

國立台北大學資訊工程學系專題報告

語音搜尋新聞的服務系統

專題組員:張凱名 賴昱賢 朱彬愷

專題編號:PRJ-NTPUCSIE-108-012

執行期間:108 年 9 月 至 109 年 6 月

1. 摘要

現代人獲取新聞資訊的方法不外乎透過觀看電視與用手機查詢或是利用手機上的 siri、ok google 等個人語音助理，但這些都需要透過眼睛來觀看，所以對於盲人或是老花眼等視力上有困難的人來說在獲取新聞的資訊上則比起一般的人困難許多。所以我們希望能夠研究讓盲人以及視力障礙的人們能夠輕鬆獲取新聞資訊的語音服務。

我們在語音服務的載體上選擇了互動式機器人，以達到陪伴的效果，而現今市面上的互動式機器人有注重陪伴兒童的說故事互動，提醒大人的代辦事項的 Zenbo 智慧居家機器人，以及重視情感陪伴的機器人和老人遠端視訊照護的 Pepper 機器人。我們使用了 Zenbo 機器人。

2. 簡介

(1) 研製背景

機器人與一般語音助理可以做到將搜尋的關鍵字丟到瀏覽器上讓使用者能觀看搜尋結果，但結合語音搜尋服務就能夠提供盲人朋友與有視力障礙和閱讀困難者使用語音的方式得

到新聞資訊，例如對 Zenbo 詢問 XXX 的新聞，zenbo 會用語音回答包含有

XXX 的新聞標題，並能夠再進一步朗讀出該新聞的內容。

(2) 目標

在本專題中，希望能將新聞查詢功能與語音對答做整合，並且結合雲端儲存的相關數據，達到提供盲人朋友語音搜尋新聞的服務。

(3) 預估成果

- 建立多種語音指令使互動更人性化。
- 運用網頁爬蟲與雲端運算讓機器人能獲取使用者詢問的相關數據。

3. 專題進行方式

(1) 實作平台

1. Zenbo 機器人

Zenbo 機器人是華碩旗下的家用機器人，本身搭載 Android 系統，機器人頭部除了有 10.1 吋的觸摸屏幕，還有 1300 萬畫素拍照相機、紅外線控制器與觸摸感測器；機身則有防掉落感測器與多組聲納感測器。

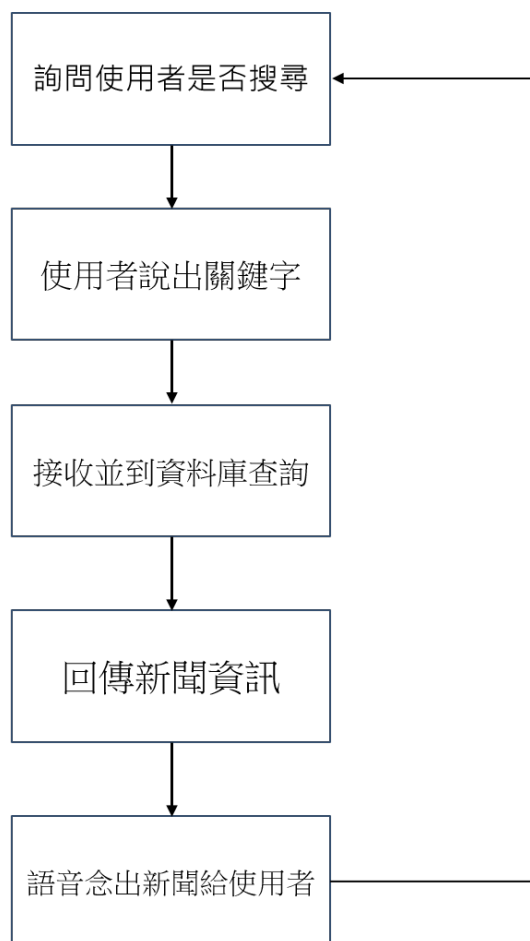
2. RESTful Server

(2) 系統分析與設計摘要

在本專題中，將依下列方法及步驟來時作並完成計畫內容。

1. 系統整體操作設計圖在本次的計畫中機器人要在固定場所與使用

者作接觸，對於機器人與使用者之間的互動需要事先設計和模擬。因此我們將機器人實際在運作時的情境以下的流程圖呈現：



圖一：本計畫機器人運作流程圖

機器人在特定點等待使用者查詢，首先機器人會詢問使用者想要搜尋甚麼新聞，使用者說出與搜尋之關鍵字後，接收並且傳至後端執行查詢資料，從後端回傳新聞資訊至機器人並唸出給使用者聽。

2. 中介層

為了要達成將語音搜尋接收之字串能做處理並一併到資料庫搜尋新聞之資訊。因此我們在遠端的主機架設了一個伺服器當作中介層，不僅可勝任繁雜程序的處理，將來也能更輕易地擴展其他功能。此處我們選擇使用

Python 的 Flask 網路應用框架(Web Application Framework)作為一個 REST 風格的伺服器。比起其他程式語言和框架，選擇 Flask 的原因為 Python 語言的強大及眾多套件資源方便擴展各式功能，以及 Flask 的方便簡約和高自由度。我們的伺服器也結合了 MongoDB 資料庫，將網頁爬蟲之新聞內容資訊存進資料庫以供使用者搜尋時用到。網頁爬蟲部分是定期將我們需要的數據從各大新聞網站擷取下來整理分析並且處理掉相同內容之新聞後，存入資料庫並提供給機器人端做應用。本計畫整體上是用戶端透過 URL 進行 HTTP 請求方法存取資料而運作。例如接收到的關鍵字字串對指定的 URL 進行 POST 方法傳輸關鍵字並得到回傳新聞內容之結果。

3. 關鍵字的處理程序

使用者說出的一段話如何切出我們需要用來搜尋的關鍵字，這裡我們使用 jieba 斷詞並且搭配維基百科的索引目錄詞庫。舉例[我來到台北大學]經過 jieba 原本詞庫的斷詞之後會變成[我/來到/台北/大學]，可以看到[台北大學]被切成了[台北/大學]，這樣的結果並不是我們所想要的輸出，所以我們爬取維基百科的索引目錄詞庫用來當作 jieba 斷詞的詞庫，以增加名詞的豐富度，讓斷詞能斷得更完整，這時[我來到台北大學]就能準確的切成[我/來到/台北大學]。

4. 判斷指令種類

後端收到機器人透過 flask 傳過來的指令後，會先進行初步的判斷，判斷收到的指令是否為日期及新聞種類的新聞或是關鍵字的查詢，並透過切字後的結果之長度決定使用者是想

查詢日期及新聞種類的資料或是關鍵字的新聞資料查詢。

5. 查詢資料庫資訊

我們這邊採用正則表示法 (regular expression) 到資料庫搜尋資料，例如：新聞標題為[王建民連續兩年獲得美聯勝投王]，如果我們的關鍵字為以上標題中任何一個字串的話，就符合正則表示法的搜尋條件。EX: 假設我們搜尋的關鍵字如果為{王建民, 美聯, 勝投王}，只要有出現任何一個上述的字或字串，就可以找到該筆資料。

語法	正則表達式	範例
整數	[0-9]+	3704
有小數點的實數	[0-9]+\.[0-9]+	7.93
英文詞彙	[A-Za-z]+	Code
變數名稱	[A-Za-z_][A-Za-z0-9_]*	_counter
Email	[a-zA-Z0-9_]+@[a-zA-Z0-9_]+	ccc@kmit.edu.tw
URL	http://[a-zA-Z0-9_./]+	http://ccc.kmit.edu.tw/mybook/

圖二：正則表示式的範例

6. 爬取新聞進資料庫

爬蟲程式使用到 python 裡面的 urllib 套件。

urllib 介紹

urllib 是 python 內置的 HTTP 請求庫，是 python 提供的一個用於發起和處理 http 請求和響應的框架。

後期的一些框架，比如：requests、scrapy 等都是基於它

包括以下四個模塊：

urllib.error 異常處理模塊

urllib.request 請求模塊

urllib.parse url 解析模塊

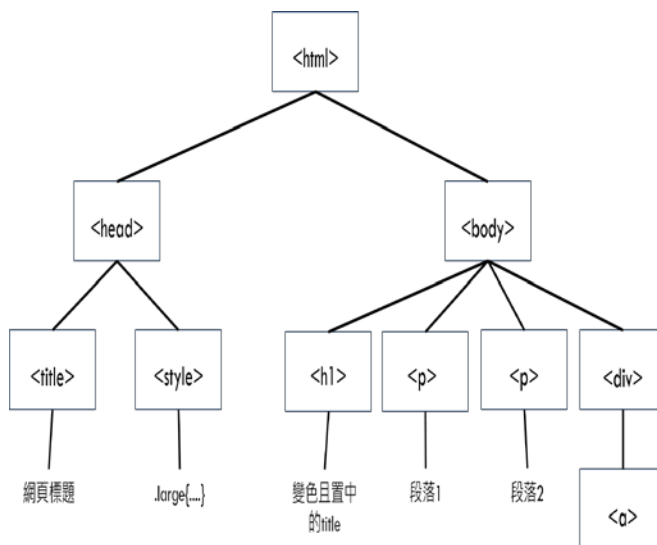
urllib.robotparser robots.txt 解析模塊

爬取的步驟首先使用 urllib.request 獲取頁面內容，用 url

和 header 例項化一個 urllib.request.Request(url, headers)，然後套件裡面的 urlopen() 函式與新聞網頁取得連線並獲取網頁原始碼。

而在取得新聞網頁的原始碼時為了讓網站認為爬蟲程式是一個人類的使用者必須添加 headers，而不同網頁需要的 header 的內容不一樣，參照檢查裡 request header 來構造。例如 dribbble 需要 Referer，即從哪一個頁面跳轉到這個當前頁面的，一般填寫網站相關頁面網址就可以。

取得的網頁原始碼(html 文件)使用 BeautifulSoup 套件轉換成樹狀的結構。

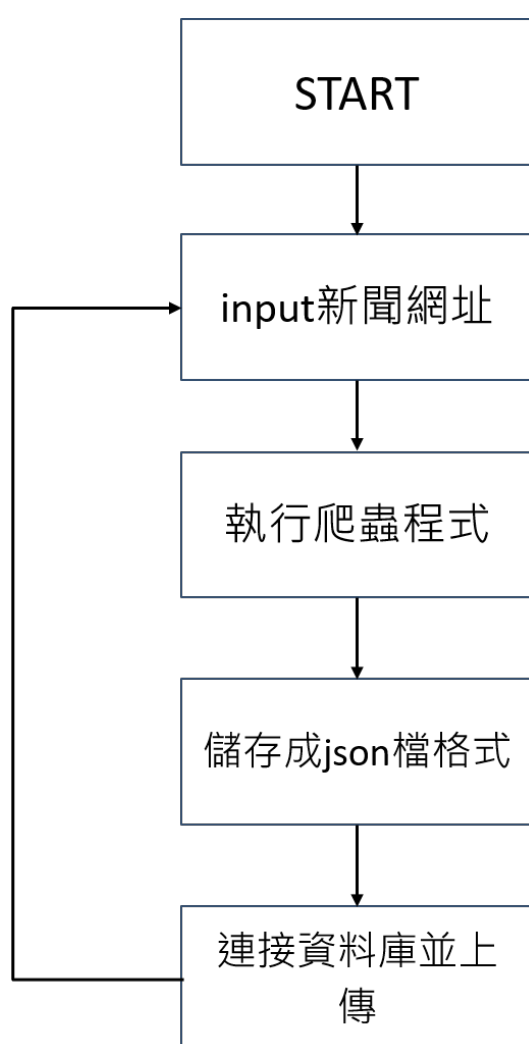


圖三：網頁原始碼樹狀圖

如圖三所示，將網頁原始碼轉換成樹狀圖之後就能夠找到需要爬取的新聞標題、網頁內文等內容的標籤，這時使用可以根據標籤名，屬性，內容查詢文件的標準選擇器以及通過雖

然功能弱但是速度快，使用尋找標籤名例如：尋找“<p>”這個標籤名的標籤選擇器來獲得需要爬取的標籤的內容，並儲存成 json 檔。

爬取新聞的流程為固定時段爬取當日新聞，並儲存成 json 檔，然後連接資料庫將該則新聞上傳至資料庫上，流程圖如下



7. 開發過程中遇到的困難如下

- 困難：
資料從機器人端到 flask 時會遇到 Unicode 跟 Urlcode 的編碼問題。

- 解決辦法：
在後端先進行兩層式的 decode 才能得到原本使用者詢問的關鍵字，不然會出現 [%53%53%53] 等等的亂碼，導致後端無法正確搜尋。
- 困難：
判斷指令是搜尋指定日期新聞或是搜尋關鍵字新聞時不知道該以甚麼方式作依據來做搜尋。
- 解決辦法：
根據 jieba 切字後的 list 長度來作為判斷依據。例如搜尋關鍵字時經過切字後得到的 list 結果不會像搜尋日期時的切字結果那麼多，所以就可以根據長度作主要判斷依據。

4. 主要成果與評估

在搜尋資料上，從機器人接收到使用者說出的命令後到機器人說出新聞資訊，反應時間控制在兩秒內。一方面後端可以先儲存所有符合使用者說出指令字串之新聞資訊內容，也不會讓使用者等太久。

此服務系統功能可以讓盲人朋友與老花朋友們等視力與閱讀有困難的人輕鬆地透過語音的溝通方式獲得想要知道的新聞資訊內容。

在往後，此功能可以紀錄使用者搜尋過的新聞種類，並經由權重做進一步的分析，讓爬取新聞的特定種類做更進一步的搜尋，達到該種類更深入的新聞資訊探討。除了搜尋新聞以外，也預期可以搜尋每日的股票資訊，包括當日的成交數量或收盤行情等等，成為一個稱職的股票資訊小幫

手。

而本計畫目標環境在於家中、養老機構或其他公共的室內場所都可以運作，當走靠近機器人旁時，機器人會主動詢問使用者想要獲得什麼樣的資訊內容，然後增加資料庫的資訊內容，不在侷限在新聞上而可以更為廣泛的符合每位使用者欲查詢之內容需求。希望將上述功能做一定的整合，將服務的對象範圍擴大，帶來新的商機。

5. 結語與展望

現在市面上的 ok google、siri、amazon 的 alexa 以及微軟的 cortana 等語音助理都非常的成熟及應用廣泛。我們欲將語音助理的資料查詢功能做進一步深層的搜尋並唸出及更多應用，以達到視力困難者也可使用之個人助理，並結合每位使用者的個人喜好做更進一步的資訊儲存，以達到客製化功能。

6. 銘謝

國立台北大學資訊工程研究所分散及行動計算實驗室

7. 參考文獻

[1]

<https://clay-atlas.com/blog/2019/09/23/python-chinese-tutorial-tokenizer-jieba/>

[2]

<https://iter01.com/159404.html>

[3]

<https://blog.gtwang.org/programming/python-beautiful-soup-module-scrape-web-pages-tutorial/3/>

[4]

<https://blog.gtwang.org/programming/python-beautiful-soup-module-scrape-web-pages-tutorial/3/>

[5]

<http://practical-dbm.blogspot.com/2018/04/html-parsing.html>

[6]

<https://www.runoob.com/python3/python3-reg-expressions.html>

[7]

<https://www.runoob.com/mongodb/mongodb-tutorial.html>

[8]

<https://www.runoob.com/python3/python-mongodb.html>

[9]

<https://medium.com/datainpoint/flask-web-api-quickstart-3b13d96cccc2>

[10]

<https://scalegrid.io/blog/how-to-connect-your-mongodb-deployments-to-robo-3t-gui-at-scalegrid>

[11]

<https://litotom.com/2017/05/25/ch5-4-activity-lifecycle/>

[12]

<http://blog.tonycube.com/2011/11/android-get-post.html>

[13]

<https://blog.csdn.net/hnyzwtf/article/details/51099075>

[14]

<http://blog.tonycube.com/2011/08/asynctask.html>

[15]

<https://zenbo.asus.com/developer/documents/Zenbo-Junior-SDK/DialogSystem>

[16]

<https://zenbolab.asus.com/>