%BC%8F%E6%89%8B%E5%86%8C/%E9%93%BE%E6%8E%A5#%E5%9F%BA%E6%9C%AC%E8%A7%80%E5%BF%B5>



圖表 7 管道鏈結(以[[漢語|中文]]為例)

資料來源：維基百科 (2019)。維基百科：格式手冊/連結。取自 <https://zh.wikipedia.org/wiki/Wikipedia:%E6%A0%BC%E5%BC%8F%E6%89%8B%E5%86%8C/%E9%93%BE%E6%8E%A5#%E5%9F%BA%E6%9C%AC%E8%A7%80%E5%BF%B5>

維基百科所扮演的角色，可不只作為自然語言處理的斷詞擴增語彙集而已，它更是個包羅萬象的知識庫。然而，使用維基百科作為回答參考問題的知識庫之前，首先面臨的是如何將維基百科的分類條目轉化為具有語義關係的概念網路形式，歸納出含有語義關聯的關係（例如 IsA 關係），從而使機器能夠大規模自動處理知識概念的層次關係。

在 Ponzetto & Strube (2007)、Nastase & Strube (2008) 和 Ponzetto & Navigli (2009) 的研究中，利用維基百科的分類 (Category) 系統得知父類別與子類別之間的相關性及概念上的層次關係，並用以建構知識概念的語意網路。其中，Ponzetto & Strube (2007) 提出了從維基百科的分類條目中推導出「IsA 關係」的四種處理方法，其方法詳述如下：

1. 基於句法結構的方法 (Syntax-based methods)

此方法是先使用 Stanford parser 將父類別與子類別的名稱進行句法分析，再把擁有相同的句法結構的片段進行字串匹配，若兩者共享相同類別的字串，則其

類別間的關係可標記為 isa (Klein & Manning, 2003)。例如：BRITISH COMPUTER SCIENTISTS 隸屬於 (isa) COMPUTER SCIENTISTS。

2. 基於連結性的方法 (Connectivity-based methods)

此方法仰賴分類的網路結構及其連結性，從實例的概念描述中識別兩者於現實世界中的關係。當一個頁面其條目名稱為複數時，將其與相對應的類別視為 isa 關係。例如：ALBERT EINSTEIN 隸屬於 (isa) NATURALIZED CITIZENS OF THE UNITED STATES 類別之下。

3. 基於詞彙-句法結構的方法 (Lexico-syntactic based methods)

當類別名稱經由命名實體識別 (Named Entity Recognition) 和句法結構分析後，符合 Y's X、Y with X、Y such as X...等任一種模式，便可將之視為 isa 關係，意即 X isa Y。

4. 基於語義推論的方法 (Inference-based methods)

該方法應用推理的方式，處理類別概念的語義層次關係。若 X isa Y 且 Y isa Z，則經推論可得 X isa Z，意即 X 類別隸屬於 (isa) Z 類別。

綜上所述，維基百科於自然語言處理的相關研究中，不只利用其鏈結結構所產生的強相關性來區別兩概念間語義上的關聯，還藉由使用管道鏈結 (piped links) 的表示方式來處理自然語言上詞義消歧 (Word Sense Disambiguation, WSD) 的問題。不僅如此，藉由維基百科的分類 (Category) 系統所歸納而得的語義關係 (例如 IsA 關係)，有助於機器大規模地建構具有層次關係的知識鏈結網路，以回答較複雜的參考問題。

針對維基百科(Wikipedia)具備的多種特性，本研究將之作為斷詞(Word Segmentation)的訓練語料、解答名詞術語的學科知識庫，並且應用其分類架構所產生的關聯於問句的查詢擴展(Query Expansion)上，幫助系統進行詞彙理解、同義詞關聯以及知識檢索。

第三章 建構意圖式聊天機器人模型

本研究探討圖書館參考諮詢機器人的建置並加以評估，藉由維基百科 (Wikipedia) 以及自然語言處理技術的應用，期望能夠讓讀者以「自然語言」查詢館務相關問題或是獲取學科方面的知識。

本章首先介紹本文建構之圖書館參考諮詢機器人的整體系統架構，而後依序說明建置此一聊天機器人所使用的開發平台與工具、演算法以及系統建置流程。其中，系統建置流程分為三大階段闡述，分別為訓練語料蒐集與分析、訓練語料預處理以及 LineBot 系統功能建置。

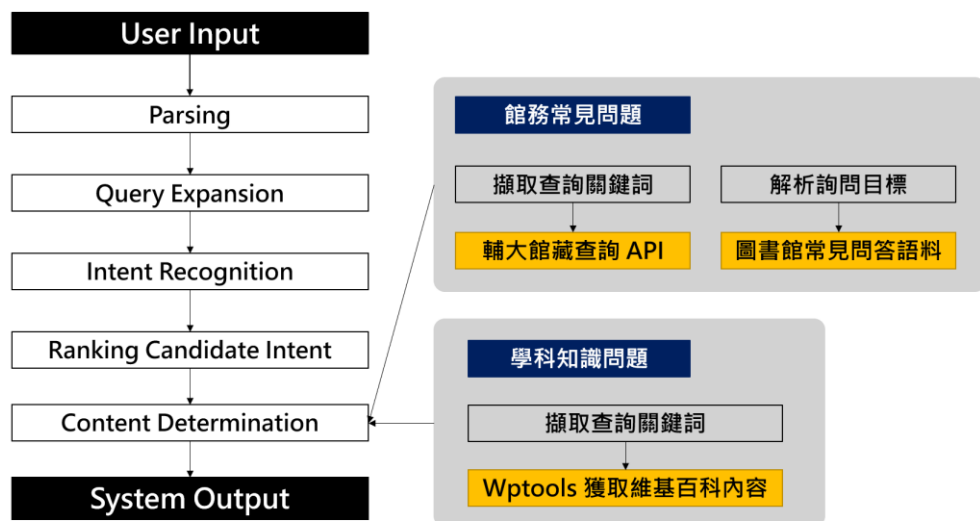
3.1 系統架構

本研究建置一個基於詢問意圖(intent-based)來回應讀者問題的圖書館參考諮詢機器人，此種方式建置的聊天機器人相較單純以「關鍵詞」來匹配，或是預設制定「樣板式」問句的問法，更能夠由系統彈性地判斷讀者問句的真正意思，使得問句中不必出現特定的關鍵詞或問法樣板，便能匹配到相應的內容作為系統答覆。

為建置「基於詢問意圖(intent-based)」的聊天機器人，其基礎必須建立在對問題的準確理解上，除了要能夠判斷讀者的詢問目的，還要能從句子中擷取重要訊息，因此自然語言處理(Natural Language Processing, NLP)技術以及蒐集大量的問答訓練語料是不可或缺的(Rosruen & Samanchuen, 2018)。

本圖書館參考諮詢機器人的系統架構，如圖表 8 所示。當系統接收到由 LINE 介面傳送來的讀者問題後，將會先對問題進行解析(Parsing)。而解析(Parsing)的部分包含了文本斷詞(Word Segmentation)、詞性標註(Part-of-speech Tagging)、命名實體辨識(Named Entity Recognition, NER)。藉由上述的初步處理程序，我們得到了被切割成一個一個最小詞彙單位(Token)的句子。有鑑於讀者問題中使用的詞彙可能不在所蒐集訓練語料的範疇內，如此可能導致與意圖集群的匹配結果不佳，對此系統針對訓練語料中的高頻詞彙進行詞彙擴展(Query

Expansion)。接著，在「意圖辨識 (Intent Recognition)」此一階段，系統為讀者問題進行「意圖相似度」匹配，並依相似度高低排序找出最有可能的意圖集群。最終，根據所歸屬的意圖集群進一步使用「自然語言處理技術」擷取詢問目標，並使用對應的查詢工具(輔大館藏查詢 API、圖書館常見問答語料和 Wptools 套件)檢索取得內容來輸出，例如：當讀者詢問的是某館藏是否存在於輔大圖書館時，這時系統就會擷取館藏的書名，並使用輔大館藏查詢 API 來查找；當讀者詢問的是查找館藏以外的館務常見問題時，像是圖書館的開放時間、研究生的借書資格、書籍不在架上...等問題，系統便會依據讀者詢問的精細程度再進一步地解析其詢問目標，並回傳與其詢問目標相關的解答作為回覆；而當讀者詢問的是學科知識問題時，則會擷取學科術語或專有名詞，再透過 Wptools 的 Python 套件來獲取維基百科條目的內容。



圖表 8 基於意圖辨識結果之回應內容生成流程
(研究者自行整理)

3.2 開發平台與工具

一、LINE Bot

在台灣，LINE 一直是最受國人歡迎且普遍使用的通訊軟體。透過其推出的 Messaging API (訊息應用程式介面)，開發者能夠更方便快速地讓訊息在 LINE Platform 和自家的服務系統(Bot server)之間進行傳遞，以建置一個特定應用場景的聊天機器人。

本研究所建置的圖書館參考諮詢機器人，其後端伺服器是以 Django 框架搭建而成，用來接收 LINE 介面的使用者訊息(request)，並進行邏輯運算處理。接著，處理完的結果會透過 Ngrok 所建立的內網伺服器與外網(網際網路)溝通通道，傳送至 LINE 介面讓使用者讀取系統回覆的訊息。

二、Python 函式庫：Beautiful Soup 和 Wptools

Python 是一種高階程式語言，以高可讀性和簡潔語法為特色，擁有龐大且豐富的第三方函式庫，以及跨平台支援的優勢。它被廣泛用於 Web 開發、人工智慧和數據分析與處理...等領域的應用上。

本系統是以 Python 編寫而成，並採用 Beautiful Soup 和 Wptools 第三方函式庫分別進行網路資料爬蟲以及擷取維基百科內容資料的工作。Beautiful Soup 此一個函式庫模組能夠讓使用者方便且快速地解析網頁，並提取出 HTML 或 XML 標籤中的內容；Wptools 則是藉由對 MediaWiki API 發出 request 來獲取維基百科的條目、摘要、資訊框內容(Infobox data)、Wikidata、分類(Category) ...等資料。

三、中文自然語言處理工具：CkipTagger

CkipTagger 是以 Python 建構的一個函式庫模組，由中研院的中文詞知識庫小組(CKIP Lab) 於 2019 年釋出，是一套結合斷詞(Word Segmentation)、詞性標注(Part of Speech Tagging)、實體辨識(Named Entity Recognition) 的一個國產開源套

件，不只支援繁體中文斷詞及不限長度的句子，還允許使用者針對個人需求自由制定參考/強制詞典的功能，更以超越 Jeiba 的斷詞準確度為特色¹。

四、MongoDB

MongoDB 是一種非關聯式資料庫(NoSQL)，以鍵—值對(key-value pair)的形式儲存每一筆資料。相較於傳統的關聯式資料庫(RDBMS)，NoSQL 具有彈性、易於擴展的優點，因此被廣泛應用於大規模資料的儲存應用上。

3.3 演算法

建置圖書館參考諮詢機器人的過程中，本研究主要使用了三種演算法，分別為用於凸顯關鍵詞重要性並調整功能詞權重的 TF-IDF、允許調整意圖細緻程度並協助分群的 DBSCAN 以及自行訓練詞向量並找尋相關或相似詞作為查詢擴展的 Word2Vec，以下將一一說明。

3.3.1 特徵詞加權演算法：TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) 是一種特徵詞加權演算法，經常被應用於文字探勘、自然語言處理以及推薦系統計算詞彙重要性 (Vijayarani, Ilamathi & Nithya, 2015; Beel, Gipp, Langer & Breiting, 2016)。此一演算法不只考慮了詞彙在單一文本出現頻率，同時也一併考量其出現於語料庫所有文本的情況。

TF-IDF 是由 TF (Term Frequency) 和 IDF (Inverse Document Frequency) 相乘而得，其計算方式有各種變體 (Domeniconi, Moro, Pasolini & Sartori, 2015)。TF 和 IDF 的公式，如下所示：

$$\text{Term Frequency (TF)} = P(w|d)$$

$$\text{Inverse Document Frequency (IDF)} = \log \left(\frac{|D|}{|\{d \in D: w \in d\}|} \right)$$

¹ CkipTagger : <https://github.com/ckiplab/ckiptagger>

基本上，TF 計算的是詞彙在一個文本內的出現頻率；DF 則是計算在多少文本中曾出現過該詞彙，而由 DF 衍伸而來的 IDF 被用來反映並調整詞彙在所有文本出現頻率來評估其重要性。因此，TF-IDF 對於字詞權重調整的概念是：當一個詞彙多次出現在文本中時，應提高其權重，代表該詞彙越能作為該文本之關鍵語彙；然而，若該詞彙在所有文本都出現時，則應降低其權重，因為這表示該詞彙並不能有效作為區別該文本與其他文本的指標。

Term Frequency Factor	Notation	Description
1/0	BIN	Presence or absence of terms in the document
term frequency	TF	Number of times the term occurs in the document
$\log(1+tf)$	log TF	Logarithm of tf
$1 - \frac{r}{r+tf}$	ITF	Inverse term frequency, usually $r = 1$
$S(V_a) = \frac{1-d}{(1-d) + d \cdot \sum_{V_b \in Conn(V_a)} \frac{S(V_b)}{Conn(V_b)}}$	RW	Given a graph $G = (V, E)$, let $Conn(V)$ be the set of vertices connected to V . Typical value for d is 0.85.

圖表 9 TF (Term Frequency) 變體

資料來源：Domeniconi, G., Moro, G., Pasolini, R., & Sartori, C. (2015, July). A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of $tf \cdot idf$. In *DATA* (pp. 26-37).

本研究將 TF-IDF 演算法用於聊天機器人建置過程中的兩個地方，分別是 DBSCAN 分群前的特徵詞權重調整，以及 Chatbot 接收讀者輸入後的意圖匹配處理。針對不同的使用階段，本研究亦對 TF-IDF 公式進行些微調整。

3.3.2 Clustering 演算法：DBSCAN

Clustering 被中譯為分群、集群或聚類，是一種根據資料本身的特徵將其他同樣也具有相似特徵的資料逐步聚合為一個集群的過程。這類 Clustering 演算法具有以下特性：同一集群內的資料相似度較高，而集群與集群之間的相似度則較低 (Kanagala & Krishnaiah, 2016)。常見的分群演算法，包含 K-Means、K-Medoids、DBSCAN、Hierarchical clustering... 等。

本研究使用的 DBSCAN (Density Based Spatial Clustering of Applications with Noise) 是由 Martin Ester 等人於 1996 年提出的一種基於密度的分群演算法，是為了挖掘任意形態特徵的資料而發明的 (Ester, Kriegel & Sander, 1996; Rong, Yan &

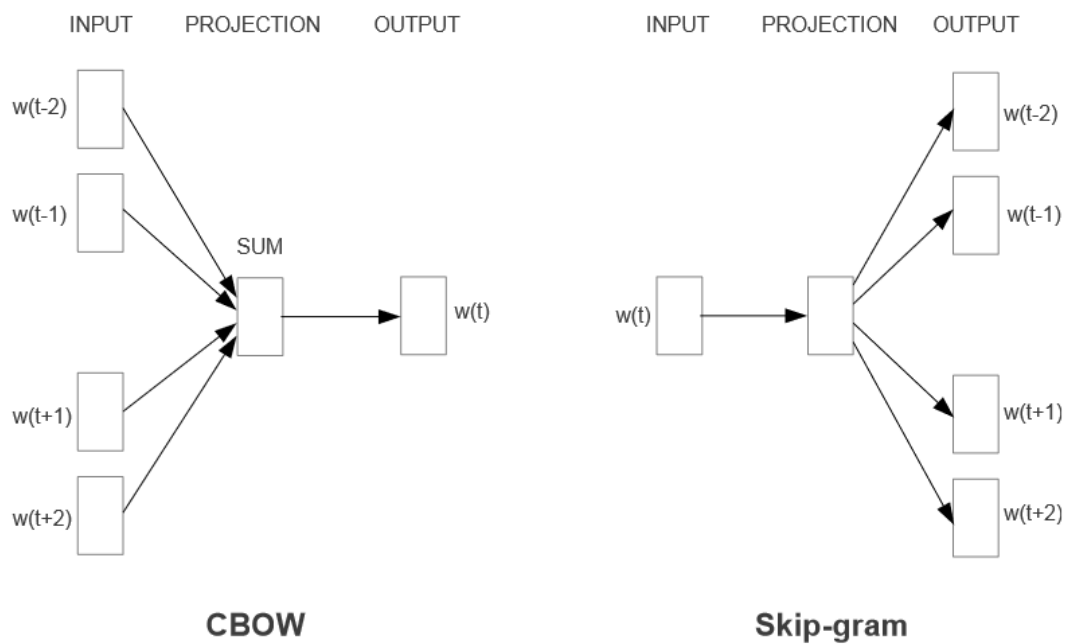
Guo, 2004)。DBSCAN 的分群概念是依據資料點在特徵空間中分佈的密集程度來進行聚合，亦即當某一區域的資料點特徵相近且密度超過一定門檻，則此區域內的所有資料點被歸為同一群。

DBSCAN 演算法可藉由兩個參數來調整分群結果，其一為設定半徑範圍(ϵ 或 Eps，代表相似度)，其二為設定半徑範圍內必須達到最小資料數目的門檻值 (MinPts)。簡單來說，其做法建立在以某資料點為核心點的基礎上，若在指定半徑範圍內擁有超過指定數量的樣本點(高密度)，便形成一個集群；而未被分配至任一集群的資料點，則被視為雜訊(noise)。其優點在於不需要事先規定劃分的集群數，而是由 DBSCAN 依據事先設定的參數自動分群，而且不受極端值(noise)影響。

3.3.3 詞向量演算法：Word2Vec

Word2Vec 是一種用來產生詞向量(Word Embedding)的技術，由 Google 的 Tomas Mikolov 研究團隊於 2013 年開發，其包含 CBOW 和 Skip-gram 兩種模型架構(Mikolov, Chen, Corrado & Dean, 2013)。相較於過去在自然語言處理利用 Bag of Words 概念來處理單詞時，把單詞看作一個一個獨立的個體並以編碼向量的形式來表示(如：One-hot encoding)，不僅忽略單詞的出現順序，也無法涵蓋詞彙在上下文所包含的語意關係資訊；Word2Vec 的出現解決了上述問題，也讓電腦得以藉由類神經網路從大量訓練語料中自動捕捉單詞之間精確的句法(syntactic)和語意(semantic)關係來更好地表達單詞，從而獲得相對低維度且富含語意特徵的詞向量。

Word2Vec 的原理源自於語言學中的分佈假說(Distributional hypothesis)，即具有相似分佈情形(上下文)的詞彙可能表達相似的含意(Harris, 1954)。因而，開發出 CBOW 和 Skip-gram 兩種模型架構來訓練詞向量。CBOW 模型是利用上下文來當作類神經網路的輸入，再輸出所預測的中心詞(當前的字詞)；而 Skip-gram 恰好相反，它是根據中心詞(當前的字詞)來預測其上下文(見圖表 10)。



圖表 10 Word2Vec 的兩種模型架構

資料來源：Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

本研究使用的 Word2Vec 是 300 維、語彙等級(Word based)之繁體中文預訓練詞嵌入模型，由臺北醫學大學的自然語言處理實驗室利用 Python 的 gensim 套件所訓練並公開於網路上。上述已被預訓練的 Word2Vec 模型，將會與維基百科的分類一同被應用於 Chatbot 的問句查詢擴展(Query Expansion)，協助找尋相關或相似詞，以彌補訓練語料之同義詞或同類概念的不足。

3.4 系統建置流程

為建置此一參考諮詢機器人，本系統的建置流程共可分為三大階段，分別為訓練語料蒐集與分析、訓練語料預處理、LineBot 系統功能建置。

3.4.1 訓練語料蒐集與分析

開始建置圖書館參考諮詢機器人之前，首先必須蒐集圖書館常見問答集作為訓練語料，以便藉此了解讀者在參考諮詢時經常使用的詞彙或是習慣問法。因此，本階段使用了 Python 的 Beautiful Soup 函式庫進行網路爬蟲作業，抓取台灣各圖書館網站的常見問答集語料，其範圍包括大學圖書館、公共圖書館(包含直轄市圖書總館與省轄市立圖書館)和國家圖書館網站，共 100 間(完整圖書館名單見附錄一)，總計 4975 個問答。

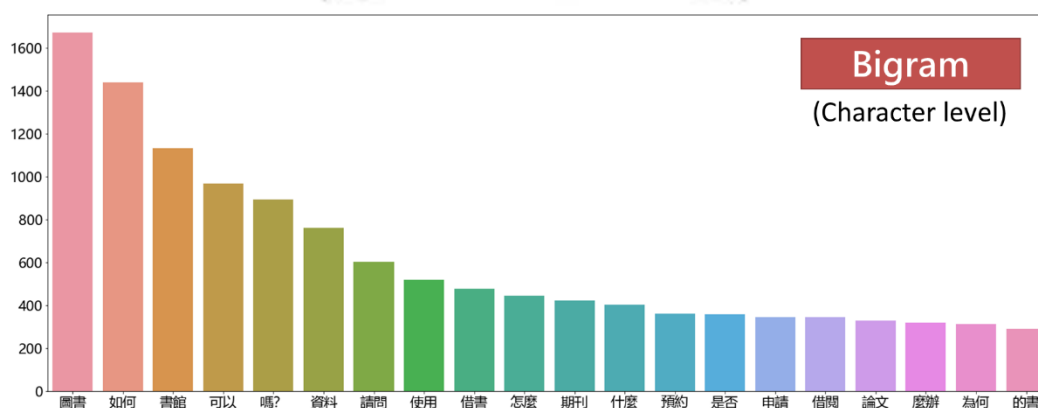
針對每間圖書館的常見問答集，本研究一筆一筆地將其儲存至 MongoDB，每筆問答紀錄包含 LibraryName (圖書館名稱)、ID、Question (問題)、Answer (答案)、Category (所屬類別)、Keyword (關鍵詞)以及 RelatedQ (相關問題)共計七個欄位，如圖表 11 所示。接著便可針對問答語料文本進行分析和預處理工作，以挖掘出語料本身的結構或模式(pattern)。

LibraryName String	ID Int32	Question String	Answer String	Category String	Keyword Null	RelatedQ Null
"國家圖書館"	1	"國家圖書館開放時間為何?"	"(1) 本館 (臺北市中山南路29號) 部	"開放時間"	null	null
"國家圖書館"	2	"國家圖書館為何週一休館?"	"圖書館對於資訊系統軟、硬體設備之"	"開放時間"	null	null
"國家圖書館"	3	"國家圖書館為何四樓以上各專室只開"	"本館四樓以上專科閱覽室有錄影資料"	"開放時間"	null	null
"國家圖書館"	4	"為何國家圖書館入館年齡需為年滿16歲"	"國家圖書館係以政府機關(構)、法人"	"辦理與人館資格"	null	null
"國家圖書館"	5	"如何辦理國家圖書館之閱覽證?需準"	"(1) 依據<國家圖書館閱覽服務規定"	"辦理與人館資格"	null	null
"國家圖書館"	6	"為何國家圖書館不外借圖書與人館部"	"為維持圖書館的秩序、保障館內閱覽"	"辦理與人館資格"	null	null
"國家圖書館"	7	"館內能否可以使用行動電話?"	"讀者不可在圖書公共空間使用行動"	"人館規範"	null	null
"國家圖書館"	8	"人館利用館藏資源時，可否再借包攜"	"本館依「圖書辦法」為公開出版品之"	"人館規範"	null	null
"國家圖書館"	9	"國家圖書館的館藏資源主要包含哪些?"	"(1) 中外文圖書、期刊、報紙、雜誌"	"典藏資源"	null	null
"國家圖書館"	10	"如何查詢利用國家圖書館之「電子館"	"因應資訊科技的發展，本館近年來所"	"典藏資源"	null	null
"國家圖書館"	11	"國家圖書館所建置提供的電子資料庫"	"為方便讀者利用電腦及網路查詢資料"	"典藏資源"	null	null
"國家圖書館"	12	"國家圖書館收藏之博碩士論文典藏地"	"(1) 碩士論文：a. 陳列原則：三年"	"典藏資源"	null	null
"國家圖書館"	13	"為何有些博碩士論文不開放或拖"	"有的博、碩士論文之作者，可能基於"	"典藏資源"	null	null
"國家圖書館"	14	"如何推薦圖書資料給國家圖書館?"	"可親自到本館各服務台填寫「國家圖"	"典藏資源"	null	null
"國家圖書館"	15	"圖書館是否提供到館之前事先在館外"	"已持有本館閱覽證讀者可利用本館館"	"圖書詢問"	null	null

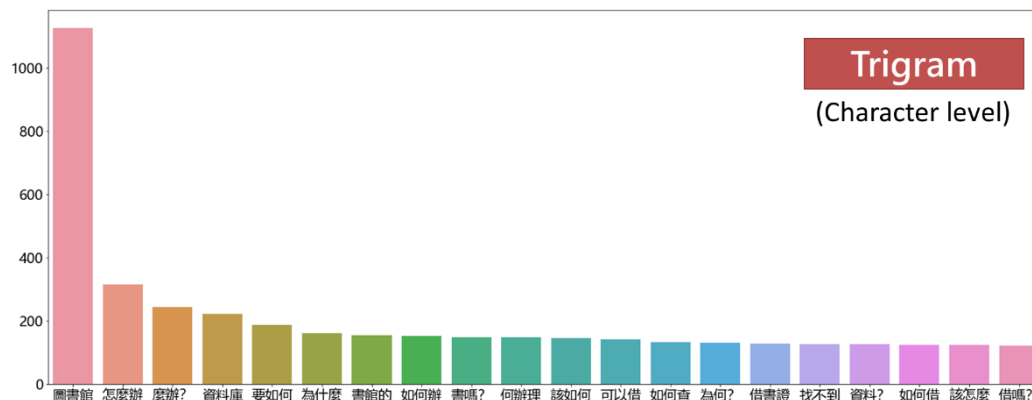
圖表 11 圖書館常見問答集的語料蒐集範例
(研究者自行整理)

對於語料中的語彙分析，常見方法有 N 元語法(N-gram)、字詞頻率以及共現詞分析(co-occurrence analysis)...等。N-gram 是一個基於滑動窗口(sliding window)概念，從文本中提取連續 N 個字符(Character) 或單詞(Word) 序列的語言模型(Majumder, Mitra & Chaudhuri, 2002)；字詞頻率的分析是以量化方式統計字詞在文本中的分布情形；共現詞指的是以一定頻率共同出現且不連續於同一篇文章的字詞。透過對共現詞進行分析，將有助於找到經常伴隨出現的字詞組合，從而得知特定主題下具有關聯性或相似性的習慣用語(Jiao, Liu & Jia, 2007)。

將問答語料以字符(Character)層次進行分析，其結果如圖表 12(Bigram)和圖表 13(Trigram)所示。從圖表 12 顯示較高頻的 20 個字詞中，「圖書」出現頻率超過 1600 次，可知在所蒐集的問答語料中「圖」與「書」這兩個字總是一同連續出現，是個常見的字符組合。另外，亦發現「書館」這個組合雖然出現頻率相當高，卻不是一個能完整表達某種概念的詞彙。若是再對問答語料進行進一步的 Trigram 分析，將會觀察到一個有趣的現象—原本在 Bigram 是高頻且分開狀態的「圖書」和「書館」，如今在 Trigram 的狀態下以一定頻率連著出現並組合在一起的「圖書館」一詞將更能表達完整的含意。由此可知，藉由 N-gram 將能挖掘並提取常見的連用語組合(例如：資料庫、借書證、找不到...)。而後本研究將會把這些連用語組合以自訂字典的方式加入至 CkipTagger 的中文斷詞上，這有助於從根本上提升自然語言理解的準確度。

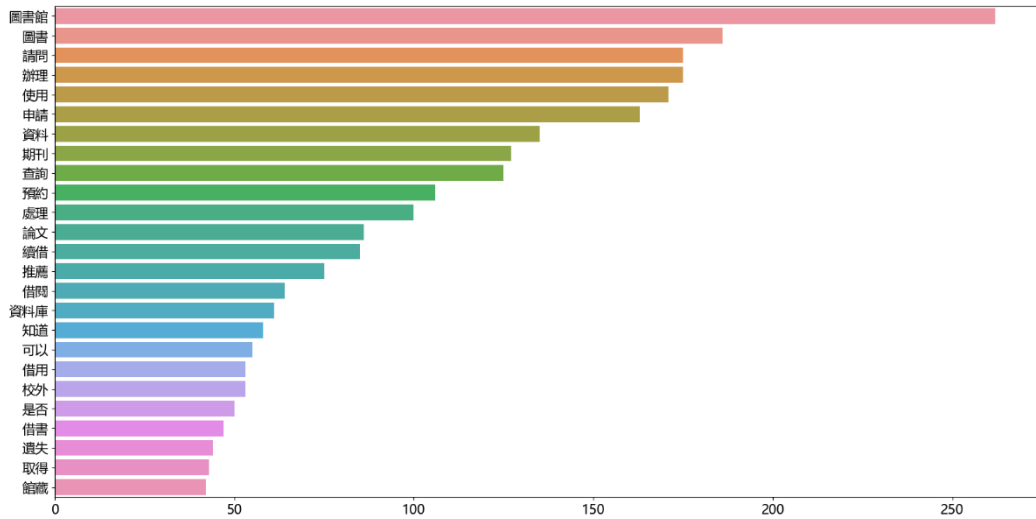


圖表 12 字符(Character)層次的 Bigram 次數統計長條圖
(研究者自行整理)

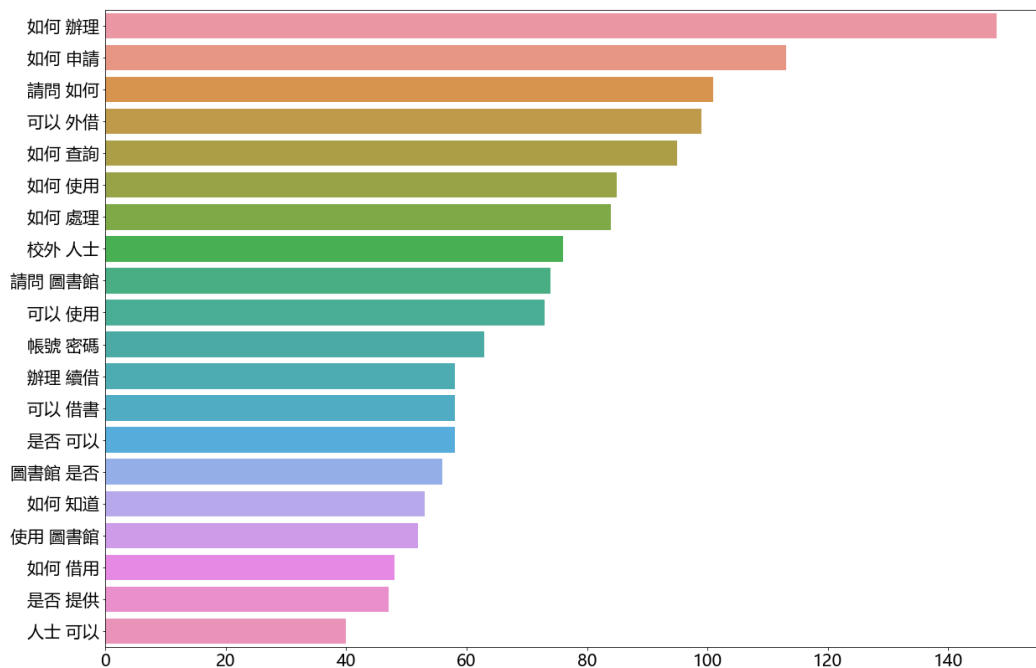


圖表 13 字符(Character)層次的 Trigram 次數統計長條圖
(研究者自行整理)

另外，假如以共現詞(co-occurrence)的角度來分析問答語料庫，可發現在文本中詞彙和詞彙間在使用上亦存在著某種關聯。在語法結構上，這些詞彙在文本中可能是緊連著出現，也可能不連續但會同時出現。例如：針對「如何」一詞的進行共現詞分析，可以發現該詞彙往往與「圖書館」、「圖書」、「請問」和「辦理」...等字詞在同一個問句內出現(見圖表 14)。假如再搭配詞彙(Word)層次的 Bigram 分析，可以知道「辦理」、「申請」、「查詢」和「使用」...等字詞經常和「如何」一詞緊鄰出現(見圖表 15)。綜上所述，可見「如何」經常作為一個問句的開頭，而且讀者經常使用「如何」來詢問「圖書館」、「圖書」、「期刊」、「論文」和「借書」...等圖書館的業務範圍。



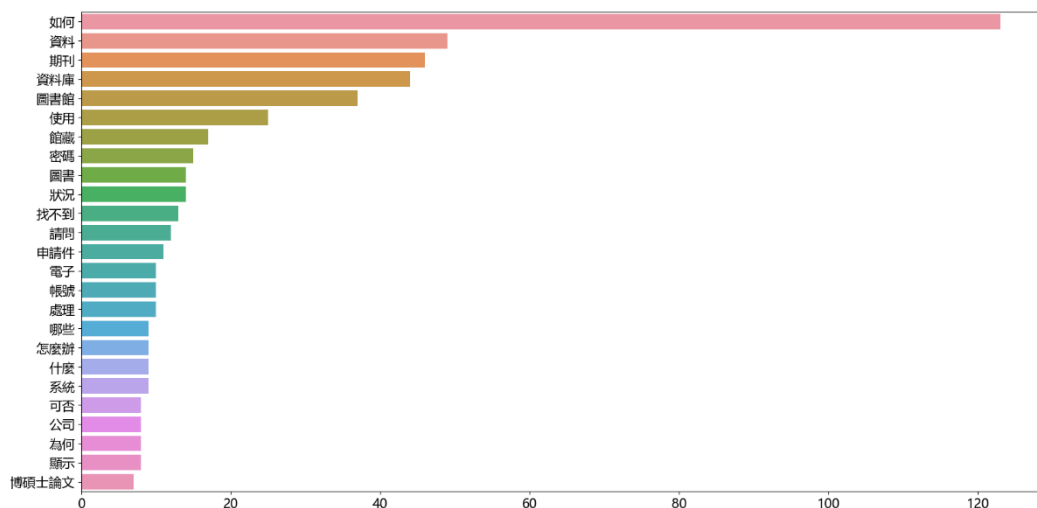
圖表 14 「如何」一詞之共現詞分析的次數統計長條圖
(研究者自行整理)



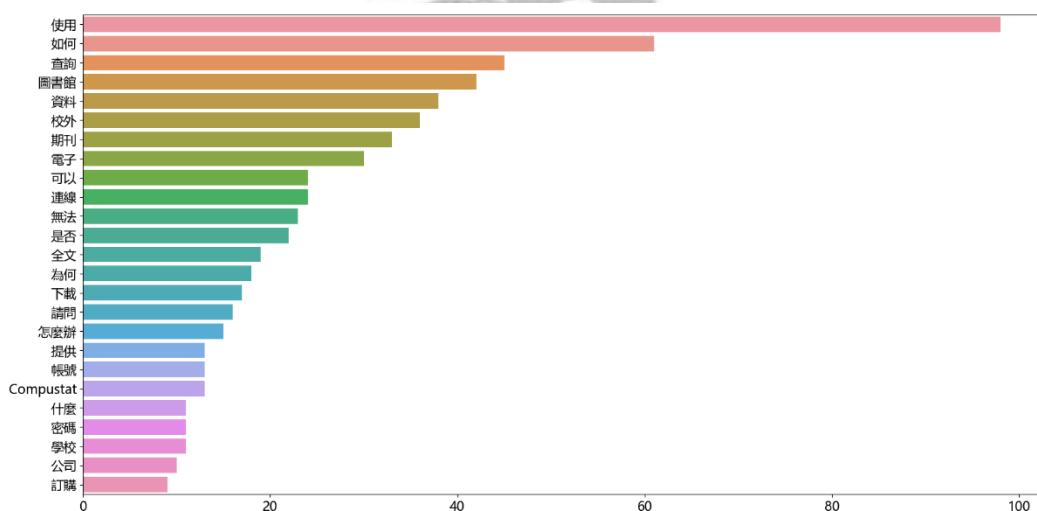
圖表 15 字詞(Word)層次的 Bigram 次數統計長條圖
(研究者自行整理)

若是想進一步找出與「查詢」一詞共現的詞彙，則可知共現詞包含了「資料」、「期刊」、「資料庫」、「館藏」、「找不到」和「碩博士論文」...等(見圖表 16)。其中，「資料庫」一詞在問句裡又與「使用」、「連線」、「全文」和

「Compustat」等詞彙有著某種相關性(見圖表 17)。經過層層的共現詞分析，讓詞彙間的關聯性顯而易見。而藉著多個詞彙的共現性，對於確認並限縮問句所要表達的詢問意圖或許會有莫大的幫助。這種共現關係，本研究將列入考量，並應用於 DBSCAN 分群之後，作為意圖集群的微調。



圖表 16 「查詢」一詞之共現詞分析的次數統計長條圖
(研究者自行整理)



圖表 17 「資料庫」一詞之共現詞分析的次數統計長條圖
(研究者自行整理)

除了對問答語料進行語法結構的分析之外，本研究亦試圖提取具有相似問法或相同詢問目的(意圖)的問句，以期讓最終建置的圖書館參考諮詢機器人能夠辨識讀者的意圖並給予相應的答覆。為構建基於意圖辨識的機器人，需要將具有相似詢問目的的問句歸類至同一群。而審視儲存於 MongoDB 的原始蒐集問答語料庫後，發現「Category (所屬類別)」此一欄位無法被直接作為意圖來使用，理由如下：由於每間圖書館常見問答集的分類不同、分類的粗細程度也不一，而且部分圖書館給予分類類別別，他館則沒有(見圖表 18)；又或者由於人工分類的主觀因素，使得問答語料的分類類別並不適當。因此，本研究將回歸問句本身，透過理解問句本身的語意重心來解析訓練語料庫中間句間的相似性，並經由 DBSCAN 分群演算法的使用，獲得統一分群標準的意圖集群。隨後，再藉由分析各意圖集群的共現關係來進行細部微調。

問句：如何辦理圖書續借？		類別：借還書
相似度	問句	類別
1.0000	如何辦理圖書續借？	流通(借還書服務)
0.8870	如何辦理續借？	借書問題
0.8870	如何辦理續借？	false (尚未分類)
0.8870	如何辦理續借？	圖書流通問題
0.8870	如何辦理續借？	借閱服務
0.8721	如何辦理圖書續借？何時可以辦理續借？	續借
0.8367	如何續借圖書？	false (尚未分類)
0.8367	如何續借圖書？	圖書借閱
0.8082	圖書可以續借嗎？要如何辦理續借呢？	借書、還書及預約
0.7969	請問如何辦理續借？	圖書館借閱查詢問題
0.7963	如何辦理續借？	false (尚未分類)
0.7963	如何辦理續借？	續借
0.7883	要如何辦理續借？	借閱問題
0.7575	如何辦理續借呢？	續借問題
0.7226	我如何在線上辦理圖書續借？	入館閱覽及借閱服務
0.7189	如何辦理圖書續借？何時可以辦理續借？什麼樣的情況下無法辦理續借？	續借圖書
0.7025	請問該如何辦理續借？	圖書館

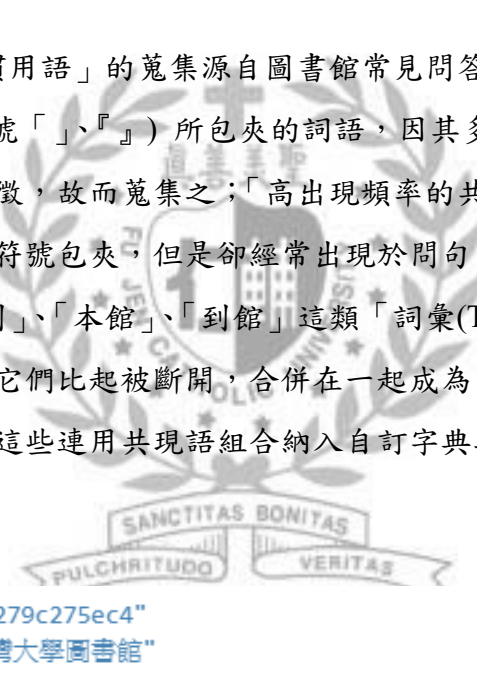
圖表 18 整併相似類別與修正問句分類之範例
(研究者自行整理)

3.4.2 訓練語料預處理

為了讓聊天機器人正確地理解一個問句，需要從表達語意的最小單位「詞彙(Token)」著手，才能從根本上提升自然語言理解的準確度，從而建構出一個自然語言理解的處理器。

對於中文處理而言，能正確標示出文本中詞彙(含詞組)對於理解自然語言是最重要的第一步。為了改善斷詞結果，本研究將維基百科條目、圖資領域慣用語(見圖表 19) 以及高出現頻率的共現 N 字詞 (N-gram collocations) 加入 CkipTagger 斷詞器的自訂字典之中。

其中，「圖資領域慣用語」的蒐集源自圖書館常見問答集中被具有開頭和結尾之標點符號(如：引號「」、「」)所包夾的詞語，因其多半為名詞短語，具有特別強調其重要性的特徵，故而蒐集之；「高出現頻率的共現 N 字詞」蒐集的是沒有開頭和結尾之標點符號包夾，但是卻經常出現於問句中的相連 N 字詞，如圖表 20 例句中的「找不到」、「本館」、「到館」這類「詞彙(Token)」極易被斷詞器斷開，就語意上來說，它們比起被斷開，合併在一起成為「詞組」才能更好地表達完整的意思，故而將這些連用共現語組合納入自訂字典之中。



```
_id: "5f714dea3cd0e4279c275ec4"  
LibraryName: "國立台灣大學圖書館"  
ID: "206"  
Question: "持 校友證 有何優惠?"  
Answer: "可於本館開放時間入館查閱及使用資料，並比照本校讀者可將個人背包及書籍攜入。  
Category: "借閱規則及服務"  
Keyword: "["校友"]"  
RelatedQ: "["校友可以借書嗎?","如何辦理?","校友可以使用圖書館資料庫嗎?","畢業學
```

圖表 19 將引號內的詞納入圖資領域慣用語之範例
(研究者自行整理)



圖表 20 尚未考量 N 字詞的共現關係之斷詞範例
(研究者自行整理)

把訓練語料的句子經中研院釋出的 CkipTagger 工具進行中文斷詞(Word segmentation)後，接下來使用 TF-IDF (Term Frequency – Inverse Document Frequency) 演算法來對字詞進行重要程度的加權，公式如下：

$$TF = tf(t, d)$$

$$IDF = idf(t, D) = \log\left(\frac{1+n}{1+df(D, t)}\right) + 1$$

$$TF_IDF = tf_idf(t, d, D) = tf(t, d) \times idf(t, D)$$

因為對於一個句子來說，不管是問句或是敘述句，在傳達資訊時必定會有訊息焦點，亦即語意重心，所以句子內每一個單詞 (Token) 其重要性應有大小之別，賦予不同的權重。

隨後，本研究將已斷詞的問句與答句透過 TF-IDF 詞彙加權調整，得到能夠投射至向量空間的句子向量，方便後續藉由分群演算法把具有特定樣式(pattern)或同樣關鍵詞的句子劃分為同一群，從而獲得一個擁有一致分群標準且具有相同詢問意圖的集群。

事實上，本研究嘗試過 Transformers²、K-Means 和 DBSCAN 三種演算法。以「在架上找不到書，該怎麼辦？」此一問句為例，其重新分類或分群的結果如圖表 21、圖表 22 和圖表 23 所示。從 Transformers 演算法重新分類預測的結果中，可得

² <https://simpletransformers.ai/docs/multi-class-classification/>

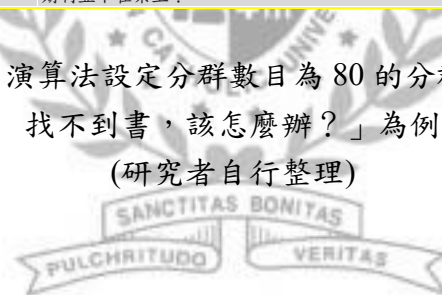
與問句相關的類別為「館藏使用相關問題」，此類別下總計 17 個文本(問句)，經判斷後得知僅有 4 個文本是相關的，表示與該類別真正相關的文本僅佔比 24%；而從 K-Means 和 DBSCAN 演算法的分群結果中，得知其各自文本相關度的佔比為 67%和 47%。實驗結果顯示，Transformers 演算法之問句重新分類結果最差，而 K-Means 和 DBSCAN 的分群結果較好，其中又以 K-Means 最優。假如進一步分析所屬集群中較不相關的語料，則可發現經 Transformers 分類預測的不相關語料與問句之間在詢問意圖上毫無規則或關聯可言；相較之下，透過 K-Means 和 DBSCAN 進行分群的不相關語料，其與問句之間仍是比較相關的，從中可看到這些不相關語料大多保留著「不在架上」、「找不到」或「期刊」的字眼。

演算法	Transformers
所屬集群/類別	館藏使用相關問題
所屬集群/類別之語料總數	17個
與問句相關的語料佔所屬集群下的比例	4/17 = 0.2353
所屬集群內與問句相關的語料範例	在架上找不到書，該怎麼辦？ 找不到書，圖書館可以幫忙找？ 資料不在架上，遍尋不獲，該怎麼辦？ 資料不在架上，遍尋不獲，該怎麼辦？
與問句不相關的語料佔所屬集群下的比例	13/17 = 0.7647
所屬集群內與問句不相關的語料範例	想要贈書給圖書館，要如何處理？ 我想要學習如何查尋資料，要如何參加？ 圖書館的電腦可以使用隨身碟？ 如何知道圖書館有何新的館藏？ 忘記【我的帳戶】登入帳號密碼，怎麼辦？ 各類資料放置在何處？ 如何知道圖書館有何新的館藏？ 各類資料放置在何處？ 我想找考古題？ 如何知道圖書館有何新的館藏？ 如何推介資料/圖書館處理推介資料的流程？ 參考書/博碩士論文/教師課程指定參考書可以外借嗎？ 考古題都只有題目，是否可以提供詳解呢？

圖表 21 使用 Transformers 演算法的重新分類結果，以問句「在架上找不到書，該怎麼辦？」為例
(研究者自行整理)

演算法	K-Means
所屬集群/類別	Cluster 4
所屬集群/類別之語料總數	30個
與問句相關的語料佔所屬集群下的比例	$20/30 = 0.6667$
所屬集群內與問句相關的語料範例	館藏查詢圖書為「在架上」但卻找不到書，其原因？該如何處理？ 常常要找書，系統明明顯示在架上，可是就是找不到？如果又急著要用，怎麼辦？ 「歡迎外借」的書，為什麼在架上找不到呢？ 請問在網路上查書顯示在架上，但到書架上卻找不到書怎麼辦？ 為何在架上找不到「在架」的書？是否遺失了呢？ 為何在架上找不到「在架」的書？是否遺失了呢？ 在架上找不到書，該怎麼辦？ 為何圖書館藏狀態顯示【在架上】，卻在書架上找不到該圖書？ 資料顯示在架上，但實際卻找不到？ 書況顯示「在架上」的書找不到，怎麼辦？ 為什麼使用圖書館查詢系統顯示書本在架上，實際到架上尋找的時候卻找不到書？ 資料不在架上，遍尋不獲，該怎麼辦？ 系統顯示資料未借出，但不在架上，要怎麼辦？ 查詢館藏目錄，電腦顯示「在架上」，可是為何找不到書？該怎麼辦呢？ 資料不在架上，遍尋不獲，該怎麼辦？ 為何在架上找不到我要的書籍或刊物？ 在網路上查書顯示在架上，但到書架上卻找不到書怎麼辦？ 館藏記錄上有，但書不在架上怎麼辦？ 我要借的書不在架上，怎麼辦？ 我在電腦查到書在架上，為何找不到？
與問句不相關的語料佔所屬集群下的比例	$10/30 = 0.3333$
所屬集群內與問句不相關的語料範例	我怎麼在架上會找不到語言學習類的期刊呢？ 為什麼現期的期刊卻在架上找不到？ 請問期刊不在架上該怎麼辦？ 為什麼從電腦目錄上查到的期刊卷期，卻在架上找不到？ 請問期刊不在架上該怎麼辦？ 我想找的期刊，在架上卻找不到？ 請問現期期刊不在架上該怎麼辦？ 為什麼在架上找不到以前的期刊？ 請問期刊不在架上怎麼辦？ 期刊並不在架上？

圖表 22 使用 K-Means 演算法設定分群數目為 80 的分群結果，以問句「在架上找不到書，該怎麼辦？」為例
(研究者自行整理)



演算法	DBSCAN
所屬集群/類別	Cluster 284
所屬集群/類別之語料總數	17個
與問句相關的語料佔所屬集群下的比例	8/17 = 0.4706
所屬集群內與問句相關的語料範例	書找不到怎麼辦？ 我找不到我想借的書，怎麼辦？ 書找不到怎麼辦？ 在架上找不到書，該怎麼辦？ 資料不在架上，遍尋不獲，該怎麼辦？ 資料不在架上，遍尋不獲，該怎麼辦？ 我要借的書不在架上，怎麼辦？ 找不到書，該怎麼辦？
與問句不相關的語料佔所屬集群下的比例	9/17 = 0.5294
所屬集群內與問句不相關的語料範例	為什麼現期的期刊卻在架上找不到？ 請問期刊不在架上該怎麼辦？ 請問期刊不在架上該怎麼辦？ 期刊並不在架上？ 找不到館藏期刊怎麼辦？ 請問期刊不在架上怎麼辦？ 我想找的期刊，在架上卻找不到？ 為什麼在架上找不到以前的期刊？ 請問現期期刊不在架上該怎麼辦？

圖表 23 使用 DBSCAN 演算法設定半徑 0.8、最小資料數目為 2 的分群結果，以問句「在架上找不到書，該怎麼辦？」為例
(研究者自行整理)

根據上述問句「在架上找不到書，該怎麼辦？」所屬集群的例子來說明，儘管 K-Means 的分群效果比 DBSCAN 好，但是卻不見得對其他集群的語料都是那麼的相關。以圖表 24 所列「與圖書館開放時間相關」的問句來看，可見它們在 K-Means 演算法下被分為 2 個集群，而在 DBSCAN 演算法下則被分為 8 個集群。若細究各個集群的詢問意圖，顯然由 DBSCAN 分群而得的語料較為精細，有助於幫助判斷更為細緻的詢問內容。因此，本研究選擇基於密度的 DBSCAN 演算法來得到較一致且準確的意圖集群，以便建置一個能夠判斷讀者詢問意圖的圖書館參考諮詢機器人。

問句	K-Means 演算法下的所屬集群	DBSCAN 演算法下的所屬集群
從何處可以得知圖書館開放時間及休館日?	Cluster 59	Cluster 118
從何處可以得知圖書館開放時間及休館日?	Cluster 59	Cluster 118
請問可由何處得知圖書館開放時間及休館日?	Cluster 59	Cluster 118
圖書館開放時間?國定假日有開館嗎?	Cluster 59	Cluster 158
圖書館開放時間?國定假日有開館嗎?	Cluster 59	Cluster 158
圖書館開放時間?國定假日有開館嗎?	Cluster 59	Cluster 158
請問圖書館的開放時間?假日或寒暑假有開館嗎?	Cluster 59	Cluster 158
影印室的開放時間?	Cluster 59	Cluster 327
影印室開放時間?	Cluster 59	Cluster 327
請問視聽區的開放時間?	Cluster 59	Cluster 440
請問多媒體視聽區的開放時間?	Cluster 59	Cluster 440
24H閱讀區開放時間?	Cluster 59	Cluster 515
國家圖書館開放時間為何?	Cluster 59	Cluster 65
請問圖書館開放時間?	Cluster 59	Cluster 65
請問圖書館開放時間?	Cluster 59	Cluster 65
請問圖書館的開放時間是?	Cluster 59	Cluster 65
圖書館開放時間?	Cluster 59	Cluster 65
請問圖書館開放時間?	Cluster 59	Cluster 65
圖書館開放時間為何?	Cluster 59	Cluster 65
請問圖書館的開放時間?	Cluster 59	Cluster 65
圖書館的開放時間?	Cluster 59	Cluster 65
圖書館開放時間?	Cluster 59	Cluster 65
非圖書館開放時間如何還書?	Cluster 59	Cluster 65
請問圖書館開放時間?	Cluster 59	Cluster 65
圖書館開放時間	Cluster 59	Cluster 65
開放時間	Cluster 59	Cluster 65
請問圖書館開放時間?	Cluster 59	Cluster 65
圖書館開放時間?	Cluster 59	Cluster 65
圖書館開放時間?	Cluster 59	Cluster 65
請問圖書館開放時間為何?	Cluster 59	Cluster 65
期中、期末考週有延長開館嗎?	Cluster 67	Cluster 157
期中、期末考期間，週末有延長開館嗎?	Cluster 67	Cluster 157
期中、期末考週有延長開館嗎?	Cluster 67	Cluster 157
期中、期末考週有延長開館嗎?	Cluster 67	Cluster 157
期中、期末考週有延長開館嗎?	Cluster 67	Cluster 157
桃園市立圖書館各館開館時間?	Cluster 67	Cluster 85
圖書館開館的時間為何?	Cluster 67	Cluster 85
請問高空大圖書館開館時間?	Cluster 67	Cluster 85
圖書館開館時間及借還書櫃台服務時間為何?	Cluster 67	Cluster 85
請問高空大圖書館開館時間?	Cluster 67	Cluster 85
圖書館的開館時間?	Cluster 67	Cluster 85
中市圖各分館開館時間?	Cluster 67	Cluster 85

圖表 24 使用 K-Means 和 DBSCAN 演算法的分群結果，以圖書館開放時間相關的問句為例
(研究者自行整理)

綜上所述，本研究選擇 DBSCAN 演算法應用於意圖分群的理由，主要有四項：其一，Transformers 多類別分類任務的做法是對原始類別進行特徵學習後再重新預測其類別，因此在分類標準不一的情況下學習特徵傾向會預測為擁有較多訓練語料的大類別，而大幅影響預測結果。反觀 DBSCAN 和 K-Means，由於分群是根據問答語料本身的句法結構或語意資訊來進行聚類，因此結果較為準確；其二，DBSCAN 不需要事先確定集群數目，而是依據所設定的參數自動分群，因此勝過

需要預先決定集群數目的 K-Means；其三，DBSCAN 不受極端值(noise)影響，因為極端值不會被硬性的歸類為任意一群，也就比較不會有不相似的問答語料歸類在集群內，勝過強硬將每一個問答語料歸類於某一集群的 K-Means；其四，DBSCAN 藉由設定半徑範圍(ϵ 或 Eps)，決定集群的數量和大小，從而彈性地調整集群內詢問問法的嚴謹或籠統，即界定詢問意圖的細緻度。

經 DBSCAN 分群後，我們得到較一致且準確的意圖集群。然而，由於參數設定得較為嚴苛，使得集群內的訓練語料傾向擁有相同的問法，而那些同樣具有相同意圖但不同問法的語料則會被另外自成一群，導致得到較多的意圖集群，共計 528 個。

為了將那些「具有相同意圖但不同問法」的集群合併在同一個意圖集群內，首先會對每個意圖集群的字詞進行 TF-IDF 加權運算，得到每群前 6 名的關鍵詞，而這些關鍵詞代表各意圖集群的語意重心。隨後，藉著這些語意重心的共現關係，篩選出具有相似關鍵詞分佈的意圖集群，再以人工來判斷該合併哪些意圖集群(見圖表 25)。

透過合併相似的意圖集群，增加各意圖集群下訓練語料的多樣性，有助於提升後續在意圖辨識(Intent Recognition) 階段的準確度，同時也方便基於詢問意圖(intent-based)的圖書館參考諮詢機器人藉由判定讀者意圖來決定要做什麼動作來回應。

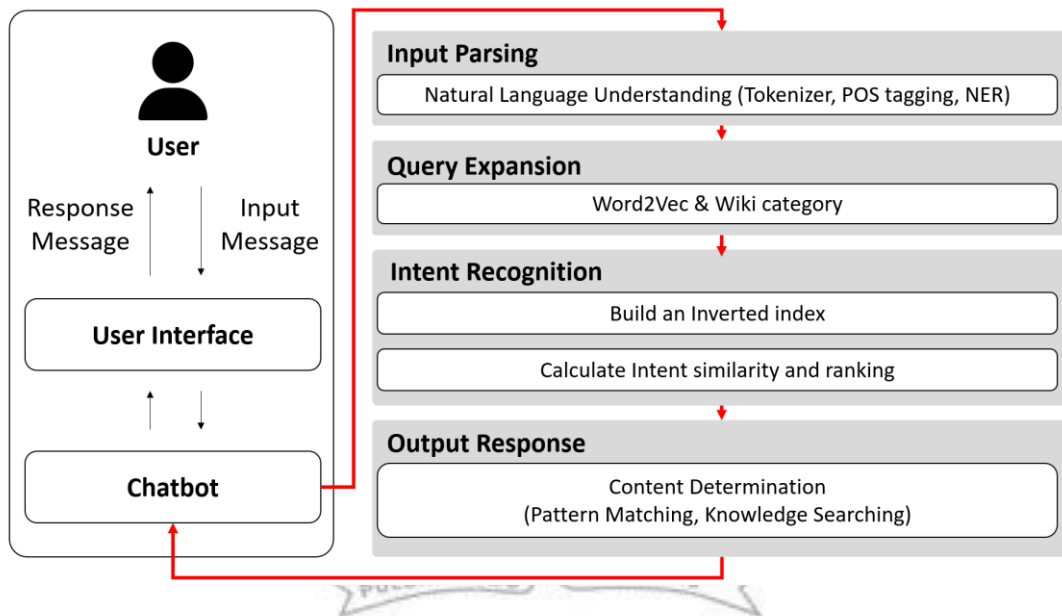
集群	加權後的前6名關鍵詞	關鍵詞權重	所屬集群之訓練語料範例	合併後的集群
Cluster 65	開放時間	0.9012	請問圖書館開放時間？	Cluster 65
	圖書館	0.3341		
	請問	0.1951		
	？	0.1336		
	為何	0.0927		
	非	0.0538		
Cluster 85	開館	0.6571	請問高空大圖書館開館時間？	Cluster 65
	時間	0.5674		
	圖書館	0.2141		
	高空	0.1951		
	大	0.1722		
	各	0.1588		
Cluster 139	開放	0.6591	圖書館開放館室時間為何？	Cluster 65
	時間	0.5429		
	為何	0.2862		
	館室	0.2783		
	圖書館	0.2731		
	的	0.1436		
Cluster 158	假日	0.5188	圖書館開放時間？國定假日有開館嗎？	Cluster 158
	開館	0.4759		
	開放時間	0.4364		
	國定	0.3891		
	有	0.2228		
	圖書館	0.1809		
Cluster 118	休館日	0.5304	從何處可以得知圖書館開放時間及休館日？	Cluster 118
	開放時間	0.4187		
	得知	0.3861		
	何處	0.3534		
	及	0.3022		
	從	0.2946		
Cluster 327	影印室	0.7534	影印室的開放時間？	Cluster 327
	開放時間	0.6338		
	？	0.1275		
	的	0.1035		
Cluster 440	視聽區	0.6543	請問多媒體視聽區的開放時間？	Cluster 440
	多媒體	0.4500		
	請問	0.3085		
	開放時間	0.3000		
	？	0.2240		
	的	0.1961		
Cluster 157	期中	0.4694	期中、期末考週有延長開館嗎？	Cluster 157
	期末考	0.4405		
	開館	0.4040		
	週	0.3360		
	、	0.2558		
	有	0.1891		
Cluster 515	閱讀區	0.5313	24H閱讀區開放時間？	Cluster 515
	24	0.4985		
	人士	0.2287		
	開放時間	0.2097		
	校外	0.1818		
	？	0.1566		

圖表 25 合併「具有相同意圖但不同問法」集群的範例，以圖書館開放時間相關的問句為例
(研究者自行整理)

3.4.3 LineBot 系統功能建置

當上述的「訓練語料蒐集與分析」以及「訓練語料預處理」完成後，便能開始建置圖書館參考諮詢機器人的系統功能模組，如圖表 26 所示。

讀者透過 LINE 使用者介面輸入的問題，系統會對該問題進行自然語言理解以及邏輯判斷處理。相關的處理步驟，依序包括問題解析(Input Parsing)、查詢詞彙擴展(Query Expansion)、詢問意圖辨識(Intent Recognition) 以及輸出回覆(Output Response)。



圖表 26 系統功能模組
(研究者自行整理)

3.4.3.1 問題解析 (Input Parsing)

「問題解析(Input Parsing)」的目的在於透過最小語意單位(Token)來理解讀者的問題，因此對於中文這種沒有分隔符號隔開單詞的語言來說，改善其斷詞結果將有助於更好地理解一個句子的語意，且斷詞結果的正確與否與最終的意圖匹配結果息息相關。因此，此一階段系統導入維基百科條目、圖資領域慣用語、高出現頻率的共現 N 字詞 (N-gram collocations)，來改善 CkipTagger 中文斷詞器，從而獲得不錯的斷詞、詞性標註和命名實體識別結果。

3.4.3.2 查詢詞彙擴展 (Query Expansion)

經解析後的問句，有時我們會發現某些詞彙並不在訓練語料的範圍內，若貿然進入後續的「詢問意圖辨識(Intent Recognition)」階段，則往往會得到不符預期且相當奇怪的意圖匹配結果。因此，有必要在「詢問意圖辨識(Intent Recognition)」階段之前進行「查詢詞彙擴展(Query Expansion)」。

在「查詢詞彙擴展(Query Expansion)」階段，本研究將從詞彙的同類概念(同義詞)著手，藉由維基百科的分類架構以及語詞等級(Word based)的繁體中文 Word2Vec 模型來建構一個擴展詞字典。

透過維基百科的分類架構，幫助我們獲得父類別與子類別之間上下階層的類別附屬關係，以及同一階層上對於某類別的實際概念舉例。舉例來說，圖表 27 的維基百科分類架構上，假如從上下階層的角度來看待「零食」一詞，可獲得口香糖、日本零食、甜甜圈、糖果、醃製水果以及零食品牌的 6 個子分類，而這些子分類下又各自涵蓋與其類別相關的概念和例子；若查看與「零食」相同階層的 26 個頁面，則可看出品客、三角巧克力、樂事、墨西哥玉米片、魷魚絲...等詞彙皆為「零食」一詞的實例。



圖表 27 維基百科的分類架構，以「零食」為例 (研究者自行整理)

而藉助臺北醫學大學自然語言處理實驗室所預訓練的高品質語詞等級(Word based)繁體中文 Word2Vec 模型，將有助於我們獲得更常見、直覺以及容易聯想的詞彙，這是因為該模型是由大規模語料之上下文與中心詞的關係所訓練而得，因此語料訓練的過程中也就保留了人們的常見用語以及相似上下文所帶來的相似語意資訊。不僅如此，Word2Vec 模型的應用也補充了維基百科分類架構上對於尚未建立分類關係之條目的不足。例如圖表 28「零食」一詞的擴展詞彙包含了零嘴、餅乾、糖果...等詞彙，而這些詞彙中只有「糖果」出現在維基百科分類架構中，其他相關的詞彙像是零嘴、餅乾、點心、甜食...等卻沒有存在其中相當可惜。

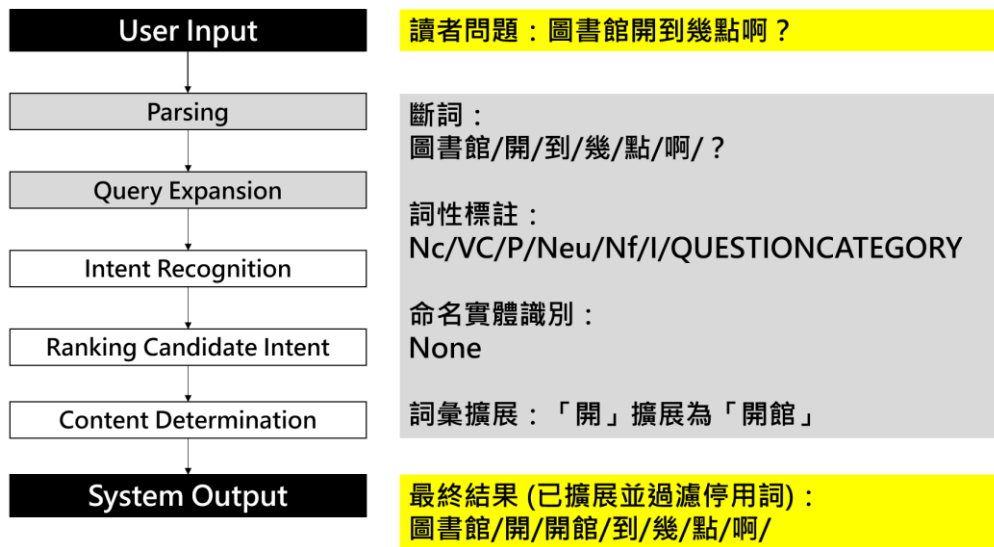
欲擴展之詞彙	前10名的相關詞彙	相似度	是否在維基百科分類架構中出現
零食	零嘴	0.7597	否
	餅乾	0.6770	否
	糖果	0.6729	是
	飲料	0.6506	否
	食物	0.6324	否
	泡麵	0.6263	否
	甜食	0.6229	否
	水果	0.6090	是
	點心	0.6016	否
	宵夜	0.5977	否

圖表 28 Word2Vec 模型的相似度計算結果，以「零食」為例
(研究者自行整理)

因此，綜合上述所言，本階段結合維基百科分類架構的多實例優點，以及 Word2Vec 模型提供的直覺、易聯想相似詞，同時也考量了本圖書館參考諮詢機器人的應用場域，以建構一個針對館務問題的擴展詞字典。此一階段，首先將篩選出與館務問題相關且高使用頻率的詞彙，並以這些詞彙為基礎利用維基百科以及 Word2Vec 模型進行詞彙擴展，協助找尋相關或相似詞，以提升對於館務相關詢問意圖的辨識度。

3.4.3.3 詢問意圖辨識 (Intent Recognition)

經過上述兩階段的處理後，將取得讀者問題的斷詞、詞性標註、命名實體識別以及詞彙擴展結果，如圖表 29 所示。接著，利用 TF-IDF 演算法為詞彙的重要性進行權重調整，再進行與意圖集群之間的相似度匹配，最後依據匹配結果的高低排序出最有可能的意圖集群。



圖表 29 經問題解析與查詢詞彙擴展的輸出結果
(研究者自行整理)

為獲取讀者問題與意圖集群之間的相似度，需要將讀者問題與每一個意圖集群進行相似度匹配計算，可想而知這樣的方式會耗費極高的運算時間和儲存空間。因此，為改善並加速匹配速度，這裡將借用文件檢索系統中最常用且兼具效率與彈性的資料結構—反向索引(Inverted Index)來實作。

反向索引(Inverted Index)的做法是預先建立單詞的索引表，索引表中記錄著單詞曾在哪些文件位置(意圖集群)中出現過，以便將來直接比對包含該單詞的文件(意圖集群)，並且僅對這些文件(意圖集群)進行相似度計算，從而降低時間複雜度(Yan, Ding & Suel, 2009)。因此，本階段將建構一個基於 TF-IDF 的反向索引。

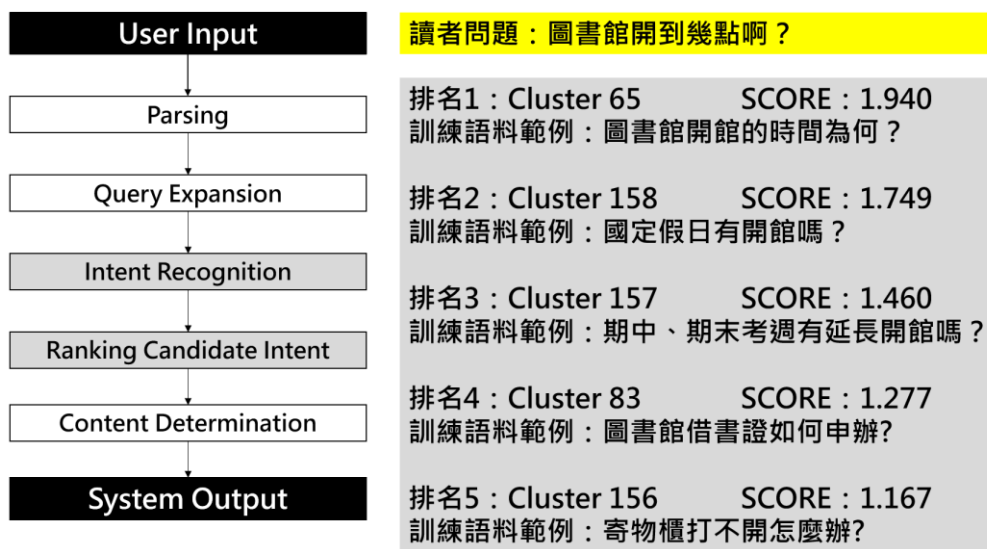
經多次實驗後，選擇如下的 TF-IDF 變體公式進行意圖集群與讀者問句之間的評分並排序。

$$TF = tf(t, d) = \log(1 + tf)$$

$$IDF = idf(t, D) = \frac{\log\left(\frac{n}{df(D, t)}\right)}{d_len}$$

$$TF_IDF = tf_idf(t, d, D) = tf(t, d) \times idf(t, D)$$

以「圖書館開到幾點啊？」讀者問題為例，此一階段的最終輸出結果如下圖表 30 所示，得到最為相似的前 5 名意圖集群，分別為第 65、158、157、83、156 個意圖集群。從中，發現只有第 65、158 個意圖集群與讀者真正的詢問目標最為相關，因此我們預期最終應該要有兩個系統回覆訊息。而為了獲得最為相關的系統回覆訊息，本研究選擇相似度評分在平均以上的意圖集群作為輸出。



圖表 30 經意圖辨識與排序的候選意圖集群結果
(研究者自行整理)

3.4.3.4 輸出回覆 (Output Response)

最為相關的意圖集群辨識出來後，接著便要開始規劃輸出給讀者的內容，這時需要自然語言處理技術的輔助，以及視讀者查詢意圖決定該從圖書館常見問答語料、輔大館藏查詢 API 還是 Wptools 維基百科知識庫提取工具中獲取解答。

在此階段將會分析圖書館常見問答語料中「問句與答句」之間的關聯，並獲得「問題類型」與「疑問詞短語」之間的對應關係(見圖表 31)。基本上，圖書館常見問答集依據詢問的目標可區分為具體時間、持續時間、規則、地點、數量、原因、方法、是否和定義這九種問題類型。根據所觀察到的規則(pattern)，發現多數的疑問詞(如：什麼、哪些…)需要搭配其前後的詞語(如：時候、資料庫…)，方能完整表達使用者想獲得的答案類型。

問題類型	疑問詞短語	關鍵疑問詞	例子
項目(具體時間)	開放時間為何	(N)+為何_D	圖書館開放時間為何?
	什麼時間/什麼時候/何時/哪個時候/何時	何時_Nd	在網路上預約書籍完成後，何時可到館取書?
持續時間	多長時間/多少時間	多久_Nd	借閱多久呢?
項目(規則)	規定為何/限制為何	(N)+為何_D	使用館際合作服務的資格限制為何?
地點	什麼地方/什麼地點/哪裡/哪兒/何處	哪裡_Ncd/何處_Nc	請問我要到哪裡找過期的期刊資料呢?
數量	多少/幾	幾本_Na	我可以借幾本書?
原因	為何	為何_D+(N+V)	為何收到逾期通知?
	為什麼	為什麼_D	請問為什麼我介購的圖書遲遲未到館呢?
方法	如何辦理/如何申請/如何利用/如何捐贈/如何得知	如何_D+(V)	如何辦理閱覽證?
	該怎麼辦呢/該怎麼辦/怎麼辦	怎麼辦_VH	我想借的書，圖書館沒有收藏，怎麼辦?
	該如何/要如何	要_D/該_D+如何_D+(V)	該如何辦理借書手續呢?
是否	是否/是否有/是否可以/是否提供	是否_D+(V)	館內是否可以使用行動電話?
	可否/可否用	可否_D	館藏圖書資料可否外借?
	有...嗎/可以...嗎/可以嗎/可以	有_V_2/可以_VH+...+嗎_T	校友可以使用圖書館資料庫嗎?
	可不可以	可不可以_D	在家或宿舍可不可以使用圖書館的資料庫呢?
定義	包含哪些/哪些/有那些/有那些	包含_VJ/有_V_2+哪些_Neqa+(N)	校外人士進入圖書館，有那些規定?
	有什麼/是什麼/什麼管道	什麼_Nep	如果圖書逾期罰款一直不處理，那麼會有什麼後果嗎?

圖表 31 九種問題類型的常見疑問詞問法
(研究者自行整理)

接下來，將細究以下九種問題類型的處理方式，並配合本研究建置的圖書館參考諮詢機器人來展示輸出結果。

3.4.3.4.1 「項目(具體時間)」問題類型

在圖書館問答的訓練語料中，這種關於詢問具體時間的問題類型，其疑問詞短語通常會出現「開放時間為何」、「什麼時間」、「什麼時候」、「何時」、「哪個時候」、「幾點」...等字詞，而且假如讀者想要更進一步詢問確切的時間的時候，往往會伴隨著「平日」、「假日」、「今天」、「週一」...等這類詞性標註為 Nd 的時間詞加入詢問。

因此，在規劃系統輸出的內容時，將會先擷取問句中詞性標註為 Nd 的時間詞，經邏輯判斷處理後，才將相應的回覆輸出(見圖表 32)。

判斷條件	可能情況	最終輸出
輸入內容是否有詞性標註為 Nd 的時間詞	現在情況	擷取讀者的所在時區和日期後，進行平日或假日情況的星期判斷後輸出。
	寒暑假或其他情況	範例見圖表 33
	平日情況	範例見圖表 34
	假日情況	範例見圖表 35
輸入內容是否有特殊字串	有「疫情」	範例見圖表 36

圖表 32 「項目(具體時間)」問題類型之回覆判斷條件
(研究者自行整理)

已讀
下午 4:32 輔大圖書館寒暑假幾點開？



輔大圖書館各館的開放時間，詳情請見：
<http://web.lib.fju.edu.tw/chi/intro/opentime>
國定及校定假日特殊開放時間：<http://web.lib.fju.edu.tw/chi/news/20200915>

下午 4:32

圖表 33 寒暑假或其他情況的「項目(具體時間)」問題類型範例
(研究者自行整理)

已讀
下午 5:04 平日圖書館的開放時間？



平日 各館的開放時間如下：
濟時樓：8:00 ~ 22:00
公博樓：8:00 ~ 21:30
國璽樓：8:00 ~ 23:00

下午 5:04

圖表 34 平日情況的「項目(具體時間)」問題類型範例
(研究者自行整理)

已讀
下午 11:27 圖書館假日時候的開館時間為何？



Saturday 各館的開放時間如下：
濟時樓：9:00 ~ 18:00
公博樓：9:00 ~ 18:00
國璽樓：8:00 ~ 23:00

Sunday 各館的開放時間如下：
濟時樓：9:00 ~ 18:00
公博樓：不開放
國璽樓：8:00 ~ 23:00

下午 11:27

已讀
下午 11:28 週六開放嗎？



Saturday 各館的開放時間如下：
濟時樓：9:00 ~ 18:00
公博樓：9:00 ~ 18:00
國璽樓：8:00 ~ 23:00

下午 11:28

圖表 35 假日情況的「項目(具體時間)」問題類型範例
(研究者自行整理)

已讀
下午 5:09 疫情期間圖書館何時開？



因短期疫情持續升高，雙北已提升至三級警戒，依防疫規定，圖書館屬應關閉場所，本館緊急於 5/15 中午 12:30 起閉館。若開放時間及借還書相關服務有變動，請隨時留意圖書館網站公告，不便之處，敬請見諒。
<http://web.lib.fju.edu.tw/chi/news/20210517>

下午 5:09

圖表 36 疫情情況的「項目(具體時間)」問題類型範例
(研究者自行整理)

3.4.3.4.2 「持續時間」問題類型

關於這種詢問持續時間的問題類型，其疑問詞短語通常會出現「多長時間」、「多少時間」、「多久」...等字詞，而且問句中必定至少會有一個詢問目標以及一個動作，並以動詞或名詞這種富含語意的詞性呈現。

由於那些能夠代表一個句子語意重心之詞性的字詞已在意圖辨識階段中被納入考量，並反映在相似度排序後的意圖集群中，因此在最終規劃輸出內容時不需要再做進一步的邏輯判斷處理，即可在找到對應的意圖集群後，直接從問答語料中輸出相應的解答(見圖表 37 和圖表 38)。

已讀
下午 5:20

預約書到館多久沒去借，會被取消該預約書的借書權利？



預約書到館將保留5天，並以電子郵件方式通知讀者，如逾保留期限未辦理借出，則將圖書上架或轉給下一位預約者。至於未留電子郵件帳號之讀者，則請自行上網查閱個人借閱狀況是否預約書已到，因此建議讀者登錄穩定之e-mail address，以便圖書館通知，而不致損失借書權利。

下午 5:20

圖表 37 「持續時間」問題類型的預約書借閱範例
(研究者自行整理)

已讀
下午 5:33

論文上傳完成後，需要多久時間得知審核結果？



論文上傳後，隔日起2個工作天內會進行審核，若通過審核，或是有問題待修改，皆會寄發E-mail信件通知，您也可登入論文提交系統查詢審核狀況。若上傳3-5天之後仍然沒有得到任何訊息回應，請與圖書館聯絡(流通櫃台#2682)。

論文上傳繳交程序完成後，圖書館人員會在三個工作天內寄出email通知，以下任一方式皆可得知審核結果：1)請至個人電子郵件信箱查看審核結果，若是收到核准的通知，便可辦理後續離校手續；若繳交論文電子檔有問題，亦會以e-mail告知疏漏之處，也請研究生參照通知內容配合修正。2)可從學位論文查詢系統搜尋個人論文是否可以搜尋到，若是，表示已通過審核，即可辦理後續離校手續。

下午 5:34

圖表 38 「持續時間」問題類型的論文審核範例
(研究者自行整理)

3.4.3.4.3 「項目(規則)」問題類型

關於這種詢問規則相關的問題類型，其疑問詞短語經常出現「規定」、「限制」、「資格」、「為什麼不能」...等字詞。這類問題類型的處理方式與「持續時間」問題類型雷同，因為那些富含語意的字詞已囊括在意圖集群的相似度計算和排序上，所以在最終規劃輸出內容時不需要再做進一步的邏輯判斷處理，即可在找到對應的意圖集群後，直接從問答語料中輸出相應的解答，如圖表 39 和圖表 40。

已讀
下午 6:17 贈書給學校有何限制?



圖書館為顧及贈書之適用性，將予以篩選，以確定是否收藏，若館藏已有相同書籍或不適本館收藏者，將轉贈他館或校內師生。不予收藏的項目如下：1. 本館已有館藏。2. 違反著作權法之規定者。3. 殘缺不全之套書。4. 書況不佳、破損，有畫線、註記、眉批。5. 內容過時或涉及色情、暴力。6. 教材練習本。7. 與本館館藏政策不符者。贈書收取時間及地點：1. 週一至週五 8:30-12:00 13:00-16:50 採編組。2. 其他時間 總館流通櫃台及分館開館時間均可代收。

下午 6:17

圖表 39 「項目(規則)」問題類型的贈書限制範例
(研究者自行整理)

已讀
下午 6:36 為什麼不能續借?



可能已被他人預約，或續借次數已滿。

下午 6:36

圖表 40 「項目(規則)」問題類型的續借問題範例
(研究者自行整理)

不過，面對與資格有關的問題，需要從問句中進一步擷取使用者所欲詢問的身分為何，再進行邏輯判斷處理並回覆，像是在借閱資格規定上圖表 41 所詢問的對象為大學生、圖表 42 是關於研究生、圖表 43 則為教師。

已讀
下午 7:02 大學生借書相關限制?



大學部學生借閱總數以三十冊為限，借期為二十八日；無人預約時得續借一次。

下午 7:02

圖表 41 「項目(規則)」問題類型的大學生借閱資格範例
(研究者自行整理)

已讀
下午 7:11 研究生的借期規定



研究生借閱總數以四十冊為限，借期為四十二日；無人預約時得續借一次。研究生自入學第二年起，因撰寫學位論文之需而提出申請者，得辦理延長借書。延長借書之借期為六十日；無人預約時得續借一次。

圖表 42 「項目(規則)」問題類型的研究生借閱資格範例
(研究者自行整理)

已讀
下午 7:00 老師身分的借閱規定為何？



本校教師借閱總數以七十五冊為限，借期為一百二十日；無人預約時得續借一次。教師以其研究計畫專案經費購買之圖書，得辦理「專案借書」，借期至該計畫結束為止，不受第四款冊數及借期之限制。

圖表 43 「項目(規則)」問題類型的教師借閱資格範例
(研究者自行整理)

3.4.3.4.4 「地點」問題類型

關於這種詢問地點相關的問題類型，其疑問詞短語經常出現「什麼地方」、「什麼地點」、「哪裡」、「哪兒」、「何處」...等字詞，而這些字詞的詞性標注通常與「Ncd」位置詞或「Nc」地方詞有關，例如：哪裡(Ncd)、何處(Nc)。面對這類的問題，我們一樣能基於意圖辨識後的集群結果直接從問答語料中輸出相應的解答，如圖表 44、圖表 45 和圖表 46。

已讀
下午 7:24 何處可以得知圖書館開放時間及休館日？



輔大圖書館各館的開放時間，詳情請見：
<http://web.lib.fju.edu.tw/chi/intro/opentime>
國定及校定假日特殊開放時間：<http://web.lib.fju.edu.tw/chi/news/20200915>

下午 7:24

圖表 44 「地點」問題類型的圖書館開館範例
(研究者自行整理)

已讀
下午 7:27 哪裡可找到歷屆研究所考古題？



本館經授權之考古題僅提供紙本試題公開陳列，不開放線上下載。公開陳列之紙本試題皆置於本館四樓影印室（服務時間），請至本館考古題網頁查詢所需考題是否公開，再行決定是否前往利用。

下午 7:27

圖表 45 「地點」問題類型的研究所考古題查找範例
(研究者自行整理)

已讀
下午 7:31 接到預約書領取通知，要至何處領取？



請您於收到本館'預約到館通知'後，攜帶學生證或教職員工證，直接至本館流通櫃檯辦理借閱手續即可。

館際合作申借之圖書或文件到館時，本館將以申請者所登錄的E-mail或電話通知，讀者於收到通知之後到圖書館辦理繳費及取件。若因故無法提供圖書或文件資料之申借服務時，本館亦會以E-mail或電話通知。

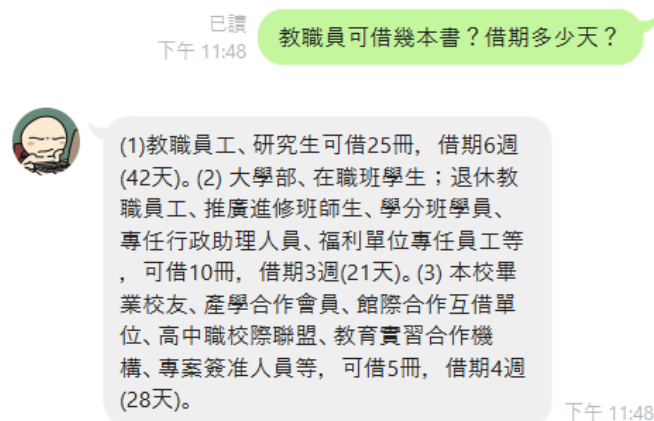
下午 7:31

圖表 46 「地點」問題類型的預約書領取範例
(研究者自行整理)

3.4.3.4.5 「數量」問題類型

關於這種詢問數量相關的問題類型，其疑問詞短語中往往有「多少」、「幾」...這類字詞，其詞性標注通常為「Neqa」數量定詞或「Neu」數詞定詞，而且這些字詞的後面必定會加上一個詞性標注為 Nf 的量詞。例如：多少(Neqa)本(Nf)、幾(Neu)冊(Nf)、幾(Neu)個(Nf)。

面對這類的問題，我們一樣能基於意圖辨識後的集群結果直接從問答語料中輸出相應的解答，如圖表 47 所示。



圖表 47 「數量」問題類型的借書範例
(研究者自行整理)

3.4.3.4.6 「原因」問題類型

關於這種詢問原因相關的問題類型，其疑問詞短語中往往有「為何」、「為什麼」...這類字詞，其詞性標注為 D 的副詞，而這些字詞多半置於問句開頭附近的位置。面對這類的問題，我們一樣能基於意圖辨識後的集群結果直接從問答語料中輸出相應的解答，如圖表 48 所示。

已讀
上午 12:15

為何找不到我要的書？



查詢線上目錄顯示仍有館藏，書架上卻找不到書，可能原因是：1. 本書剛歸還，還沒上架，可能在服務台或書車上，歡迎前往選閱；2. 本書是新書，在新書展示架上，歡迎借閱；3. 其它讀者已捷足先登，選了本書，可能在館內閱讀，還沒辦理借出；4. 各種類型的資料分別典藏在不同的樓層，請留意線上目錄查詢結果的館藏地點說明；5. 本書被放錯架位了；6. 最不幸的是... 本書已遺失..... 若您找不到所要的書，請洽服務台由工作人員協尋。

讀者於開館期間都可任意於館內取閱圖書資料，所以館藏狀況雖列為「在架」，但可能因為剛好有人在館內閱覽，或在各樓層的待歸架區、影印區或歸架書車上，工讀同學還來不及上架等因素，不在架上。服務檯備有「尋書申請單」，填好後交給服務檯人員，我們會另外安排人力蒐尋，並通知您尋書結果。

上午 12:15

圖表 48 「原因」問題類型的找不到書籍範例
(研究者自行整理)



3.4.3.4.7 「方法」問題類型

關於這種詢問方法相關的問題類型，其疑問詞短語的類型非常多樣，可能出現如下三種問法：其一，以「如何」作為問句起始的疑問詞，例如「如何辦理」、「如何申請」、「如何利用」、「如何得知」...等慣用問語；其二，含有「怎麼」的疑問詞，例如「該怎麼辦」、「要怎麼做」...等疑問詞短語；其三，含有「如何」的疑問詞，例如「該如何處理」、「要如何做」...等疑問詞短語。

面對這類的問題，我們一樣能基於意圖辨識後的集群結果直接從問答語料中輸出相應的解答，如圖表 49、圖表 50 和圖表 51 所示。

已讀
上午 12:00 找不到我要的書，怎麼辦？



1.確認館藏目錄顯示的圖書館狀態「在書架上」2.確認圖書狀態「在書架上」，但找不到書，請填寫「協尋未在架上書籍」。

1.其他讀者正在使用或未歸架：可能在閱覽桌上、影印區、書車上或其他讀者手上。2.上錯架：可在期刊架前後左右上下大略找找看。3.送裝訂中：如果是已到館的期刊，整卷、整年或部份完整而連續的卷期都不在架上時，有可能是送裝訂。4.館藏無此卷期：請利用館藏目錄確認。如仍有限期期刊利用問題，歡迎逕洽2樓參考諮詢服務櫃檯，由館員代為尋找，一旦尋獲將立即告知。

上午 12:00

圖表 49 「方法」問題類型的找不到書籍範例
(研究者自行整理)

已讀
上午 12:17 如何申請研究小間？



詳見研究小間申請方式，寒暑假可申請。

總圖書館討論室自2018年3月1日開始可以線上借用，只要您是本校教職員工學生，人數3至8人，以本校計中的帳號及密碼登入「討論室線上借用系統」，選取要使用的場地及時段，並填寫使用人數及使用者姓名，即完成線上預約申請，歡迎您多加利用。詳細借用步驟及說明請見本館首頁 > 服務項目 > 場地及設備 > 討論室借用。

上午 12:17

圖表 50 「方法」問題類型的空間申請範例
(研究者自行整理)

已讀
上午 12:20 學生證遺失，該如何處理？



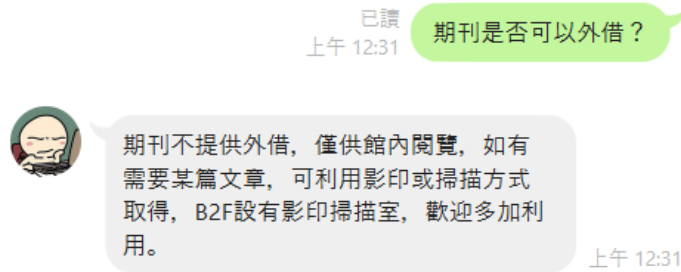
學生證確定遺失了，請儘速到教務處註冊組辦理補發事宜，同時為避免被冒用借書，請通知圖書館辦理借書證失效。

圖表 51 「方法」問題類型的證件遺失處理範例
(研究者自行整理)

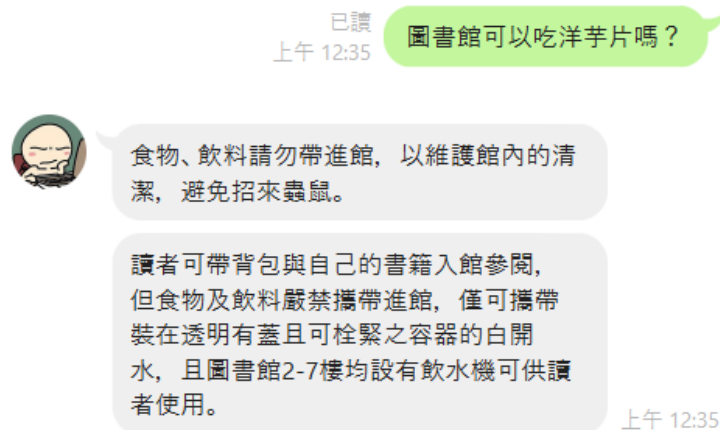
3.4.3.4.8 「是否」問題類型

關於這種詢問是否相關的問題類型，其疑問詞短語的類型也非常多樣，可能出現如下三種問法：其一，含有「是否」或「可否」的疑問詞，例如「是否」、「是否有」、「是否可以」、「是否提供」、「可否」、「可否用」...等疑問詞短語；其二，含有「有」或「可以」的疑問詞，例如「有...嗎」、「可以...嗎」、「可以嗎」、「可以」...等疑問詞短語；其三，含有「可不可以」的慣用問語。

面對這類的問題，我們一樣能基於意圖辨識後的集群結果直接從問答語料中輸出相應的解答，如圖表 52 和圖表 53 所示。



圖表 52 「是否」問題類型的期刊外借範例
(研究者自行整理)



圖表 53 「是否」問題類型的圖書館飲食範例
(研究者自行整理)

另外，針對查詢圖書館的館藏中是否有某本書，其處理方式會有些許不同。我們會先為訓練語料庫新增幾筆關於「查詢館藏」意圖集群的訓練語料(見圖表 54)，並取得該意圖集群的字詞，以便用於幫助過濾讀者問句中與查詢目標無關的字詞(如：圖書館、有、的、書、嗎、?)。接著，才把過濾後的查詢目標丟到輔大館藏 API 並返回其檢索結果連結，如圖表 55 所示。

ID String	Question String	Answer String	Category String
"1"	"我想找館藏"	"https://library.lib.fju.edu.tw"	"找館藏"
"2"	"我想找書"	""	"找館藏"
"3"	"我想找書，請問圖書館有嗎?"	""	"找館藏"
"4"	"請問圖書館有這本書嗎?"	""	"找館藏"
"4"	"圖書館有這本書嗎"	""	"找館藏"
"5"	"這本書有在圖書館的館藏裡嗎"	""	"找館藏"
"6"	"我想找館藏"	""	"找館藏"
"7"	"館藏有這本書嗎"	""	"找館藏"

圖表 54 「查詢館藏」意圖集群的訓練語料範例
(研究者自行整理)



圖表 55 「是否」問題類型的查詢館藏範例
(研究者自行整理)

3.4.3.4.9 「定義」問題類型


關於這種詢問定義相關的問題類型，其疑問詞短語通常會出現含有「包含哪些」、「哪些」、「有哪些」、「有什麼」、「是什麼」...等疑問詞。

面對這類的問題，依據所欲查找的資料來源區分，處理方式會有兩種：

其一，針對館務相關的問題，我們一樣能基於意圖辨識後的集群結果直接從問答語料中輸出相應的解答，如圖表 56 所示。

已讀
上午 12:41

電子資料庫有哪些？



請您由圖書館首頁 → 研究資源 → 點選「資料庫」，於左側的「資料類型」選擇「報紙」即可列出本館目前可查詢報紙資源之資料庫清單。各資料庫收錄的報刊內容與年代、本校訂購的範圍，請依本館資料庫系統網頁為準。以下簡介：**【臺灣地區】**1. 聯合知識庫：聯合報系八種報紙(聯合報、經濟日報、民生報、聯合晚報、星報、Upaper、美洲世界日報、歐洲日報)之線上全文。除民生報以外，其餘均自創刊迄今完整收錄。2. 臺灣新聞智慧網：收錄國內12種主要報紙之索引摘要，線上全文部份本館僅購買聯合報、聯合晚報、經濟日報、中國時報、中央日報，詳細收錄年份請見資料庫網頁說明，其餘報紙僅能閱讀標題及摘要。3. 其他：收錄中時報系（中國時報、工商時報、中時晚報）的知識贏家、收錄「自由時報」之線上全文的自由新聞網、收錄國民政府於中國大陸時期的官方媒體「中央日報」的中央日報全文影像資料

圖表 56 「定義」問題類型的資料庫類型範例
(研究者自行整理)

其二，針對查找維基百科知識庫的學科知識問題，我們會先為訓練語料庫新增幾筆關於「查詢維基百科」意圖集群的訓練語料(見圖表 57)，取得該意圖集群的字詞來過濾讀者問句中與查詢目標無關的字詞(如：圖書館、有、的、書、嗎、?)，再將查詢目標丟到維基百科知識庫並返回其條目的摘要說明和連結，如圖表 58 和圖表 59 所示。

ID String	Question String	Answer String	Category String
"1"	"請問是什麼?"	""	"查Wikipedia"
"2"	"我想查在維基百科是啥意思?"	""	"查Wikipedia"
"3"	"查wiki"	""	"查Wikipedia"
"4"	"查wikipedia"	""	"查Wikipedia"
"5"	"在維基的定義?"	""	"查Wikipedia"
"6"	"甚麼意思?"	""	"查Wikipedia"
"7"	"我想查這個術語的解釋"	""	"查Wikipedia"

圖表 57 「查詢維基百科」意圖集群的訓練語料範例
(研究者自行整理)

已讀
上午 12:44 核酸檢測是什麼？



核酸檢測是一種用於檢測遺傳物質 (RNA或DNA) 上的核酸序列的技術，該技術通常用於檢測和識別某些生物體，例如病毒或細菌。由於遺傳物質的數量通常非常少，不易檢測，所以在核酸檢測時需要透過聚合酶鏈式反應等方式複製遺傳物質，這種核酸檢測又稱為核酸擴增檢測。

<https://zh.wikipedia.org/wiki/核酸檢測>

上午 12:44

圖表 58 「定義」問題類型的查詢核酸檢測範例
(研究者自行整理)

已讀
上午 12:57 我想查meme



，又譯媒因、覓母、米姆、瀰等。目前比較公認的定義是透過模仿在人與人之間傳播的思想、行為或風格，通常是為了傳達模因所代表的特定現象、主題或意義。這個詞是在1976年，由理查·道金斯在《自私的基因》一書中所創造，將文化傳承的過程，類比成做生物學中的演化繁殖規則 (有共同先祖、隨著環境改變進化、優勝劣汰等等)。

<https://zh.wikipedia.org/wiki/meme>

上午 12:58

圖表 59 「定義」問題類型的查詢 meme 範例
(研究者自行整理)

綜合以上建置過程所述，本研究先是利用維基百科條目、圖資領域慣用語以及高出現頻率的共現 N 字詞(N-gram collocations)來改善自然語言處理技術的中文斷詞，以期更好地理解中文問句的語意。接著，透過一連串針對訓練語料的分析，挖掘出語料本身的結構，並利用這些觀察而得的模式(pattern)幫助後續在輸出回覆階段的內容規劃。

建構一個基於意圖辨識的圖書館參考諮詢機器人的過程中，依序經過語料的蒐集、分析、預處理以及功能建置的步驟，使用了 TF-IDF、DBSCAN 和 Word2Vec 三種演算法，分別幫助特徵詞加權、聚類相似意圖問法以及問句的同義詞查詢擴展。而最終的內容規劃輸出，是基於意圖辨識與排序的結果，查找對應的資料來源(圖書館常見問答語料、輔大館藏查詢 API、維基百科知識庫)，並搭配觀察到的語法結構模式(pattern)，來進行邏輯判斷處理。

第四章 系統分析與評估

4.1 評估方法說明

有鑑於讀者問題的問法相當多元且詢問意圖複雜難以預料，而且經常隨時事或時間而改變詢問的角度與方向，較難訂出所欲回答的意圖集群集合，故而在實驗成效評估上本研究僅針對受試者提出的問題採認心理學中經常使用的態度量表——李克特(Likert)量表——進行評量(Likert, 1932)，試圖了解系統輸出的意圖集群辨識結果是否與讀者真正的詢問意圖相關，以便得知本系統能否正確理解讀者的問題。更進一步地，針對藉由邏輯判斷處理而產生的系統答覆進行評價，以便知曉本系統能否切中受試者的資訊需求。

為取得問法多元、詢問意圖複雜且難以預料的讀者問題(測試文本)來作為本圖書館參考諮詢機器人之系統分析與評估的依據，乃由受試者們提供以自身方式描述的問題，總計 30 個。這 30 個問題中，其中有 18 個圖書館館務相關問題，6 個為圖書館館藏查詢相關，剩下的 6 個屬於詢問學科相關知識。測試語料彙整完成後，接著是將讀者問題(測試文本)輸入研究所建置的 Chatbot 中，獲取經由相似度計算並排序後的意圖集群辨識結果，而辨識結果包含所屬意圖集群的原始訓練語料樣本以及系統回覆結果。

接下來，本系統的評估將會基於李克特 (Likert) 五點量表設計問卷，針對每筆測試語料的排序後意圖集群辨識結果，分別取得各意圖集群之「原始訓練語料樣本」和「系統回覆結果」兩部分來進行主觀評價分析，並藉此探討本系統產生的意圖判斷結果是否與讀者認知一致，甚至更深入地分析與探討當辨識結果不如預期時的可能原因為何。

針對「原始訓練語料樣本」和「系統回覆結果」兩部分結果所欲探討的評估目標以及所採用的做法，以下詳述：

1. 原始訓練語料樣本

評估目標：

- (1) 了解「讀者問題(測試文本)」與「意圖集群的原始訓練語料樣本」之間的相似、相關或符合程度，從而得知系統能否正確理解讀者問題，即具有相似的意圖。
- (2) 搭配系統輸出的意圖集群一起觀察，若強烈相關、尚可接受的意圖集群與系統輸出的意圖集群重疊或相符程度較高，則代表系統能夠篩選出具有一定相關程度的意圖集群。

評估作法：

針對系統辨識後所屬意圖集群之原始訓練語料樣本的部份，讓受試者主觀判定並勾選測試文本與原始訓練語料樣本之間的「相關程度」，從而得知系統輸出的集群結果是否與問題的真正意圖相關。

2. 系統回覆結果

評估目標：

- (1) 了解「系統回覆結果」是否有滿足到「讀者的資訊需求」。
- (2) 比較系統輸出的順序與讀者認為的順序，從而得知系統排序結果與讀者預期結果之間的差異。也就是說，若受試者認為排名在前的系統回覆與「主題相關」且「完全滿足」或「大部分滿足」，即代表本系統的答覆內容符合受試者預期。

評估作法：

針對系統辨識後所屬意圖集群之系統回覆的部份，讓受試者依據她/他的主觀認知將系統回覆結果依據資訊需求的滿足程度做出排序，從而得知系統輸出的回覆結果是否能解答讀者問題，並且為受試者所接受。

4.2 測試語料

在受試者組成方面，本研究邀請輔仁大學的學生以及教職員共計 10 人，這 10 位受試者皆具備圖書資訊領域的專業背景，其中 3 位為從事圖資領域教學或相關事務的教職員，剩餘 7 位為圖書資訊學系各年級的大學部學生。接下來，將向受試者們蒐集 30 筆依據其自身敘述方式而得的測試語料，如圖表 60 所示。受試者的問題範圍，包含館務常見問題、館藏查詢問題以及學科知識問題。大致上，可知其細部的詢問範圍，涵蓋了開館、入館、閉館/逾期、歸還/逾期、借閱、連線、薦購、捐贈、遺失、設備安裝、操作使用、其他問題、館藏資源查詢以及學科知識查詢這 14 個方面。

詢問範圍	編號	問題描述	範圍細分
館務	1	疫情期間有開放入館嗎？	開館
	2	為何我刷卡後依舊不能進入圖書館？	入館
	3	飲料可以入館嗎？	
	4	閉館日為什麼也算在逾期時間內？	閉館/逾期
	5	為什麼已經還書了還被通知逾期？	歸還/逾期
	6	架上沒有我要的書...	借閱
	7	為什麼最新一期的雜誌都不能外借？	
	8	我想知道借書的規定	
	9	校友可以借書嗎？	
	10	可以借到別的學校的書嗎？	
	11	設定校外 VPN 連線	連線
	12	可以幫忙買書嗎？	薦購
	13	想要捐書直接給櫃檯嗎？	捐贈
	14	借的書弄不見了，怎麼辦	遺失
	15	我有東西在圖書館裡弄丟了，該去哪裡找？	
	16	圖書館電腦可不可以自行安裝其他軟體？	設備安裝
	17	還書箱如何使用？	操作使用
	18	有人用包包佔位子該怎麼辦？	其他問題
館藏查詢	19	有沒有什麼推薦的推理小說？	館藏資源 查詢
	20	學習韓文的資源？	
	21	圖書館有解剖學的書嗎？	

	22	有沒有好看的電影原著小說？	
	23	我想找深度學習的資源	
	24	食譜放在哪？	
學科知識	25	網路迷因是甚麼意思？	學科知識 查詢
	26	meme 是什麼？	
	27	新冠肺炎是什麼？	
	28	資料科學的定義是？	
	29	我想查機器學習	
	30	RNA 的解釋	

圖表 60 測試語料
(研究者自行整理)

接著，把這 30 筆測試語料鍵入 Chatbot 所得的意圖集群之排序結果，如圖表 61 所示。表格的左方，記錄著每筆測試語料的編號以及問題描述；而表格的右方，則記錄著每筆測試語料的意圖集群辨識結果，並且由相似度高到低排序，依序為 Rank1、Rank2、Rank3、Rank4 和 Rank5。其中，塗有淺色背景的部分，代表該意圖集群相應的解答被作為系統回覆輸出，反之則無。另外，關於每筆測試語料的詳細系統辨識結果，請見附錄二。

以編號 2 的測試語料為例，其問題描述為「為何我刷卡後依舊不能進入圖書館？」。經由相似度計算並排序後，系統輸出前 5 名相關的意圖集群，依序為集群 459、集群 385、集群 503、集群 171 和集群 502。再次經由「選擇相似度評分在平均以上的意圖集群作為輸出」的篩選後，將得到集群 459、集群 385 作為系統回覆的輸出。

編號	問題描述	系統的意圖集群辨識結果				
		Rank1	Rank2	Rank3	Rank4	Rank5
1	疫情期間有開放入館嗎？	387	140			
2	為何我刷卡後依舊不能進入圖書館？	459	385	503	171	502
3	飲料可以入館嗎？	387	143	329	386	503
4	閉館日為什麼也算在逾期時間內？	311	317	447	12	323
5	為什麼已經還書了還被通知逾期？	428				

6	架上沒有我要的書...	176	39	142	147	65
7	為什麼最新一期的雜誌都不能外借？	171	517	69	275	49
8	我想知道借書的規定	300	261	319	114	443
9	校友可以借書嗎？	19	221	175	256	333
10	可以借到別的學校的書嗎？	228	257	167	65	368
11	設定校外 VPN 連線	7	374	33	444	16
12	可以幫忙買書嗎？	212	362	437	258	296
13	想要捐書直接給櫃檯嗎？	254	315	467	100	359
14	借的書弄不見了，怎麼辦	315	396	285	395	32
15	我有東西在圖書館裡弄丟了，該去哪裡找？	32	247	530	38	368
16	圖書館電腦可不可以自行安裝其他軟體？	460	510	80	156	525
17	還書箱如何使用？	268	60	340	82	190
18	有人用包包佔位子該怎麼辦？	214	71	260	488	351
19	有沒有什麼推薦的推理小說？	43	195	357	372	360
20	學習韓文的資源？	452	270	233	38	54
21	圖書館有解剖學的書嗎？	530	159	226	442	325
22	有沒有好看的電影原著小說？	195	261	42	441	43
23	我想找深度學習的資源	530	452	32	438	368
24	食譜放在哪？	130	289	178	247	364
25	網路迷因是甚麼意思？	531				
26	meme 是什麼？	357				
27	新冠肺炎是什麼？	357				
28	資料科學的定義是？	531				
29	我想查機器學習	531				
30	RNA 的解釋	531				

圖表 61 測試語料的意圖集群排序結果
(研究者自行整理)

4.3 實驗結果

根據受試者針對測試語料與「原始訓練語料樣本」的相關程度評分結果，繪製成圖表 62 的熱圖(heat map)呈現。表格的左方，記錄著每筆測試語料的編號以及問題描述；而表格的右方，則記錄著受試者對於「原始訓練語料樣本」與「測試語料(受試者問題)」之間認為的相似、相關或符合程度的五種等級評價，分別為 5 (強烈相關)、4 (較為相關)、3 (尚可接受)、2 (較不相關)以及 1 (完全無關)。

對於每一筆測試語料的相關程度評價，表格內紀錄每一種排名(Rank1 至 Rank5)的每一種等級(強烈相關至完全無關)的人數統計，並依據人數的多寡由深(人數多)到淺(人數少)視覺化呈現。其中，塗有紅色背景的部分，代表經篩選後該意圖集群相應的解答被作為系統回覆輸出；而塗有藍色背景的部分，則屬於沒有被系統輸出解答的意圖集群。

從宏觀的角度來看圖表 62，可發現受試者對於這些「塗有紅色背景」作為系統回覆的意圖集群大多傾向認為原始訓練語料樣本與測試問題之間是「強烈相關」以及「較為相關」，而且評價為「強烈相關」的幾乎為 Rank1 的意圖集群；而對於那些「塗有藍色背景」沒有作為系統回覆的意圖集群則多半傾向認為原始訓練語料樣本與測試問題之間是「較不相關」甚至是「完全無關」的。

綜上所述，由於被作為系統回覆的意圖集群(塗有紅色背景)其訓練語料與讀者問題之間大多被認為具有「強烈相關」、「較為相關」，顯示本研究所建置的 Chatbot 能夠在相似度排序後，更進一步地篩選出那些具有強烈相關、較為相關的意圖集群(即與提問者真正意圖相關的意圖集群)作為系統輸出，因而推論得知本系統在理解讀者問題上，能夠正確辨識出具有相似意圖的集群。

搭配訓練語料來深究造成此結果的原因，推測可能是因為系統傾向把「含有測試語料裡出現過字詞」的意圖集群排序在前，從而造成「意思上真正相關卻使用不同字詞描述」的意圖集群其相似度不如「含有測試語料裡出現過字詞」的意圖集群。例如：以編號 8 的測試語料「我想知道借書的規定」為例，由於 Rank1 的訓練語料內含有與測試語料相同的用詞——「借書」，因此在相似度計算上會比 Rank2 含有「書」一詞以及 Rank3 含有「借閱」一詞的訓練語料之結果還要來得相似。

另外，也有一些例子(如：測試語料編號 1 和 12) 雖然在 Rank1 和 Rank2 都同樣含有測試語料裡出現過的字詞，但其在訓練語料出現的頻率影響了該字詞的權重，使得每一個意圖集群的相似度分數(Score)結果有了些微差異，因而反映在排名(Rank)上。

4.4.2 排名第一(Rank1)卻出現意圖相關最差、需求滿足最好

觀察圖表 66 後，可以發現測試語料編號 22、24、26、27 在「系統回覆」對於滿足讀者資訊需求的程度上主要分布在「主題相關，完全滿足(5)」和「主題相關，大部分滿足(4)」上，其中又以「主題相關，完全滿足(5)」的評價最多。然而，搭配圖表 67 一起觀察，則會看到那些系統回覆切合讀者資訊需求的意圖集群，其訓練語料的相關程度對於受試者而言非常兩極，有的受試者偏向認為「強烈相關」和「較為相關」；有的則偏向認為「較不相關」、甚至是「完全無關」。

編號	問題描述	Rank1					Rank2					Rank3					Rank4					Rank5				
		5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
22	有沒有好看的電影原著小說?	7	2	0	0	1																				
24	食譜放在哪?	6	4	0	0	0																				
26	meme是什麼?	8	2	0	0	0																				
27	新冠肺炎是什麼?	8	2	0	0	0																				

圖表 66 「系統回覆」對於滿足讀者資訊需求的程度，以測試語料編號 22、24、26、27 為例
(研究者自行整理)

編號	問題描述	Rank1					Rank2					Rank3					Rank4					Rank5				
		5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
22	有沒有好看的電影原著小說？	2	2	1	2	3	0	2	2	3	3	0	0	0	3	7	0	0	0	0	10	0	2	3	2	3
24	食譜放在哪？	2	2	1	1	4	0	2	1	1	6	0	0	1	0	9	0	0	0	1	9	0	0	1	0	9
26	meme是什麼？	2	3	0	0	5																				
27	新冠肺炎是什麼？	3	2	0	0	5																				

圖表 67 測試語料與「訓練語料」的相關程度評分，以測試語料編號 22、24、26、27 為例
(研究者自行整理)

若我們以測試語料編號 22 的「有沒有好看的電影原著小說？」為例，回歸訓練語料來看的話，可看到 Rank1 的訓練語料為「圖書館有沒有空中英語教室？」。對於此訓練語料的描述，推測受試者評價兩極可能的原因在於其依據的角度或理由為「句法結構相似」或是「詢問目標相似」。受試者的角度，若為「句法結構相似」的話，則傾向認為訓練語料與測試語料之間是「強烈相關」或「較為相關」的；反之，則傾向認為兩者之間是「較不相關」或「完全無關」的。

而此一測試語料之所以在「系統回覆」上能夠「主題相關，完全滿足(5)」讀者資訊需求是因為本系統在輸出系統回覆前，系統會過濾讀者問句中與查詢目標無關的字詞，再把擷取的查詢目標丟入館藏查詢 API，以取得檢索結果作為系統回覆輸出。

4.4.3 某些評價與訓練語料相關的意圖集群卻沒有輸出其系統回覆

觀察圖表 68 後發現，相較於有作為系統回覆之意圖集群(塗有紅色背景)，某些沒有作為系統回覆的意圖集群(塗有藍色背景)雖然在人數統計上其相關性稍差一些，但主要仍落於「強烈相關(5)」和「較為相關(4)」的等級。

我們推測造成某些評價與訓練語料相關的意圖集群卻沒有輸出其系統回覆的原因，是因為本研究採用的篩選公式「選擇相似度評分在平均以上的意圖集群作為系統回覆輸出」或許有些嚴苛，將門檻放得寬鬆一些有助於提高召回率 (Recall)。

編號	問題描述	Rank1				Rank2				Rank3				Rank4				Rank5								
		5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
2	為何我刷卡後依舊不能進入圖書館？	6	4	0	0	0	0	4	2	4	0	7	3	0	0	0	1	0	0	0	9	0	5	3	1	1
6	架上沒有我要的書...	9	0	1	0	0	5	5	0	0	0	0	0	0	1	9	0	0	1	2	7	1	2	3	4	0
13	想要捐書直接給櫃檯嗎？	6	4	0	0	0	0	1	0	2	7	0	0	0	2	8	2	4	4	0	0	6	3	1	0	0
14	借的書弄不見了，怎麼辦	8	2	0	0	0	6	3	1	0	0	0	1	2	3	4	0	0	0	2	8	0	0	1	4	5
15	我有東西在圖書館裡弄丟了，該去哪裡找？	7	3	0	0	0	7	3	0	0	0	0	0	1	4	5	0	2	5	2	1	0	0	0	0	10

圖表 68 測試語料與「訓練語料」的相關程度評分，以測試語料編號 2、6、13、14、15 為例
(研究者自行整理)

4.4.4 可能影響系統意圖識別結果的因素

一個能夠準確理解、判斷讀者詢問目的，進而擷取句子中的重要訊息，加以處理後再回覆的圖書館參考諮詢機器人，其建置過程的每一步驟都與最終的意圖辨識結果直接相關且環環相扣。

針對本研究中所建置的圖書館參考諮詢機器人，可能影響系統意圖識別結果的因素，如下所示：

1. 訓練語料的預處理

在自然語言的預處理中，「斷詞」為首重。「斷詞」是將一個句子以最小的語意單位來分割，藉著一個一個各自表達完整概念的字詞來讓聊天機器人理解其含意。因此，「斷詞」的準確度，直接影響聊天機器人能否正確解析讀者問題，進而賦予正確的詞性標注以及識別專有名詞。

而若是讀者在鍵入問題時出現打錯字的情況，或是使用與訓練語料差異極大的問法來詢問時，可能會對系統的意圖辨識結果產生誤差。

2. DBSCAN 意圖分群的成果

儘管選擇 DBSCAN 作為本系統意圖分群的理由相當充分，像是 DBSCAN 是以資料本身的特性藉由演算法自動分群、不需要事先確定集群數目、不受極端值(noise)影響、可藉由設定半徑範圍和群內最小資料數目來彈性決定集群的數量和大小。

然而，也因為「依據資料本身性質」此一特性，造成分群上對於「相似句法結構，但不同詢問目標之語料」的挑戰。從圖表 69 來看，可知意圖集群編號 38 主要在詢問「關於遺失東西該如何處理的問題」，而造成這樣分群結果的原因，可能是因為「遺失」、「如何處理」和「怎麼辦」這些詞的權重其影響力相較於其他字詞更為巨大，才使得它們被歸類為一群。

若加以細分，則可發現其詢問目標可進一步分為館藏、證件、圖書附件、圖書...等項目。這使得系統在辨識到與遺失相關的問題時，若找到此一意圖集群，需要更進一步地擷取讀者的詢問目標，仰賴良好的資訊擷取技術。

3. 系統篩選輸出回覆的門檻(Score)

由於本系統進一步「選擇相似度評分在平均以上」的意圖集群作為輸出，而非直接將原始的所有相關結果輸出，因此輸出門檻的高低將直接影響相應的意圖集群其回覆是否被篩選並作為系統回覆輸出。

Cluster 38 訓練語料	詢問目標
向圖書館外借的館藏遺失了，該如何處理？	館藏
學生證遺失怎麼辦？	學生證
遺失圖書館證件，我應如何處理？	證件
借書遺失該如何處理？若只遺失隨書所附之光碟呢？	書、光碟
圖書附件（如光碟）遺失如何處理？	圖書附件
資料損壞或遺失要怎麼辦？	資料
圖書遺失了如何處理？	圖書
如果借的圖書遺失了怎麼辦？	圖書
圖書遺失如何處理？	圖書
借閱圖書遺失怎麼辦？	圖書
向圖書館借閱的圖書不小心遺失了怎麼辦？	圖書
請問圖書如果找不到怎麼辦？	圖書
書丟(壞)了，怎麼辦？	書
書遺失了怎麼辦？	書
借的書找不到（遺失了）或破損了怎麼辦？	書
我在圖書館借的書丟了怎麼辦？	書
向圖書館借的書若遺失了，該怎麼辦？	書
如果向圖書館所借之書已遺失，該如何處理？	書

圖表 69 意圖集群編號 38 的訓練語料
(研究者自行整理)

第五章 結論與建議

本研究旨在基於維基百科和自然語言處理相關技術，建構一個兼具語料擴充性、不需要撰寫大量匹配規則以及能夠自動辨識讀者詢問意圖的圖書館參考諮詢聊天機器人，以期能緩解參考館員的工作量，並提供讀者即時的參考諮詢服務。本章將彙整前一章所得之實驗結果與觀察，並給予結論及未來研究之建議。

5.1 結論

5.1.1 系統建置概述

本研究建置的圖書館參考諮詢聊天機器人，主要能夠回答館藏查詢、學科知識以及圖書館館務方面的問題，其系統回覆分別對應輔大館藏查詢 API、Wptools 維基百科知識庫提取工具以及收錄到 MongoDB 的圖書館常見問答語料經查詢後回傳其結果。

而在建置的做法上，不同於制式以人工撰寫規則來對應系統接下來的動作與回答，本研究採用「讀者問題」與「意圖集群」之間的相似度匹配來自動辨識讀者詢問目的，不只降低人工撰寫規則的時間，更容許讀者以多元的問法詢問(在現有的意圖集群下新增訓練語料)，甚至允許藉由訓練語料的擴充(新增新的意圖集群並為其添加訓練語料)來增加系統可回答的問題範圍。

為達到自動辨識讀者詢問意圖並輸出回應的目標，我們先是蒐集全台圖書館的常見問答集作為本研究的訓練語料，並在訓練語料預處理的階段，透過「維基百科條目」、「圖資領域慣用語」和「高出現頻率的共現 N 字詞 (N-gram collocations)」的加入，讓 Ckriptagger 的中文斷詞結果更加準確。接著，利用 TF-IDF 特徵詞加權演算法對單詞進行權重調整，反映單詞之間具有重要程度之別，再加上 DBSCAN 分群演算法將具有相似句法結構的詢問意圖聚為一群，讓分群結果能夠保留訓練語料本身的特性，形成一群群的「意圖集群」，方便後續意圖判斷使用。

緊接著便是依序為 Chatbot 處理流程上所需的功能進行建置，對於讀者問題的處理，我們考量到可能有些字詞並不在訓練語料的範圍內，因此利用 Word2vec 詞向量演算法找尋相關或相似詞來為那些不存在的字詞進行查詢擴展。而擴展後的字詞會連同中文斷詞結果，會被用於計算與意圖集群之間的相似度分數 (Score)，評分在平均以上的意圖集群其相應的回答將作為最終給讀者的輸出回覆。當獲得篩選後的相關意圖集群，便開始規劃輸出給讀者的內容，這時需要自然語言處理技術的輔助，並視讀者查詢意圖決定該從何種資源或工具中獲取解答。

5.1.2 使用評估結果

整體而言，研究結果發現受測者大多認為經意圖辨識後所得之意圖集群其「訓練語料」與讀者問題(測試問題)之間是「強烈相關」和「較為相關」的，而且評價為「強烈相關」的幾乎為排名第一順位(Rank1)的意圖集群，表示本系統能將具有強烈相關、較為相關的意圖集群(即與提問者真正意圖相關的意圖集群)作為系統輸出，得知系統能夠正確辨識出具有相似意圖的集群。

另外，受試者對於系統的「輸出回覆」也大多認為與「主題相關且滿足資訊需求」，其中又以每一筆測試語料的 Rank1 評價最為集中，主要評價為與「主題相關，完全滿足」，顯示本系統第一筆輸出的答覆往往就能切合需求、解答問題。

綜上所述，經過受試者評價的驗證後，得出結論如下：此圖書館參考諮詢機器人能夠輸出那些真正與意圖相關的對應答覆，而且這些答覆往往與主題相關且滿足資訊需求。

5.1.3 相較於圖資領域其他中文聊天機器人的優勢

回顧與圖書館相關的中文聊天機器人，具代表性的例子包含中國清華大學圖書館的「小圖」(姚飛、紀磊、張成昱、陳武，2011)、台師大圖書館參考諮詢機器人(陳宜琳，2019) 以及台灣大學圖書館虛擬參考諮詢服務機器人(范蔚敏，2020)。以中國清華大學圖書館的「小圖」來說，其建置上雖然綜合了自然語言處

理技術、問句相似度檢索匹配、教學系統以及知識推理機制，但在問句的處理上仍是仰賴人工規則的撰寫以及依靠大量訓練語料的支撐；而台師大圖書館參考諮詢機器人，雖然也是利用訓練語料以及問句相似度檢索匹配進行建置，但在訓練語料的筆數上(僅 1540 筆)比起中國清大圖書館的「小圖」少了許多，而且在問句的處理上較為粗淺，主要藉由人工撰寫的關鍵詞匹配規則來判斷，無法良好地處理相同意圖但不同問法的問句；更別提使用 Google DialogFlow 對話平台進行建置的台大圖書館虛擬參考諮詢服務機器人，雖然它能夠藉由訓練語料的學習來預測使用者的意圖，但在對話內容的範圍和設計上卻無法保有彈性，因為其必須仰賴「人工」設計聊天機器人的對話內容、問句與答句的對應設定(Intent)、關鍵詞(Entities)的觸發機制。

反觀本研究建置的圖書館參考諮詢機器人，在降低人工撰寫規則的前提下，目前僅透過可蒐得之小規模的圖書館問答文本語料來實現基於意圖集群 (Intent) 的匹配，不只在對話內容的範圍和設計上保有彈性，我們還透過自然語言處理技術取代人工撰寫規則來擷取「詢問目標」並針對不同的意圖集群給予對應的檢索動作，甚至在此基礎上能夠輕易擴充訓練語料，有助於多元問法的納入，便於後續的系統維護與利用。

5.2 未來展望

以下針對本研究之實驗結果及研究過程中所觀察到的現象，提出幾點相關的未來研究方向與建議：

1. 提高對於錯別字和開放用語的容忍度

本研究的圖書館參考諮詢機器人沒有針對「錯別字」進行特殊處理，因此若讀者在問題輸入上打錯字的話，系統極有可能無法正確辨識讀者的詢問意圖，而造成答非所問的情況。此外，針對一些沒有被收錄至維基百科條目的開放用語，或許未來研究能夠從社群爬文著手，提高系統對於新興用語上的識別能力。

2. 對意圖集群內進行細分群

由於分群演算法是依據「資料本身的性質、特徵」進行聚類，因此可能會將「相似句法結構，但不同詢問目標之語料」也歸為一群，而造成該意圖集群內存在多種不同的詢問目標(見圖表 69)。若把該意圖集群加以細分，則有助於更細緻地回答特定詢問目標的問題，不過這同時也需要良好的資訊擷取與特徵比對技術。

3. 增加意圖集群內訓練語料的多樣性

本研究從網路上蒐集整理的圖書館常見問答集，重複性較高，較難反應讀者「多元」的諮詢問答。若是未來能夠蒐集讀者與系統之間的對話內容，或是加入更多線上參考諮詢的問答，將有助於提高訓練語料的多元與豐富性，使圖書館參考諮詢機器人「學習」到更多回應的問答，從而讓系統具彈性，甚至能夠針對描述較不全面的諮詢問題給予較好的回覆。

4. 改善答案呈現方式

針對答案回覆的做法，本研究是檢索自預存於訓練語料庫的圖書館常見問答、利用輔大館藏 API 查詢，以及使用 Wptools 維基百科知識庫提取工具取得條目的摘要和連結。

在後續研究或許可以針對不同問題類型(見圖表 31)，藉由自然語言處理以及機器學習技術，歸納不同答案類型(如：具體時間、持續時間、規則、地點、數量、原因、方法、是否和定義)的問法規則(pattern)，並且進一步對回覆內容生成技術進行改善。

5. 強化擴展訓練新議題的聊天功能

可以透過「增添新的意圖集群」並「為該集群新增訓練語料」的方式，來為此圖書館參考諮詢機器人擴展訓練新議題的聊天功能，更廣泛的回應讀者的問題。

希望以上的建議能有助於未來建構意圖式參考諮詢服務機器人之研究，提供讀者一個更佳之對話體驗。

附錄

附錄一 104 間圖書館的常見問答集蒐集名單

	館名	問題分類	類別總數	所屬等級
1	國家圖書館	Y	14	國家圖書館
2	桃園市立圖書館	Y	7	直轄市圖書總館
3	高雄市立圖書館	Y	13	直轄市圖書總館
4	新北市立圖書館	Y	12	直轄市圖書總館
5	臺中市立圖書館總館	Y	11	直轄市圖書總館
6	臺北市立圖書館	Y	5	直轄市圖書總館
7	臺南市立圖書館	Y	4	直轄市圖書總館
8	花蓮縣文化局圖書館	N		省轄市立圖書館
9	金門縣文化局圖書館	N		省轄市立圖書館
10	南投縣政府文化局圖書館	N		省轄市立圖書館
11	苗栗縣立圖書館	N		省轄市立圖書館
12	基隆市文化局圖書館	Y	2	省轄市立圖書館
13	新竹市文化局圖書館	Y	8	省轄市立圖書館
14	彰化縣立圖書館	N		省轄市立圖書館
15	澎湖縣圖書館	N		省轄市立圖書館
16	中山醫學大學圖書資訊處	Y	4	大學圖書館
17	中州學校財團法人中州科技 大學圖資處	Y	10	大學圖書館
18	中原大學張靜愚紀念圖書館	Y	11	大學圖書館
19	中國文化大學圖書館	Y	14	大學圖書館
20	中華學校財團法人中華科技 大學圖書館	Y	19	大學圖書館
21	環球學校財團法人環球科技 大學圖書資訊處	N		大學圖書館
22	元培醫事科技大學圖書館	Y	8	大學圖書館
23	元智大學圖書館	N		大學圖書館
24	文藻學校財團法人文藻外語 大學圖書館	Y	10	大學圖書館
25	世新大學圖書館	Y	3	大學圖書館

26	台北海洋學校財團法人台北海洋科技大學圖書館(校本部)	N		大學圖書館
27	台灣首府大學圖書館	N		大學圖書館
28	正修科技大學圖書資訊處	Y	4	大學圖書館
29	佛光大學圖書館	N		大學圖書館
30	吳鳳學校財團法人吳鳳科技大學圖書館	N		大學圖書館
31	私立大同大學圖書館	Y	7	大學圖書館
32	亞洲大學圖書館	Y	5	大學圖書館
33	明道學校財團法人明道大學承正圖書館	N		大學圖書館
34	東方學校財團法人東方設計大學圖書資訊處	Y	7	大學圖書館
35	東南科技大學圖書資訊處	N		大學圖書館
36	東海大學圖書館	Y	9	大學圖書館
37	長榮大學圖書資訊處-圖書館	Y	12	大學圖書館
38	南開科技大學圖書資訊服務處	N		大學圖書館
39	南臺科技大學圖書館	Y	10	大學圖書館
40	建國科技大學圖書館	Y	2	大學圖書館
41	美和學校財團法人美和科技大學圖書館	Y	9	大學圖書館
42	真理大學圖書館	Y	4	大學圖書館
43	高雄市立空中大學圖書館	N		大學圖書館
44	高雄醫學大學圖書資訊處	Y	7	大學圖書館
45	健行學校財團法人健行科技大學圖書館	Y	13	大學圖書館
46	國立中山大學圖書與資訊處	Y	4	大學圖書館
47	國立中央大學圖書館	Y	7	大學圖書館
48	國立交通大學圖書館	Y	15	大學圖書館
49	國立成功大學圖書館	Y	14	大學圖書館
50	國立宜蘭大學圖書資訊館	Y	10	大學圖書館
51	國立東華大學圖書資訊處	Y	19	大學圖書館
52	國立空中大學圖書館	N		大學圖書館

53	國立虎尾科技大學圖書館	N		大學圖書館
54	國立金門大學圖書館	Y	8	大學圖書館
55	國立屏東大學圖書館	Y	5	大學圖書館
56	國立政治大學圖書館	N		大學圖書館
57	國立高雄大學圖書資訊館	Y	4	大學圖書館
58	國立高雄餐旅大學圖書資訊處	Y	2	大學圖書館
59	國立陽明大學圖書館	Y	5	大學圖書館
60	國立雲林科技大學圖書館	Y	7	大學圖書館
61	國立勤益科技大學圖書館	Y	10	大學圖書館
62	國立嘉義大學圖書館	Y	10	大學圖書館
63	國立彰化師範大學圖書與資訊處	Y	2	大學圖書館
64	國立臺中科技大學民生校區圖書分館	N		大學圖書館
65	國立臺北大學圖書館(臺北校區)	N		大學圖書館
66	國立臺北商業大學圖書館(桃園校區)	N		大學圖書館
67	國立臺北護理健康大學圖書館	Y	6	大學圖書館
68	國立臺東大學圖書資訊館	Y	2	大學圖書館
69	國立臺南大學圖書館	N		大學圖書館
70	國立臺南藝術大學圖書館	Y	6	大學圖書館
71	國立臺灣大學圖書館	Y	18	大學圖書館
72	國立臺灣大學醫學圖書館	Y	4	大學圖書館
73	國立臺灣科技大學圖書館	Y	11	大學圖書館
74	國立臺灣海洋大學圖書暨資訊處	Y	9	大學圖書館
75	國立臺灣體育運動大學圖書資訊處(臺中)	N		大學圖書館
76	國立聯合大學圖書館	N		大學圖書館
77	國立體育大學圖書館	N		大學圖書館
78	崑山科技大學圖書資訊館	Y	10	大學圖書館
79	康寧學校財團法人康寧大學資圖中心(校本部)	Y	8	大學圖書館

80	淡江大學覺生紀念圖書館	Y	14	大學圖書館
81	逢甲大學圖書館	Y	11	大學圖書館
82	朝陽科技大學波錠紀念圖書館	Y	12	大學圖書館
83	開南大學圖書資訊處	Y	6	大學圖書館
84	慈濟學校財團法人慈濟大學圖書館	N		大學圖書館
85	慈濟學校財團法人慈濟科技大學圖書館	N		大學圖書館
86	聖約翰科技大學圖書館	N		大學圖書館
87	嘉藥學校財團法人嘉南藥理大學圖書資訊館	Y	14	大學圖書館
88	實踐大學圖書暨資訊處(高雄分館)	Y	19	大學圖書館
89	臺北市立大學圖書館	Y	12	大學圖書館
90	臺北醫學大學圖書館	N		大學圖書館
91	臺南家專學校財團法人臺南應用科技大學圖書館	Y	11	大學圖書館
92	輔仁大學學校財團法人輔仁大學圖書館	Y	8	大學圖書館
93	輔英科技大學圖書暨資訊處	Y	8	大學圖書館
94	銘傳大學圖書館	Y	4	大學圖書館
95	廣亞學校財團法人育達科技大學家聲紀念圖書館	N		大學圖書館
96	樹德科技大學圖書館	Y	19	大學圖書館
97	醒吾學校財團法人醒吾科技大學圖書館	Y	9	大學圖書館
98	靜宜大學蓋夏圖書館	Y	8	大學圖書館
99	龍華科技大學圖書館	Y	11	大學圖書館
100	嶺東科技大學圖書館	N		大學圖書館

附錄二 30 筆測試資料的系統意圖辨識結果

編號	1	
問題描述	疫情期間有開放入館嗎？	
系統輸出門檻 (Avg_score)	1.5221	
Rank1	意圖集群	Cluster 387
	相似度(Score)	1.689
	訓練語料	可以攜帶背包、書、食物或飲料入館嗎？可以攜帶食物入館嗎？請問可以攜帶袋子或背包入館嗎？請問可以攜帶飲料及食物入館嗎？換證讀者是否可攜帶背包、食物入館？請問可以攜帶水及食物進館嗎？請問可以攜帶背包入館嗎？
	系統回覆	讀者可帶背包與自己的書籍入館參閱，但食物及飲料嚴禁攜帶進館，僅可攜帶裝在透明有蓋且可栓緊之容器的白開水，且圖書館 2-7 樓均設有飲水機可供讀者使用。
	是否系統回覆	是
Rank2	意圖集群	Cluster 140
	相似度(Score)	1.605
	訓練語料	圖書館開放的時間？圖書館開放的時間為何？圖書館開放館室時間為何？
	系統回覆	因短期疫情持續升高，雙北已提升至三級警戒，依防疫規定，圖書館屬應關閉場所，本館緊急於 5/15 中午 12:30 起閉館。若開放時間及借還書相關服務有變動，請隨時留意圖書館網站公告，不便之處，敬請見諒。 http://web.lib.fju.edu.tw/chi/news/20210517
	是否系統回覆	是

編號		2
問題描述		為何我刷卡後依舊不能進入圖書館？
系統輸出門檻 (Avg_score)		2.5992
Rank1	意圖集群	Cluster 459
	相似度(Score)	4.189
	訓練語料	為何刷卡無法進入？刷卡無法進入，是否也不能借書？
	系統回覆	抱歉，造成您的不便！麻煩請您填表格，讓我們幫您建立門禁檔案。
	是否系統回覆	是
Rank2	意圖集群	Cluster 385
	相似度(Score)	2.614
	訓練語料	為何在進館時需要刷卡？為何一定要刷卡進館？
	系統回覆	本館為維護館舍、人員及圖書資料之安全，及便於統計本校各科系所學生使用圖書館之情形，特於入口處設有門禁管理系統，俾利隨時掌握讀者進館及在館情形，籲請讀者於進館時配合刷卡。
	是否系統回覆	是
Rank3	意圖集群	Cluster 503
	相似度(Score)	2.46
	訓練語料	為何我的證件無法刷卡入館？為何我的證件無法刷卡入館？
	是否系統回覆	無
Rank4	意圖集群	Cluster 171
	相似度(Score)	1.909
	訓練語料	我的資料為什麼不能續借？期刊資料為什麼不能外借？我的圖書為什麼不能續借？期刊資料為什麼不能外借？為何期刊不能外借？
	是否系統回覆	無
Rank5	意圖集群	Cluster 502
	相似度(Score)	1.824
	訓練語料	請問進入圖書館之有效證件為何？請問圖書館進館之有效證件為何？
	是否系統回覆	無

編號	3	
問題描述	飲料可以入館嗎？	
系統輸出門檻 (Avg_score)	1.8574	
Rank1	意圖集群	Cluster 387
	相似度(Score)	3.378
	訓練語料	可以攜帶背包、書、食物或飲料入館嗎？可以攜帶食物入館嗎？請問可以攜帶袋子或背包入館嗎？請問可以攜帶飲料及食物入館嗎？請問可以攜帶水及食物進館嗎？請問可以攜帶背包入館嗎？
	系統回覆	讀者可帶背包與自己的書籍入館參閱，但食物及飲料嚴禁攜帶進館，僅可攜帶裝在透明有蓋且可栓緊之容器的白開水，且圖書館 2-7 樓均設有飲水機可供讀者使用。
	是否系統回覆	是
Rank2	意圖集群	Cluster 143
	相似度(Score)	2.188
	訓練語料	食物、飲料可以帶進館嗎？食物可以帶進館嗎？食物可以帶進圖書館內嗎？食物可以帶進圖書館嗎？食物可以帶進圖書館內嗎？圖書館可以吃東西嗎
	系統回覆	食物、飲料請勿帶進館，以維護館內的清潔，避免招來蟲鼠。
	是否系統回覆	是
Rank3	意圖集群	Cluster 329
	相似度(Score)	1.303
	訓練語料	辦完離校手續後，還能持學生證入館借還嗎？辦完離校手續後，還能持學生證入館嗎？
	是否系統回覆	無
Rank4	意圖集群	Cluster 386
	相似度(Score)	1.220
	訓練語料	我是延畢生，是否可以借書？我是延畢生，是否可以持學生證刷卡入館及借書？
	是否系統回覆	無
Rank5	意圖集群	Cluster 503
	相似度(Score)	1.198
	訓練語料	為何我的證件無法刷卡入館？
	是否系統回覆	無

編號	4	
問題描述	閉館日為什麼也算在逾期時間內？	
系統輸出門檻 (Avg_score)	1.8189	
Rank1	意圖集群	Cluster 311
	相似度(Score)	2.361
	訓練語料	逾期的書罰款怎麼算？ 逾期的書罰款怎麼算？逾期的書罰款怎麼算？會通知我嗎？
	系統回覆	逾期罰款一天一冊 5 元，遇閉館日會扣除；累計至第 20 天後，每超過一天停權 2 天。在未還清罰款前，停止借書權利至繳清為止。
	是否系統回覆	是
Rank2	意圖集群	Cluster 317
	相似度(Score)	1.950
	訓練語料	本館開閉館時間？請問圖書館開閉館時間請問圖書館開閉館時間？
	系統回覆	公佈在圖書館首頁-->讀者服務-->服務時間與電話。寒暑假開放時間，另行公布在圖書館網頁-->最新消息中。
	是否系統回覆	是
Rank3	意圖集群	Cluster 447
	相似度(Score)	1.598
	訓練語料	如何續借？新到期日怎麼算？如何續借？新到期日怎麼算？ 如何續借？新到期日怎麼算？
	是否系統回覆	無
Rank4	意圖集群	Cluster 12
	相似度(Score)	1.595
	訓練語料	逾期滯還金如何計算？逾期滯還金如何計算？逾期滯還金如何計算？我借的書過期了，逾期滯還金如何計算？圖書館的圖書逾期滯還金怎麼計算？書逾期了，逾期滯還金是怎麼算的？
	是否系統回覆	無
Rank5	意圖集群	Cluster 323
	相似度(Score)	1.590
	訓練語料	請問高空大圖書館休館時間？請問高空大圖書館休館時間？
	是否系統回覆	無

編號	5	
問題描述	為什麼已經還書了還被通知逾期？	
系統輸出門檻 (Avg_score)	2.7093	
Rank1	意圖集群	Cluster 428
	相似度(Score)	3.257
	訓練語料	我今天已經拿書去還了喔，為什麼還會收到催還/到期通知？請代為查明情形，謝謝。我今天已經拿書去還了喔，為什麼還會收到催還通知？請代為查明情形，謝謝。
	系統回覆	系統的通知 mail 於每天凌晨寄發(可從 e-mail 的 sent 日期時間看出)，此封通知當日凌晨已寄出，而您是當天中午才還書的，所以當然會收到這封催還信。當您對借還紀錄有疑問時，可先利用本館網頁「借閱紀錄/續借/預約/薦購」功能，自行上線查看個人的借還書紀錄。
	是否系統回覆	是

編號	6	
問題描述	架上沒有我要的書...	
系統輸出門檻 (Avg_score)	1.6317	
Rank1	意圖集群	Cluster 176
	相似度(Score)	2.211
	訓練語料	資料未借出，可是架上找不到？找不到架上的書？館藏未借出，但架上找不到？資料顯示在書架上，可是架上找不到書？架上找不到我想借閱的書該怎麼辦？
	系統回覆	1. 可能有讀者正在使用中。2. 可能放錯位置，本館將利用整架時將其歸位〔希望讀者能夠隨手將閱畢的書刊歸位，或放置「不知如何上架」之書車上，由館方歸架〕。3. 若該書已外借，可辦理預約。4. 可至二樓櫃台申請協尋。
	是否系統回覆	是

Rank2	意圖集群	Cluster 39
	相似度(Score)	1.536
	訓練語料	在 SLIM 查到的期刊，為何架上找不到？圖書目錄查到的期刊，為何架上找不到？
	是否系統回覆	無
Rank3	意圖集群	Cluster 142
	相似度(Score)	1.494
	訓練語料	我的學生證為什麼沒有辦法刷卡進館？我的學生證為什麼沒有辦法刷卡進館？
	是否系統回覆	無
Rank4	意圖集群	Cluster 147
	相似度(Score)	1.470
	訓練語料	為什麼我沒有收到圖書館的通知單？沒有收到圖書館的通知單怎麼辦？
	是否系統回覆	無
Rank5	意圖集群	Cluster 65
	相似度(Score)	1.447
	訓練語料	我想借的書，台大圖書館沒有收藏，怎麼辦？可以幫我去其他學校圖書館借嗎？我想借的書，校內圖書館沒有收藏，怎麼辦？可以幫我去其他學校圖書館借嗎？
	是否系統回覆	無

編號		7
問題描述		為什麼最新一期的雜誌都不能外借？
系統輸出門檻 (Avg_score)		2.2523
Rank1	意圖集群	Cluster 171
	相似度(Score)	3.016
	訓練語料	我的資料為什麼不能續借？期刊資料為什麼不能外借？我的圖書為什麼不能續借？期刊資料為什麼不能外借？為何期刊不能外借？
	系統回覆	可能已被他人預約，或續借次數已滿。
	是否系統回覆	是
Rank2	意圖集群	Cluster 517
	相似度(Score)	2.588
	訓練語料	某本期刊的前一期在何處？某本期刊的前一期在何處？
	系統回覆	本館現期期刊係採鳩巢式陳列架，通常最新一期會在面板上陳列其餘當年內未裝訂的單本期刊，則必須掀開鳩巢的小門才能找到。過去幾年已裝訂期刊則放置期刊架上供讀者閱讀。
	是否系統回覆	是
Rank3	意圖集群	Cluster 69
	相似度(Score)	1.989
	訓練語料	為什麼我在校外不能從資料庫下載全文？不能看電子書？為什麼我在校外不能從資料庫下載全文？不能看電子書？
	是否系統回覆	無
Rank4	意圖集群	Cluster 275
	相似度(Score)	1.864
	訓練語料	期刊能否外借？期刊能否外借？過期雜誌和期刊能否提供外借？過期期刊能否外借？
	是否系統回覆	無
Rank5	意圖集群	Cluster 49
	相似度(Score)	1.805
	訓練語料	為什麼圖書館要換系統？圖書館為什麼忽冷忽熱？為什麼我要借的書圖書館都沒有？
	是否系統回覆	無

編號	8	
問題描述	我想知道借書的規定	
系統輸出門檻 (Avg_score)	1.6548	
Rank1	意圖集群	Cluster 300
	相似度(Score)	1.837
	訓練語料	我想去附近的大學圖書館借書可以嗎？我想去附近的大學圖書館借書可以嗎？
	系統回覆	本館有跟國立台南藝術大學及國立成功大學互有館際互借服務。
	是否系統回覆	是
Rank2	意圖集群	Cluster 261
	相似度(Score)	1.700
	訓練語料	如何知道某種期刊有沒有館藏？如何知道某種期刊有沒有館藏？想知道圖書館有沒有我要的書？如何知道圖書館有沒有我需要的電子資源？要如何知道某種期刊有沒有館藏？
	系統回覆	期刊依學科主題，分別放置於各館，各館期刊擺放位置請參見圖書館網頁 [http://web.lib.fju.edu.tw/chi/intro/floor] 找尋前，請您查詢本館館藏目錄 [http://140.136.208.1/]：已知期刊名稱：請直接以刊名查詢館藏目錄 [http://140.136.208.1/]。沒有確定的期刊名稱，只是想看某一類的期刊：先用本館館藏目錄的「關鍵字查詢」，輸入主題辭彙，並將「查詢範圍」設定為「期刊」。
	是否系統回覆	是
Rank3	意圖集群	Cluster 319
	相似度(Score)	1.682
	訓練語料	借閱資格及相關規定借閱相關規定？
	系統回覆	借閱冊數及期限如下：1. 大學部學生借閱總數以三十冊為限，借期為二十八日；無人預約時得續借一次。 2. 研究生借閱總數以四十冊為限，借期為四十二日；無人預約時得續借一次。 3. 研究生自入學第二年起，因撰寫學位論文之需而提出申請者，得辦理延長借書。延長借書之借期為六十日；無人預約時得續借一次。 4. 本校教師借閱總數以七十五
	是否系統回覆	是

		冊為限，借期為一百二十日；無人預約時得續借一次。 5. 本校職員工借閱總數以三十冊為限，借期為二十八日；無人預約時得續借一次。 6. 教師以其研究計畫專案經費購買之圖書，得辦理「專案借書」，借期至該計畫結束為止，不受第四款冊數及借期之限制。 7. 校友及退休人員借書以五冊為限，借期為二十八日，不得續借。 8. 其他具借閱資格者之借閱總數，以十冊為限，借期為十四日，不得續借。 借書不還者，除依第九條計算罰款外，情節重者應予三個月以上停權，如繳保證金者不予發還。
	是否系統回覆	是
Rank4	意圖集群	Cluster 114
	相似度(Score)	1.607
	訓練語料	我想知道自己借了那些書，何時到期？是否有逾期？我想知道自己目前借了哪些書，何時到期，是否有欠逾期滯還金？
	是否系統回覆	無
Rank5	意圖集群	Cluster 443
	相似度(Score)	1.448
	訓練語料	我如何知道所預約的書已經歸還？如何知道我的預約資料已經到館了？如何知道文獻傳遞申請的資料已經到館？
	是否系統回覆	無

編號	9	
問題描述	校友可以借書嗎？	
系統輸出門檻 (Avg_score)	1.7328	
Rank1	意圖集群	Cluster 19
	相似度(Score)	1.818
	訓練語料	繳交年費享有借書權利的校友，可以透過三校聯盟到台師大、台科大的圖書館借書嗎？繳交年費享有借書權利的校友，可以透過三校聯盟到台師大、台科大的圖書館借書嗎？
	系統回覆	校友及退休人員借書以五冊為限，借期為二十八日，不得續借。
	是否系統回覆	是
Rank2	意圖集群	Cluster 221
	相似度(Score)	1.768
	訓練語料	校友可以回學校圖書館使用電腦設備嗎？校友可以回學校圖書館嗎？校友可以在館內使用電腦設備嗎？校友可以回學校圖書館使用座位系統設備嗎？
	系統回覆	1. 已辦理校友借書證者可以憑證免費使用。2. 未辦校友借書證者以每小時一佰元計費。
	是否系統回覆	是
Rank3	意圖集群	Cluster 175
	相似度(Score)	1.710
	訓練語料	臨時學生證可以借書嗎？臨時學生證可以借書嗎？
	是否系統回覆	無
Rank4	意圖集群	Cluster 256
	相似度(Score)	1.710
	訓練語料	兼任老師可以借書嗎？兼任老師可以借書嗎？
	是否系統回覆	無
Rank5	意圖集群	Cluster 333
	相似度(Score)	1.657
	訓練語料	休學生可不可以憑學生證借書？準研究生可以借書嗎？研究所休學生可以借書嗎？休學生借書休學生或準研究生可以借書嗎？
	是否系統回覆	無

編號		10
問題描述		可以借到別的學校的書嗎？
系統輸出門檻 (Avg_score)		2.5235
Rank1	意圖集群	Cluster 228
	相似度(Score)	3.196
	訓練語料	逢甲的學生也可以去別的學校借書嗎？嶺東的教職員或學生，可以到別的學校借書嗎？
	系統回覆	本館與多所大學互為聯盟合作館，您可於平日或寒暑假，親自前往聯盟館借閱，詳細辦法，請參考本館與其它圖書館圖書互借服務說明。若不想親自前往，也可以透過全國文獻傳遞服務辦理。
	是否系統回覆	是
Rank2	意圖集群	Cluster 257
	相似度(Score)	2.918
	訓練語料	可以到其他學校借書嗎？可以去其他學校借書嗎？可否到到其他學校借書？
	系統回覆	可以到與本館有合作之圖書館借書，詳情可詢問流通櫃檯館員。
	是否系統回覆	是
Rank3	意圖集群	Cluster 167
	相似度(Score)	2.244
	訓練語料	預約總館的書，可以在其他分館取書嗎？
	是否系統回覆	無
Rank4	意圖集群	Cluster 65
	相似度(Score)	2.187
	訓練語料	我想借的書，台大圖書館沒有收藏，怎麼辦？可以幫我去其他學校圖書館借嗎？我想借的書，校內圖書館沒有收藏，怎麼辦？
	是否系統回覆	無
Rank5	意圖集群	Cluster 368
	相似度(Score)	2.072
	訓練語料	我是別的學校的學生，請問我可以帶自己的(NB)電腦去成大圖書館用網路找資料嗎？我是別的學校的學生，請問我可以帶自己的(NB)電腦去圖書館用網路找資料嗎？
	是否系統回覆	無

編號	11	
問題描述	設定校外 VPN 連線	
系統輸出門檻 (Avg_score)	2.7573	
Rank1	意圖集群	Cluster 7
	相似度(Score)	4.780
	訓練語料	設定校外連線 VPN 後網路就斷了？如何設定校外連線 VPN？
	系統回覆	依照〈如何設定校外連線 VPN？〉的步驟，結果網路斷掉。此時，有可能目前計中的 VPN 服務真的當機了。您可洽詢臺大計中確認目前 VPN 服務是否正常：(02) 3366-5022 或 3366-5023 / cchelp@ntu.edu.tw。若計中 VPN 服務正常，則有可能是：您的防毒軟體阻擋了 VPN 連線，請嘗試關閉防毒軟體。您目前所在的網路環境阻擋了 VPN 連線，請洽詢所在處的網管人員。若無網管人員，煩請洽詢計中：(02) 3366-5022 或 3366-5023 /cchelp@ntu.edu.tw。請明確描述您的操作步驟、遭遇的問題，以利技術人員判斷。您可進入「開始」→「所有程式」→「附屬應用程式」→「命令提示字元」，輸入指令「tracert 173.194.72.105」→鍵盤「Enter」，這是用來追蹤您的網路斷在哪個點。您可詢問網管人員或計中人員是否需要以下資訊或畫面，以利判斷問題。請檢查所在位置的網路是否正常連線（如無線網路 wifi 訊號、網路線等）。
	是否系統回覆	是
Rank2	意圖集群	Cluster 374
	相似度(Score)	3.098
	訓練語料	在校外連線資料庫查詢，需要做什麼設定？
	系統回覆	當您要在校外連線查詢，並不需要在電腦上做任何的設定，直接以借書證的帳號密碼登入即可。若確定帳號密碼無誤，但仍無法連線，請嘗試以下列方法連線：請試以備用連線清單上的資料庫網址連線。請利用電算中心提供之 SSL VPN 服務，以 VPN 連線後，再查詢圖書館之資料庫。
是否系統回覆	是	
Rank3	意圖集群	Cluster 33

	相似度(Score)	2.512
	訓練語料	校外連線時，未成功安裝 Pulse secure 程式？校外連線 VPN 的驅動程式 Pulse Secure 安裝失敗，怎麼辦？
	是否系統回覆	無
Rank4	意圖集群	Cluster 444
	相似度(Score)	1.745
	訓練語料	校友或休學學生可以使用校外連線服務嗎？請問校友或休學學生可以使用校外連線服務嗎？
	是否系統回覆	無
Rank5	意圖集群	Cluster 16
	相似度(Score)	1.652
	訓練語料	哪些身分可用 VPN？休（退）學、畢業、校友、校外人士是否可用？哪些身分可用 VPN？休（退）學、畢業、校友、校外人士是否可用？
	是否系統回覆	無



編號		12
問題描述		可以幫忙買書嗎？
系統輸出門檻 (Avg_score)		1.5001
Rank1	意圖集群	Cluster 212
	相似度(Score)	1.865
	訓練語料	可以多買一些熱門院線片嗎？可以多買一些熱門院線片嗎？可以多買一些熱門院線片嗎？
	系統回覆	依本館採購政策，在經費許可範圍下採購電影及影集以國內外經典名片、得獎、入圍或被提名之作品為原則。
	是否系統回覆	是
Rank2	意圖集群	Cluster 362
	相似度(Score)	1.715
	訓練語料	有些書借用的人多，借閱率高的書，是否可以多買幾本？有些借閱率高的書，是否可以多買幾本？
	系統回覆	圖書館圖書經費之應用必須合理化，以期能作更有效的應用。目前圖書館受限於經費及館藏空間，且希望能夠提供讀者更多種類的資訊，因此以不買複本為原則。
	是否系統回覆	是
Rank3	意圖集群	Cluster 437
	相似度(Score)	1.298
	訓練語料	我遺失的書在出版社或書局已經買不到了，怎麼辦？
	是否系統回覆	無
Rank4	意圖集群	Cluster 258
	相似度(Score)	1.283
	訓練語料	新書展示區的書可以借嗎？新書/暢銷書展示區的書可以外借嗎？
	是否系統回覆	無
Rank5	意圖集群	Cluster 296
	相似度(Score)	1.191
	訓練語料	可否請圖書館代為向他館借書？我是否可以請圖書館代為購書？圖書館可否代為買書？可否代為向他館借書？
	是否系統回覆	無

編號	13	
問題描述	想要捐書直接給櫃檯嗎？	
系統輸出門檻 (Avg_score)	1.3533	
Rank1	意圖集群	Cluster 254
	相似度(Score)	1.488
	訓練語料	如何贈書給圖書館?贈書方式請問圖書館是否接受贈書?如果可以,何時可贈書?圖書館接受贈書嗎?圖書館是否接受贈書?任何書贈送,圖書館均會典藏嗎?想要贈書給圖書館,要如何處理?圖書館是否接受贈書?任何書贈送,圖書館均會典藏嗎?圖書館接受哪些贈書?如何辦理贈書?圖書館是否接受贈書?任何書均可贈送嗎?請問圖書館有接受贈書嗎?
	系統回覆	(一) 如果到學校,可直接將書籍交給圖書館借還書櫃台。(二) 若不方便到學校,可直接將書籍郵寄到學校圖書資訊處採編組。
	是否系統回覆	是
Rank2	意圖集群	Cluster 315
	相似度(Score)	1.434
	訓練語料	我的書不見了怎麼辦?我跟圖書館借的書不見了,該怎麼處理?借閱的書不見了,該怎麼處理?借閱的光碟不見了,該怎麼處理?借的書不見了要如何處理呢?借的書不見了怎麼辦?我想捐書給圖書館,要怎麼處理?我借的書不見了,怎麼辦?
	系統回覆	當您不小心將圖書遺失時,請務必盡快到圖書館辦理圖書掛失,並購買相同書籍賠償。在您賠償圖書之前所產生的逾期滯留金仍會依規定計算。若遺失的圖書已絕版,無法購買,則需依相關辦法賠償,請詳閱「圖書館閱覽及借閱規則」(http://testweb.tf.edu.tw/sites/default/files/dli_files/圖書館閱覽及借閱規則.pdf)。
是否系統回覆	是	
Rank3	意圖集群	Cluster 467
	相似度(Score)	1.364
	訓練語料	討論室借用的時間已到期,如果要繼續使用可以直接續借嗎?討論室借用的時間已到,如果要繼續使

		用可以直接續借嗎？
	系統回覆	需先將鑰匙歸還後，借用人憑教職員證或學生證於一樓流通櫃台重新辦理借用手續。
	是否系統回覆	是
Rank4	意圖集群	Cluster 100
	相似度(Score)	1.358
	訓練語料	如何捐贈圖書資料給國家圖書館？如果我想要捐贈書刊資料給圖書館，要如何處理？如何捐贈圖書？我想捐贈圖書/資料？我如何捐贈圖書或其他資料給圖書館？如何捐贈圖書？捐贈圖書館的書刊資料，圖書館會如何處理？
	系統回覆	圖書館竭誠歡迎各界捐贈，只要符合本館館藏發展政策的書刊資料，均為納藏對象。舉凡古籍善本、抄本、手稿、絕版書、善書、族譜、藝術類圖書、研討會論文集、藝文海報、視聽資料，主題與東南亞研究、新移民研究、東亞研究相關之中外文著作，以及國外出版品（含大陸出版品）內容與臺灣相關者皆在歡迎之列。社會賢達如有藏書捐贈，歡迎來電告知。
	是否系統回覆	是
Rank5	意圖集群	Cluster 359
	相似度(Score)	1.121
	訓練語料	我想贈送書刊資料給圖書館，要如何處理？我想贈送書刊資料給圖書館，要如何處理？
	是否系統回覆	無

編號	14	
問題描述	借的書弄不見了，怎麼辦	
系統輸出門檻 (Avg_score)	1.9007	
Rank1	意圖集群	Cluster 315
	相似度(Score)	3.128
	訓練語料	我的書不見了怎麼辦？我跟圖書館借的書不見了，該怎麼處理？借閱的書不見了，該怎麼處理？借閱的光碟不見了，該怎麼處理？借的書不見了要如何處理呢？借的書不見了怎麼辦？我想捐書給圖書館，要怎麼處理？我借的書不見了，怎麼辦？
	系統回覆	當您不小心將圖書遺失時，請務必盡快到圖書館辦理圖書掛失，並購買相同書籍賠償。在您賠償圖書之前所產生的逾期滯留金仍會依規定計算。若遺失的圖書已絕版，無法購買，則需依相關辦法賠償，請詳閱「圖書館閱覽及借閱規則」 (http://testweb.tf.edu.tw/sites/default/files/dli_files/圖書館閱覽及借閱規則.pdf)。
	是否系統回覆	是
Rank2	意圖集群	Cluster 396
	相似度(Score)	1.693
	訓練語料	遺失書怎麼辦？請問遺失書怎麼辦？
	是否系統回覆	無
Rank3	意圖集群	Cluster 285
	相似度(Score)	1.643
	訓練語料	為什麼現期的期刊卻在架上找不到？請問期刊不在架上該怎麼辦？書找不到怎麼辦？找不到館藏期刊怎麼辦？我找不到我想借的書，怎麼辦？我想找的期刊，在架上卻找不到？請問現期期刊不在架上該怎麼辦？在架上找不到書，該怎麼辦？為什麼在架上找不到以前的期刊？請問期刊不在架上怎麼辦？期刊並不在架上？資料不在架上，遍尋不獲，該怎麼辦？我要借的書不在架上，怎麼辦？
	是否系統回覆	無
Rank4	意圖集群	Cluster 395
	相似度(Score)	1.522
	訓練語料	請問學生可借多少冊書？借期多久？ 請問學生可

		借多少冊書？借期多久？
	是否系統回覆	無
Rank5	意圖集群	Cluster 32
	相似度(Score)	1.516
	訓練語料	我的東西不見了，請問圖書館有撿到嗎？失物招領應該去哪裡找？我的東西不見了，請問圖書館有撿到嗎？失物招領應該去哪裡找？
	是否系統回覆	無

編號	15	
問題描述	我有東西在圖書館裡弄丟了，該去哪裡找？	
系統輸出門檻 (Avg_score)	2.611	
Rank1	意圖集群	Cluster 32
	相似度(Score)	4.072
	訓練語料	我的東西不見了，請問圖書館有撿到嗎？失物招領應該去哪裡找？我的東西不見了，請問圖書館有撿到嗎？失物招領應該去哪裡找？
	系統回覆	如果您的東西在圖書館遺失或忘記取回，可親洽各服務櫃檯詢問是否有人拾獲，亦可透過圖書館網站→服務項目→場地及設備→「失物招領」看看拾獲物清單是否有您的物品，如果有人拾獲物品至櫃檯，館員皆會於此系統建檔，方便讀者們查詢各分館、各樓層的遺失物。
	是否系統回覆	是
Rank2	意圖集群	Cluster 247
	相似度(Score)	2.460
	訓練語料	東西遺忘在館內，要去哪邊認領？東西遺忘在館內，要去哪邊認領？
	是否系統回覆	無
Rank3	意圖集群	Cluster 530
	相似度(Score)	2.456
	訓練語料	我想找館藏我想找書我想找書，請問圖書館有嗎？請問圖書館有這本書嗎？圖書館有這本書嗎這本書有在圖書館的館藏裡嗎我想找館藏館藏有這本書嗎放在哪
	是否系統回覆	無

	意圖集群	Cluster 38
	相似度(Score)	2.250
Rank4	訓練語料	<p>向圖書館外借的館藏遺失了，該如何處理？學生證遺失怎麼辦？圖書館的燈壞了，怎麼辦？圖書館沒有我需要的書和期刊時怎麼辦？圖書遺失了如何處理？借書遺失該如何處理？若只遺失隨書所附之光碟呢？如果向圖書館所借之書已遺失，該如何處理？圖書遺失如何處理？圖書附件（如光碟）遺失如何處理？資料損壞或遺失要怎麼辦？北醫沒有我要的期刊文獻，只有其他圖書館有，要怎麼辦？我要看的書學校沒有買？如果借的圖書遺失了怎麼辦？遺失物品要去哪裡找呢？向圖書館借的書若遺失了，該怎麼辦？我想借的書學校沒有買，怎麼辦？借閱圖書遺失怎麼辦？【分館】如何找私人遺失物品？向圖書館借閱的圖書不小心遺失了怎麼辦？書遺失了怎麼辦？遺失圖書館證件，我應如何處理？請問圖書如果找不到怎麼辦？請問圖書遺失了怎麼辦？向圖書館借的書遺失了，該如何處理？請問我在圖書館借閱的圖書遺失了，怎麼辦？我的借書證遺失了，該怎麼辦？我的東西在圖書館遺失了，怎麼辦？向圖書館借的書遺失了怎麼辦？借的書找不到（遺失了）或破損了怎麼辦？我在圖書館借的書丟了怎麼辦？如果買不到書，怎麼辦？書丟(壞)了，怎麼辦？圖書遺失怎麼辦？請問圖書遺失應如何處理？在圖書館遺失物品，該如何處理？若我借閱的圖書遺失了該怎麼辦？借書證件遺失了該怎麼辦？我向圖書館借的書遺失了，怎麼辦？學生證遺失可否借書？借書證遺失了，怎麼辦？若將圖書館所借的書遺失了，怎麼辦？學生證遺失了，還可以借書嗎？遺失圖書怎麼辦？如果只有遺失附件光碟呢？學校圖書館沒有我要的期刊該怎麼辦？圖書館沒有想看的書怎麼辦？證件遺失怎麼辦？如果不小心將所借之館藏遺失了該怎麼辦？我的私人物品在圖書館內遺失，該如何處理？我向圖書館借的互換借書證遺失了，怎麼辦？圖書館沒有我要找的資源，該怎麼辦？圖書館沒有我要的期刊，怎麼辦？向圖書館借的書遺失或破損，該如何處理？如果借閱的圖書遺失了，該如何</p>

		處理?借閱的書遺失了怎麼辦?所借的書遺失了,怎麼辦?借閱證遺失該怎麼辦?圖書遺失或損壞怎麼辦?圖書、視聽資料遺失怎麼處理?證件遺失並被冒借圖書怎麼辦?
	是否系統回覆	無
Rank5	意圖集群	Cluster 368
	相似度(Score)	1.812
	訓練語料	我是別的學校的學生,請問我可以帶自己的(NB)電腦去成大圖書館用網路找資料嗎?我是別的學校的學生,請問我可以帶自己的(NB)電腦去圖書館用網路找資料嗎?
	是否系統回覆	無



編號	16	
問題描述	圖書館電腦可不可以自行安裝其他軟體?	
系統輸出門檻 (Avg_score)	2.0356	
Rank1	意圖集群	Cluster 460
	相似度(Score)	3.804
	訓練語料	請問是否可自行攜帶軟體光碟至本館電腦安裝?請問可以安裝自行攜帶的軟體於圖書館的電腦上使用嗎?
	系統回覆	不可以。由於牽涉軟體版權及電腦管理問題,本館電腦僅供資料檢索館藏使用,唯本館採購視聽資料之互動光碟例外。
	是否系統回覆	是
Rank2	意圖集群	Cluster 510
	相似度(Score)	2.074
	訓練語料	圖書館網頁上 Endnote 書目管理軟體的版本有很多種,請問我應該安裝那一種版本呢?請問要如何知道我的電腦可否安裝最新版本的 Endnote 書目管理軟體?

	系統回覆	由於期刊格式會隨著年代而有所改變，故建議下載安裝最新的版本較好，但如果您習慣使用某一個版本，也可以在論文完成後要插入參考書目之前，將資料備份後先更新 Endnote 版本再插入參考書目，這樣才能確保期刊格式是正確的。
	是否系統回覆	是
Rank3	意圖集群	Cluster 80
	相似度(Score)	1.717
	訓練語料	安裝 EndNote 要輸入認證密碼(product key)? 請問為何我在安裝 Endnote 書目管理軟體時系統要求輸入 ProductKey 呢? 哪裡可以下載書目管理軟體 EndNote? 安裝 EndNote 要輸入認證密碼(product key)嗎?
	是否系統回覆	無
Rank4	意圖集群	Cluster 156
	相似度(Score)	1.305
	訓練語料	請問可借用圖書館的電腦及印表機列印作業嗎? 可以用圖書館的電腦列印作業嗎? 可以用圖書館的電腦列印作業嗎? 可以用圖書館的電腦列印作業嗎?
	是否系統回覆	無
Rank5	意圖集群	Cluster 525
	相似度(Score)	1.278
	訓練語料	請問圖書館公用電腦之登記使用時間? 請問圖書館提供的電腦之登記使用時間?
	是否系統回覆	無

編號	17	
問題描述	還書箱如何使用？	
系統輸出門檻 (Avg_score)	2.2554	
Rank1	意圖集群	Cluster 268
	相似度(Score)	4.167
	訓練語料	請問還書箱如何使用？還書箱
	系統回覆	還書箱提供不方便入館或閉館期間需要還書的讀者使用，圖書館會在次一開館日刷還上架。使用還書箱的讀者可於次日 10:30 後連結本館網頁查詢個人借閱紀錄確認還書狀況。另多媒體資料如 DVD、VCD、錄音帶....等，請勿投入還書箱，以免損壞。
	是否系統回覆	是
Rank2	意圖集群	Cluster 60
	相似度(Score)	2.648
	訓練語料	投入還書箱之書如何認定其還書日？投入還書箱之書如何認定其還書日？投入還書箱之書如何認定其還書日？
	系統回覆	讀者擲入還書箱之書，其歸還日期以次一開館日認定。若讀者將因此產生逾期情形，請提早歸還。還書箱乃為便利讀者在閉館期間還書而設置，當讀者將圖書擲入還書箱時，還書手續其實尚未完成，因此透過還書箱還書的讀者，請於次一工作日連上本館網頁查詢個人借閱紀錄或致電擲入還書箱所在地之圖書館洽詢，以確認還書手續是否完成。
	是否系統回覆	是
Rank3	意圖集群	Cluster 340
	相似度(Score)	2.375
	訓練語料	視聽資料可以在還書箱歸還嗎？還書箱可以歸還光碟嗎？視聽資料的歸還可否投入還書箱？是否每種欲歸還之資料均可投入還書箱？還書箱可以歸還光碟嗎？
	系統回覆	不行，請至二樓借還書櫃台歸還。若投入還書箱將依照還書箱使用規則停權一個月(每件，可累計)。若因誤投，造成資料損壞，需負賠償責任，除依本館賠償規則辦理賠償之外並停權一個月。
	是否系統回覆	是

Rank4	意圖集群	Cluster 82
	相似度(Score)	1.046
	訓練語料	如何查詢、使用電子書？如何查詢、使用電子書？ 如何查詢電子書？怎麼看？如何查詢、使用電子書？
	是否系統回覆	無
Rank5	意圖集群	Cluster 190
	相似度(Score)	1.041
	訓練語料	非書資料可以在總館二樓還嗎？非書資料可以在總館二樓還嗎？
	是否系統回覆	無



編號	18	
問題描述	有人用包包佔位子該怎麼辦？	
系統輸出門檻 (Avg_score)	2.1381	
Rank1	意圖集群	Cluster 214
	相似度(Score)	5.396
	訓練語料	同學佔位子怎麼辦？同學在自習室佔位子怎麼辦？ 有同學佔位子怎麼辦？
	系統回覆	請向櫃台館員索取「時間條」，取得「時間條」十五分鐘後，原使用者仍未回座，即喪失使用權
	是否系統回覆	是
Rank2	意圖集群	Cluster 71
	相似度(Score)	1.502
	訓練語料	我有一些私人藏書，想贈送給圖書館，該怎麼辦？ 私人藏書想送給圖書館，該如何處理？我有一些私人藏書，想送給圖書館，該怎麼辦？我有一些私人藏書，想送給圖書館，該怎麼辦？
	是否系統回覆	無
Rank3	意圖集群	Cluster 260
	相似度(Score)	1.279
	訓練語料	如何借用視聽區十人視聽室？請問視聽區要幾人才可以借用？
	是否系統回覆	無
Rank4	意圖集群	Cluster 488
	相似度(Score)	1.278
	訓練語料	讀者的館際合作帳號使用已過期，該怎麼辦？讀者的館際合作帳號使用已過期，該怎麼辦？
	是否系統回覆	無
Rank5	意圖集群	Cluster 351
	相似度(Score)	1.235
	訓練語料	寒暑假期間如無法親自到館還書，該怎麼辦？無法親自到館還書，該怎麼辦？無法親自借書時，怎麼辦？寒暑假或圖書館閉館期間，如無法親自到館還書，該怎麼辦？
	是否系統回覆	無

編號		19
問題描述		有沒有什麼推薦的推理小說？
系統輸出門檻 (Avg_score)		1.7819
Rank1	意圖集群	Cluster 43
	相似度(Score)	2.259
	訓練語料	有沒有館員或師長推薦閱讀的書單呢？有沒有館員或師長推薦閱讀的書單呢？有沒有館員或師長推薦閱讀的書單呢？
	系統回覆	點擊以下連結有推薦閱讀的書單： http://web.lib.fju.edu.tw/chi/100books/g
	是否系統回覆	是
Rank2	意圖集群	Cluster 195
	相似度(Score)	1.882
	訓練語料	圖書館有沒有「空中英語教室」？圖書館有沒有「空中英語教室」？
	系統回覆	查詢館藏的關鍵字擷取： 推理小說 https://library.lib.fju.edu.tw:444/search~S0*cht/?searchtype=Y&searcharg=推理小說
	是否系統回覆	是
Rank3	意圖集群	Cluster 357
	相似度(Score)	1.668
	訓練語料	請問什麼是 VOD？請問什麼是 VOD？
	是否系統回覆	無
Rank4	意圖集群	Cluster 372
	相似度(Score)	1.593
	訓練語料	我想推薦學術期刊或資料庫？如何推薦學術期刊？ 推薦學術期刊需注意些什麼？
Rank5	是否系統回覆	無
	意圖集群	Cluster 360
	相似度(Score)	1.507
	訓練語料	請問推薦的書籍，如何知道圖書館的處理方式與進度？推薦的書籍，請問如何知道處理進度？

編號	20	
問題描述	學習韓文的資源？	
系統輸出門檻 (Avg_score)	0.7693	
Rank1	意圖集群	Cluster 452
	相似度(Score)	1.758
	訓練語料	語文學習期刊可以借閱？語文學習期刊的光碟片可以外借嗎？語文學習期刊要到那裡借閱？
	系統回覆	1.借期 7 天(含假日)，逾期滯還金，每日 2 元。2.當期期刊不外借。3.如有遺失、毀損，需自行購買原卷期賠償或以該期刊定價賠款，若無法查得該期刊購入價格，則以 300 元之金額賠償。4.請遵照著作權法規範，勿拷貝複製。
	是否系統回覆	是
Rank2	意圖集群	Cluster 270
	相似度(Score)	0.974
	訓練語料	請問可否自行攜帶視聽資料到多媒體中心使用？請問如何使用多媒體中心的視聽資料？請問可否自行攜帶視聽資料到多媒體視聽區使用？如何在多媒體學習中心內使用視聽資料？可以帶自己的資料來多媒體學習中心使用嗎？
	系統回覆	根據著作權法明令規定，圖書館為公共場所，所使用之資料需為”公開播映版”，而一般讀者所購買之資料則大都屬家用版，所以若讀者拿自己的資料到七樓多媒體區來觀賞恐有觸犯法律之嫌。因此基於「智慧財產權」之規範及管理上考量，多媒體區統一不對外開放讀者帶自己的片子來使用本區設備。
是否系統回覆	是	
Rank3	意圖集群	Cluster 233
	相似度(Score)	0.593
	訓練語料	退休教職員工要進入圖書館，該如何辦理？退休教職員工如何辦理借閱圖書館資源呢？交大退休的教職員工該如何辦理借書證？成大退休的教職員工該如何辦理借書證？教職員工如何申請借書證？南應大退休的教職員工該如何辦理借書證？退休的教職員工該如何辦理借書證？

	是否系統回覆	無
Rank4	意圖集群	Cluster 38
	相似度(Score)	0.262
	訓練語料	<p>向圖書館外借的館藏遺失了，該如何處理？學生證遺失怎麼辦？圖書館的燈壞了，怎麼辦？圖書館沒有我需要的書和期刊時怎麼辦？圖書遺失了如何處理？借書遺失該如何處理？若只遺失隨書所附之光碟呢？如果向圖書館所借之書已遺失，該如何處理？圖書遺失如何處理？圖書附件（如光碟）遺失如何處理？資料損壞或遺失要怎麼辦？北醫沒有我要的期刊文獻，只有其他圖書館有，要怎麼辦？我要看的書學校沒有買？如果借的圖書遺失了怎麼辦？遺失物品要去哪裡找呢？向圖書館借的書若遺失了，該怎麼辦？我想借的書學校沒有買，怎麼辦？借閱圖書遺失怎麼辦？【分館】如何找私人遺失物品？向圖書館借閱的圖書不小心遺失了怎麼辦？書遺失了怎麼辦？遺失圖書館證件，我應如何處理？請問圖書如果找不到怎麼辦？請問圖書遺失了怎麼辦？向圖書館借的書遺失了，該如何處理？請問我在圖書館借閱的圖書遺失了，怎麼辦？我的借書證遺失了，該怎麼辦？我的東西在圖書館遺失了，怎麼辦？向圖書館借的書遺失了怎麼辦？借的書找不到（遺失了）或破損了怎麼辦？我在圖書館借的書丟了怎麼辦？如果買不到書，怎麼辦？書丟（壞）了，怎麼辦？圖書遺失怎麼辦？請問圖書遺失應如何處理？在圖書館遺失物品，該如何處理？若我借閱的圖書遺失了該怎麼辦？借書證件遺失了該怎麼辦？我向圖書館借的書遺失了，怎麼辦？學生證遺失可否借書？借書證遺失了，怎麼辦？若將圖書館所借的書遺失了，怎麼辦？學生證遺失了，還可以借書嗎？遺失圖書怎麼辦？如果只有遺失附件光碟呢？學校圖書館沒有我要的期刊該怎麼辦？圖書館沒有想看的書怎麼辦？證件遺失怎麼辦？如果不小心將所借之館藏遺失了該怎麼辦？我的私人物品在圖書館內遺失，該如何處理？我向圖書館借的互換借書證遺失了，怎麼辦？圖書館沒有我要找的資源，該怎麼辦？圖書館沒有我要的期刊，怎麼辦？向圖書館借的書遺失或破損，該如何處理？如果借閱的圖書遺失了，該如何處理？借</p>

		閱的書遺失了怎麼辦?所借的書遺失了,怎麼辦?借閱證遺失該怎麼辦?圖書遺失或損壞怎麼辦?圖書、視聽資料遺失怎麼處理?證件遺失並被冒借圖書怎麼辦?
	是否系統回覆	無
Rank5	意圖集群	Cluster 54
	相似度(Score)	0.260
	訓練語料	總圖的密集書庫在哪裡?「密集書庫」的書可以借嗎?密集書庫的書要如何調閱?
	是否系統回覆	無

編號	21	
問題描述	圖書館有解剖學的書嗎?	
系統輸出門檻 (Avg_score)	1.4886	
Rank1	意圖集群	Cluster 530
	相似度(Score)	1.824
	訓練語料	我想找館藏我想找書我想找書,請問圖書館有嗎?請問圖書館有這本書嗎?圖書館有這本書嗎這本書有在圖書館的館藏裡嗎我想找館藏館藏有這本書嗎放在哪
	系統回覆	查詢館藏的關鍵字擷取: 解剖學 https://library.lib.fju.edu.tw:444/search~S0*cht/?searchtype=Y&searcharg=解剖學
	是否系統回覆	是
Rank2	意圖集群	Cluster 159
	相似度(Score)	1.445
	訓練語料	圖書館開放時間?國定假日有開館嗎?圖書館開放時間?國定假日有開館嗎?圖書館開放時間?國定假日有開館嗎?請問圖書館的開放時間?假日或寒暑假有開館嗎?
	是否系統回覆	無
Rank3	意圖集群	Cluster 226
	相似度(Score)	1.415

	訓練語料	圖書館有收藏過期的報紙嗎？圖書館有收藏過期的報紙嗎？圖書館有收藏過期的報紙嗎？圖書館有收藏過期的報紙嗎？圖書館有收藏過期的報紙嗎？
	是否系統回覆	無
Rank4	意圖集群	Cluster 442
	相似度(Score)	1.380
	訓練語料	圖書館有提供掃瞄器嗎？圖書館有提供掃瞄器嗎？
	是否系統回覆	無
Rank5	意圖集群	Cluster 325
	相似度(Score)	1.378
	訓練語料	請問高空大圖書館借的書可以在其他市圖分館還書嗎？請問高空大圖書館借的書可以在其他市圖分館還書嗎？
	是否系統回覆	無

編號	22	
問題描述	有沒有好看的電影原著小說？	
系統輸出門檻 (Avg_score)	1.4412	
Rank1	意圖集群	Cluster 195
	相似度(Score)	1.882
	訓練語料	圖書館有沒有「空中英語教室」？圖書館有沒有「空中英語教室」？
	系統回覆	查詢館藏的關鍵字擷取： 電影原著小說 https://library.lib.fju.edu.tw:444/search~S0*cht/?searchtype=Y&searcharg=電影原著小說
	是否系統回覆	是
Rank2	意圖集群	Cluster 261
	相似度(Score)	1.363
	訓練語料	如何知道某種期刊有沒有館藏？如何知道某種期刊有沒有館藏？想知道圖書館有沒有我要的書？如何知道圖書館有沒有我需要的電子資源？如何知道某種期刊有沒有館藏？要如何知道某種期刊有沒有館藏？
	是否系統回覆	無

Rank3	意圖集群	Cluster 42
	相似度(Score)	1.346
	訓練語料	請問期刊在那裡？請問期刊在那裡？請問期刊在那裡？
	是否系統回覆	無
Rank4	意圖集群	Cluster 441
	相似度(Score)	1.308
	訓練語料	請問視聽區的開放時間？請問如何利用多媒體視聽區的館藏資料？請問多媒體視聽區的開放時間？請問多媒體視聽區的資料可以外借嗎？
	是否系統回覆	無
Rank5	意圖集群	Cluster 43
	相似度(Score)	1.307
	訓練語料	有沒有館員或師長推薦閱讀的書單呢？有沒有館員或師長推薦閱讀的書單呢？有沒有館員或師長推薦閱讀的書單呢？
	是否系統回覆	無

編號	23	
問題描述	我想找深度學習的資源	
系統輸出門檻 (Avg_score)	1.6661	
Rank1	意圖集群	Cluster 530
	相似度(Score)	2.146
	訓練語料	我想找館藏。我想找書，請問圖書館有嗎？請問圖書館有這本書嗎？圖書館有這本書嗎？這本書有在圖書館的館藏裡嗎？我想找館藏館藏有這本書嗎放在哪
	系統回覆	查詢館藏的關鍵字擷取： 深度學習資源 https://library.lib.fju.edu.tw:444/search~S0*cht/?searchtype=Y&searcharg=深度學習資源
	是否系統回覆	是
Rank2	意圖集群	Cluster 452
	相似度(Score)	1.758
	訓練語料	語文學習期刊可以借閱？語文學習期刊的光碟片可以外借嗎？語文學習期刊要到那裡借閱？

	系統回覆	1.借期 7 天(含假日)，逾期滯還金，每日 2 元。2.當期刊不外借。3.如有遺失、毀損，需自行購買原卷期賠償或以該期刊定價賠款，若無法查得該期刊購入價格，則以 300 元之金額賠償。4.請遵照著作權法規範，勿拷貝複製。
	是否系統回覆	是
Rank3	意圖集群	Cluster 32
	相似度(Score)	1.506
	訓練語料	我的東西不見了，請問圖書館有撿到嗎？失物招領應該去哪裡找？
	是否系統回覆	無
Rank4	意圖集群	Cluster 438
	相似度(Score)	1.475
	訓練語料	為什麼我要找的期刊常不在期刊架上？
	是否系統回覆	無
Rank5	意圖集群	Cluster 368
	相似度(Score)	1.446
	訓練語料	請問我可以帶自己的電腦去成大圖書館用網路找資源嗎？我是別的學校的學生，請問我可以帶自己的(NB)電腦去圖書館用網路找資料嗎？
	是否系統回覆	無

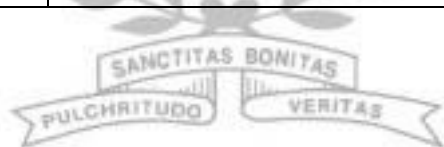
編號	24	
問題描述	食譜放在哪？	
系統輸出門檻 (Avg_score)	1.4417	
Rank1	意圖集群	Cluster 130
	相似度(Score)	2.084
	訓練語料	請問我要到哪裡找過期的期刊資料呢？過期期刊放在哪裡？館內的過期期刊放在哪裡？圖書館哪裡有影印機？請問過期期刊及各類型資料放在哪？過期學術性期刊放在哪邊？報紙放在哪裡？請問館內的過期期刊放在哪裡？圖書館報紙放在哪裡？請問圖書館的報紙放在哪裡？過期期刊、報紙放在哪裡？合訂本期刊放在哪裡？期刊放在哪裡？圖書館的影印機放在那裡？過期期刊放在哪？圖書館影印機放在哪裡？

	系統回覆	查詢館藏的關鍵字擷取： 食譜 https://library.lib.fju.edu.tw:444/search~S0*cht/?searchtype=Y&searcharg=食譜
	是否系統回覆	是
Rank2	意圖集群	Cluster 289
	相似度(Score)	1.390
	訓練語料	當年度學術性期刊放在那兒?當年度期刊放在那兒?
	是否系統回覆	無
Rank3	意圖集群	Cluster 178
	相似度(Score)	1.270
	訓練語料	畢業紀念冊放在那裡?可以外借嗎?畢業紀念冊放在那裡?可以外借嗎?
	是否系統回覆	無
Rank4	意圖集群	Cluster 247
	相似度(Score)	1.232
	訓練語料	東西遺忘在館內, 要去哪邊認領?東西遺忘在館內, 要去哪邊認領?
	是否系統回覆	無
Rank5	意圖集群	Cluster 364
	相似度(Score)	1.232
	訓練語料	Adobe Acrobat Writer 中文版要在哪下載? Adobe Acrobat Writer 中文版要在哪下載?
	是否系統回覆	無

編號	25	
問題描述	網路迷因是甚麼意思？	
系統輸出門檻 (Avg_score)	1.2756	
Rank1	意圖集群	Cluster 531
	相似度(Score)	1.888
	訓練語料	請問是什麼?我想查在維基百科是啥意思?查 wiki 查 wikipedia 在維基的定義?甚麼意思?我想查這個術語的解釋
	系統回覆	查詢 Wikipedia 的關鍵字擷取： 網路迷因 網路迷因（英語：），是指一夕間在網際網路上被大量宣傳及轉播，一舉成為備受注目的事物，亦可稱為網路爆紅事物。 https://zh.wikipedia.org/wiki/網路迷因
	是否系統回覆	是

編號	26	
問題描述	meme 是什麼？	
系統輸出門檻 (Avg_score)	1.8474	
Rank1	意圖集群	Cluster 357
	相似度(Score)	2.619
	訓練語料	請問什麼是 VOD？請問什麼是 VOD？
	系統回覆	查詢 Wikipedia 的關鍵字擷取： meme 又譯媒因、覓母、米姆、瀾等。目前比較公認的定義是透過模仿在人與人之間傳播的思想、行為或風格，通常是為了傳達模因所代表的特定現象、主題或意義。這個詞是在 1976 年，由理查·道金斯在《自私的基因》一書中所創造，將文化傳承的過程，類比成做生物學中的演化繁殖規則（有共同先祖、隨著環境改變進化、優勝劣汰等等）。 https://zh.wikipedia.org/wiki/meme
	是否系統回覆	是

編號	27	
問題描述	新冠肺炎是什麼？	
系統輸出門檻 (Avg_score)	1.9959	
Rank1	意圖集群	Cluster 357
	相似度(Score)	3.272
	訓練語料	請問什麼是 VOD？請問什麼是 VOD？
	系統回覆	<p>查詢 Wikipedia 的關鍵字擷取： 新冠肺炎</p> <p>系統回覆 2019 冠狀病毒病，是一種由嚴重急性呼吸系統綜合症冠狀病毒 2 型（縮寫：）引發的傳染病。該病已知的首名病人 2019 年末於中華人民共和國湖北省武漢市確診，其後此病在全球各國大規模爆發並急速擴散，成為人類歷史上致死人數最多的流行病之一。截至 2021 年 5 月 24 日，全球已有個國家和地區累計報告逾 1.67 億名確診病例，逾 346.5 萬名患者死亡，目前仍在持續擴散中。世界各國對該病病死率的估計值差異甚大，截止 2021 年 2 月 8 日，多數國家該病的觀測病死率在 0.5%-5.0%之間，全球初步修正病死率約為 2.9%。</p> <p>https://zh.wikipedia.org/wiki/新冠肺炎</p>
	是否系統回覆	是



編號	28	
問題描述	資料科學的定義是？	
系統輸出門檻 (Avg_score)	1.0160	
Rank1	意圖集群	Cluster 531
	相似度(Score)	1.385
	訓練語料	請問是什麼？我想查在維基百科是啥意思？查 wiki 查 wikipedia 在維基的定義？甚麼意思？我想查這個術語的解釋
	系統回覆	查詢 Wikipedia 的關鍵字擷取： 資料科學

		<p>資料科學（英語：）是一門利用資料學習知識的學科，其目標是透過從資料中提取出有價值的部分來生產資料產品。它結合了諸多領域中的理論和技術，包括應用數學、統計、模式識別、機器學習、資料視覺化、資料倉庫以及高效能計算。資料科學透過運用各種相關的資料來幫助非專業人士理解問題。</p> <p>資料科學技術可以幫助我們如何正確的處理資料並協助我們在生物學、社會科學、人類學等領域進行研究調研。此外，資料科學也對商業競爭有極大的幫助。</p> <p>https://zh.wikipedia.org/wiki/資料科學</p>
	是否系統回覆	是

編號	29	
問題描述	我想查機器學習	
系統輸出門檻 (Avg_score)	1.4354	
Rank1	意圖集群	Cluster 531
	相似度(Score)	2.161
	訓練語料	請問是什麼?我想查在維基百科是啥意思?查 wiki 查 wikipedia 在維基的定義?甚麼意思?我想查這個術語的解釋
	系統回覆	查詢 Wikipedia 的關鍵字擷取： 機器學習 https://zh.wikipedia.org/wiki/機器學習
	是否系統回覆	是

編號	30	
問題描述	RNA 的解釋	
系統輸出門檻 (Avg_score)	0.7872	
Rank1	意圖集群	Cluster 531
	相似度(Score)	1.188
	訓練語料	請問是什麼?我想查在維基百科是啥意思?查 wiki 查 wikipedia 在維基的定義?甚麼意思?我想查這個術語的解釋
	系統回覆	查詢 Wikipedia 的關鍵字擷取： rna 核糖核酸，是一類由核糖核苷酸透過 3',5'-磷酸二酯鍵聚合而成的線性大分子。自然界中的 RNA 通常是單鏈的，且 RNA 中最基本的四種鹼基為腺嘌呤 (A)、尿嘧啶 (U)、鳥嘌呤 (G)、胞嘧啶 (C)，相對的，與 RNA 同為核酸的 DNA 通常是雙鏈分子，且其含有的含氮鹼基中將 RNA 的尿嘧啶替換為胸腺嘧啶 (T)。 https://zh.wikipedia.org/wiki/rna
	是否系統回覆	是

中文參考文獻

- 陆晓曦 (2009)。Wiki 在高校图书馆中的应用初探。《情报探索》(1)，99-101。取自 <https://www.ixueshu.com/document/c5951ac9594994ee21d6184a21c9c424318947a18e7f9386.html>
- 郭鸿昌 (2008)。Wiki 在图书馆中的应用与实践。《新世纪图书馆》，2008(5)，31-33。
- 黄菁、杨帆 (2009)。Wiki 知识共享与企业 Wiki 理论初探 [The Preliminary Theory Research on Wiki Knowledge Sharing and Enterprise Wiki]。《图书与情报》，2009(1)，55-60。
- 曾元顯 (2012)。自然語言處理 natural language processing。《圖書館學與資訊科學大辭典》。取自 <http://terms.naer.edu.tw/detail/1678997/>
- 姚飞、纪磊、张成昱、陈武 (2011)。实时虚拟参考咨询服务新尝试—清华大学图书馆智能聊天机器人。《现代图书情报技术》，27(4)，77-81。
- 陳宜琳 (2019)。國立臺灣師範大學圖書館參考諮詢機器人建置與評估 (碩士論文)。取自華藝線上圖書館系統。doi:10.6345/NTNU201900975
- 范蔚敏 (2020)。聊天軟體機器人技術應用於大學圖書館虛擬參考諮詢服務建置過程與評估研究。《圖書館學與資訊科學》，46(1)，4-31。doi:10.6245/JLIS.202004_46(1).0001



英文參考文獻

- Al-Zubaide, H., & Issa, A. A. (2011, November). Ontbot: Ontology based chatbot. In *International Symposium on Innovations in Information and Communications Technology* (pp. 7-12). IEEE.
- Augello, A., Pilato, G., Vassallo, G., & Gaglio, S. (2009, March). A semantic layer on semi-structured data sources for intuitive chatbots. In *2009 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 760-765). IEEE.
- Beel, J., Gipp, B., Langer, S., & Breiteringer, C. (2016). Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4), 305-338.
- Banchs, R. E., & Li, H. (2012, July). IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 37-42).
- Chatbot. (2019). Oxford Dictionary, available at: <https://www.lexico.com/definition/chatbot> (accessed 26 May 2021)
- Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explor. Newsl.*, 19(2), 25-35. doi:10.1145/3166054.3166058
- Chen, H., He, T., Ji, D., & Quan, C. (2005). *An unsupervised approach to Chinese word sense disambiguation based on Hownet*. Paper presented at the International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 4, December 2005: Special Issue on Selected Papers from CLSW-5.
- Colby, K. M., Hilf, F. D., Weber, S., & Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3, 199-221. doi:https://doi.org/10.1016/0004-3702(72)90049-5
- Charras, F., Duplessis, G. D., Letard, V., Ligozat, A. L., & Rosset, S. (2016). Comparing system-response retrieval models for open-domain and casual conversational agent. In *Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT@ IVA2016)*.
- Domeniconi, G., Moro, G., Pasolini, R., & Sartori, C. (2015, July). A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf. idf. In *International Conference on Data Management Technologies and Applications* (pp. 39-58). Springer, Cham.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96,

- No. 34, pp. 226-231).
- Fogarolli, A. (2009, 14-16 Sept. 2009). *Word Sense Disambiguation Based on Wikipedia Link Structure*. Paper presented at the 2009 IEEE International Conference on Semantic Computing.
- Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070), 900-901. doi:10.1038/438900a
- Gill, J. K. (2019). Artificial Intelligence Overview and Applications. [Web blog message]. Retrieved from <https://www.xenonstack.com/blog/artificial-intelligence/>
- Gandhe, S., & Traum, D. (2013, August). Surface text based dialogue models for virtual humans. In *Proceedings of the SIGDIAL 2013 Conference* (pp. 251-260).
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.
- Hsieh, Y.-L., Liu, S.-H., Chen, K.-Y., Wang, H.-M., Hsu, W.-L., & Chen, B. (2016, oct). *運用序列到序列生成架構於重寫式自動摘要(Exploiting Sequence-to-Sequence Generation Framework for Automatic Abstractive Summarization)[In Chinese]*, Tainan, Taiwan.
- Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). *Convolutional neural network architectures for matching natural language sentences*. Paper presented at the Advances in neural information processing systems.
- Hussain, S., & Athula, G. (2018, May). Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)* (pp. 698-703). IEEE.
- Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing*. 3rd Edition. Retrieved from <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
- Jafarpour, S., Burges, C. J., & Ritter, A. (2010). Filter, rank, and transfer the knowledge: Learning to chat. *Advances in Ranking*, 10, 2329-9290.
- Jiao, H., Liu, Q., & Jia, H. B. (2007, December). Chinese keyword extraction based on N-gram and word co-occurrence. In *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)* (pp. 152-155). IEEE.
- Kanagala, H. K., & Krishnaiah, V. J. R. (2016, January). A comparative study of K-Means, DBSCAN and OPTICS. In *2016 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6). IEEE.
- Kane, D. A. (2016). The role of chatbots in teaching and learning. *E-Learning and the Academic Library: Essays on Innovative Initiatives*, 131.

- Kamps, J., & Koolen, M. (2008). *The importance of link evidence in Wikipedia*. Paper presented at the European Conference on Information Retrieval.
- Klein, D., & Manning, C. D. (2003). *Fast exact inference with a factored model for natural language parsing*. Paper presented at the Advances in neural information processing systems.
- Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, 81(1), 205-223. doi:<https://doi.org/10.1006/brln.2001.2518>
- Leuski, A., & Traum, D. (2011). NPCEditor: Creating virtual human dialogue using information retrieval techniques. *Ai Magazine*, 32(2), 42-56.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Mihalcea, R. (2007). *Using wikipedia for automatic word sense disambiguation*. Paper presented at the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference.
- Mnasri, M. (2019). Recent advances in conversational NLP: Towards the standardization of Chatbot building. *arXiv preprint arXiv:1903.09025*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Majumder, P., Mitra, M., & Chaudhuri, B. B. (2002, November). N-gram: a language independent approach to IR and NLP. In International conference on universal knowledge and language.
- Nastase, V., & Strube, M. (2008). *Decoding Wikipedia Categories for Knowledge Acquisition*. Paper presented at the AAAI.
- Ponzetto, S. P., & Navigli, R. (2009). *Large-scale taxonomy mapping for restructuring and integrating Wikipedia*. Paper presented at the Twenty-First International Joint Conference on Artificial Intelligence.
- Ponzetto, S. P., & Strube, M. (2007). *Deriving a large scale taxonomy from Wikipedia*. Paper presented at the AAAI.
- Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., . . . Chu, W. (2017). *Alime chat: A sequence to sequence and rerank based chatbot engine*. Paper presented at the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).
- Rambow, O., Bangalore, S., & Walker, M. (2001). *Natural language generation in*

- dialog systems*. Paper presented at the Proceedings of the first international conference on Human language technology research, San Diego.
- Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1), 57-87.
- Riaz, M., Haider, W., & Rehman, S. (2019). *Automatic Detection of Ambiguous Software Requirements: An Insight*.
- Rong, Q. S., Yan, J. B., & Guo, G. Q. (2004). Research and implementation of clustering algorithm based on DBSCAN [J]. *Computer Applications*, 4, 45-46.
- Rosruen, N., & Samanchuen, T. (2018, December). Chatbot utilization for medical consultant system. In *2018 3rd technology innovation management and engineering science international conference (TIMES-iCON)* (pp. 1-5). IEEE.
- Song, Y., Yan, R., Li, X., Zhao, D., & Zhang, M. (2016). Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.
- Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236), 433-460. doi:10.1093/mind/LIX.236.433
- Vinyals, O., & Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.
- Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language teaching*, 31(2), 57-71.
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.
- Weizenbaum, J. (1983). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 26(1), 23-28.
- Wikipedia. (2019). Wikipedia, the free encyclopedia. Retrieved from <https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=921973305>
- Yan, H., Ding, S., & Suel, T. (2009, April). Inverted index compression and query processing with optimized document ordering. In *Proceedings of the 18th international conference on World wide web* (pp. 401-410).
- Young, J. R. (2019, Jun 14). Bots in the Library? Colleges Try AI to Help Researchers (But With Caution). Retrieved from <https://www.edsurge.com/news/2019-06-14-bots-in-the-library-colleges-try-ai-to-help-researchers-but-with-caution>