# 物件導向程式設計期末報告

## 主題:日本旅遊遊記檢索系統


組員:

B06505002 張在然

B06505004 莊博翰

b06505006 陳奕舟

b06505047 陳銘杰

## 一、 系統架構及功能簡介

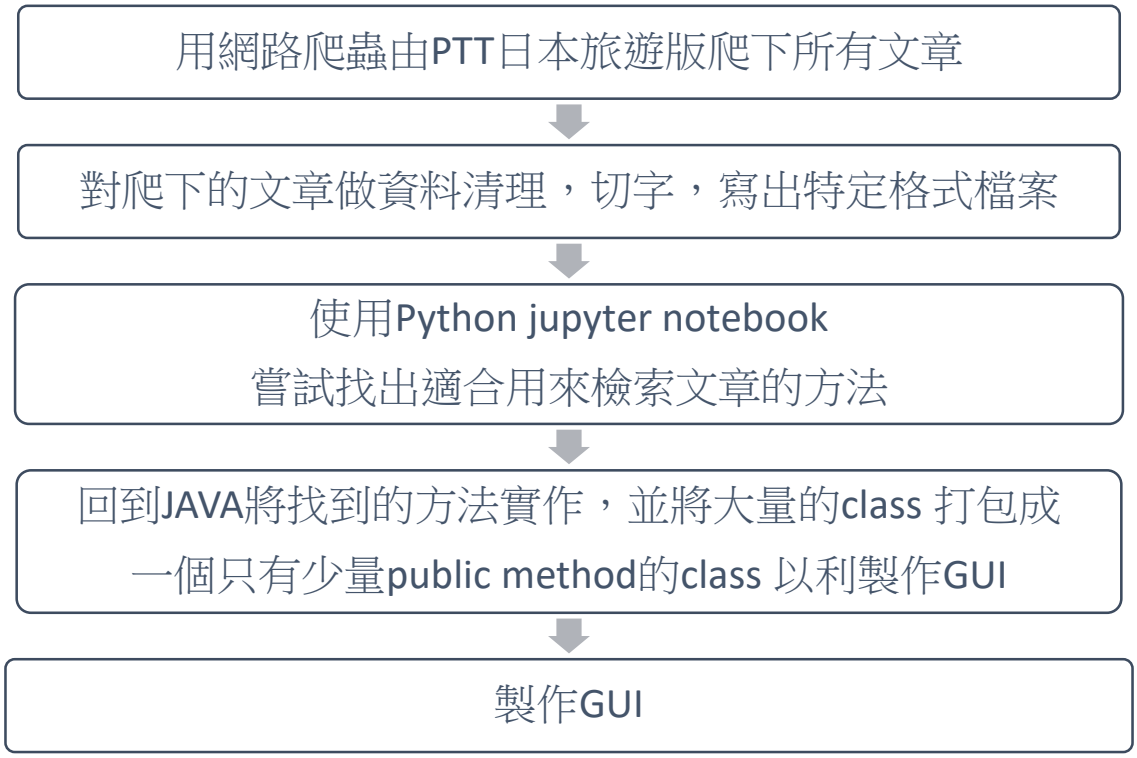
基本上我們想要製作一個類似 PTT 離線版，可以搜尋文章、看文章，此外還有一個進階功能是可以按照特定主題對文章重新排序如(假設主題為食物)：

(1) 優先顯示 "有關食物的段落多的" 文章

(2) 將有關食物的段落從遊記中挑出來顯示

當然還要有堪用的 GUI 介面。

## 二、 專題設計流程:



## 三、 使用到的套件,工具:

Jsoup(爬蟲)

SQlite-JDBC(資料庫)

hanlp(中文最短距離切字)

opencsv(開 CSV 檔)

java swing(GUI)

Python jupyter notebook

MAVEN pom.xml：

```
<dependency>
      <groupId>junit</groupId>
      <artifactId>junit</artifactId>
      <version>3.8.1</version>
      <scope>test</scope>
    </dependency>
    <dependency>
     <groupId>com.hankcs</groupId>
     <artifactId>hanlp</artifactId>
     <version>portable-1.7.1</version>
     </dependency>
     <dependency>
         <groupId>org.xerial</groupId>
         <artifactId>sqlite-jdbc</artifactId>
         <version>3.25.2</version>
     </dependency>
          <dependency>
         <groupId>com.opencsv</groupId>
         <artifactId>opencsv</artifactId>
         <version>4.0</version>
     </dependency>
```

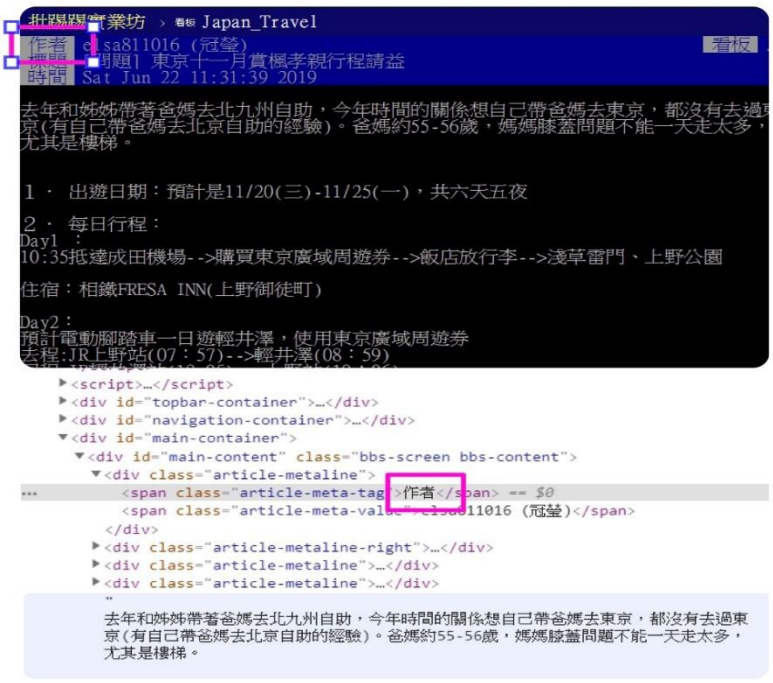
# 四、 資料獲取(網路爬蟲):

流程：

1. 以頁為單位在 ptt 日本旅遊版將原始碼爬取下來

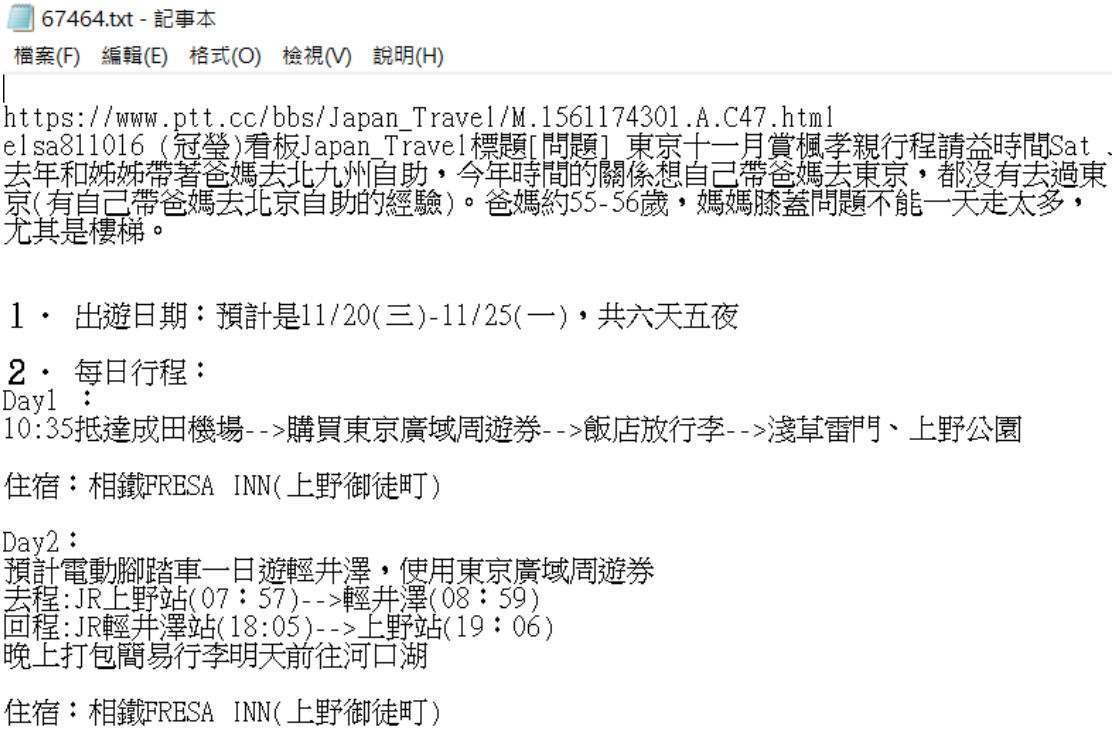

2. 探訪同一頁的 20 篇文章，將每一個頁面都爬下

3. 以「作者」為關鍵字進行字串切割並提取資料



4. 應莊柏翰要求將每篇文章的網址插在文章最開頭

5. 存成文字檔，以下為其中一篇



67464.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

https://www.ptt.cc/bbs/Japan_Travel/M.1561174301.A.C47.html
elsa811016 (冠瑩)看板Japan_Travel標題[問題] 東京十一月賞楓孝親行程請益時間Sat 、
去年和姊姊帶著爸媽去北九州自助，今年時間的關係想自己帶爸媽去東京，都沒有去過東
京(有自己帶爸媽去北京自助的經驗)。爸媽約55-56歲，媽媽膝蓋問題不能一天走太多，
尤其是樓梯。

1． 出遊日期：預計是11/20(三)-11/25(一)，共六天五夜

2． 每日行程：
Day1 ：
10:35抵達成田機場-->購買東京廣域周遊券-->飯店放行李-->淺草雷門、上野公園

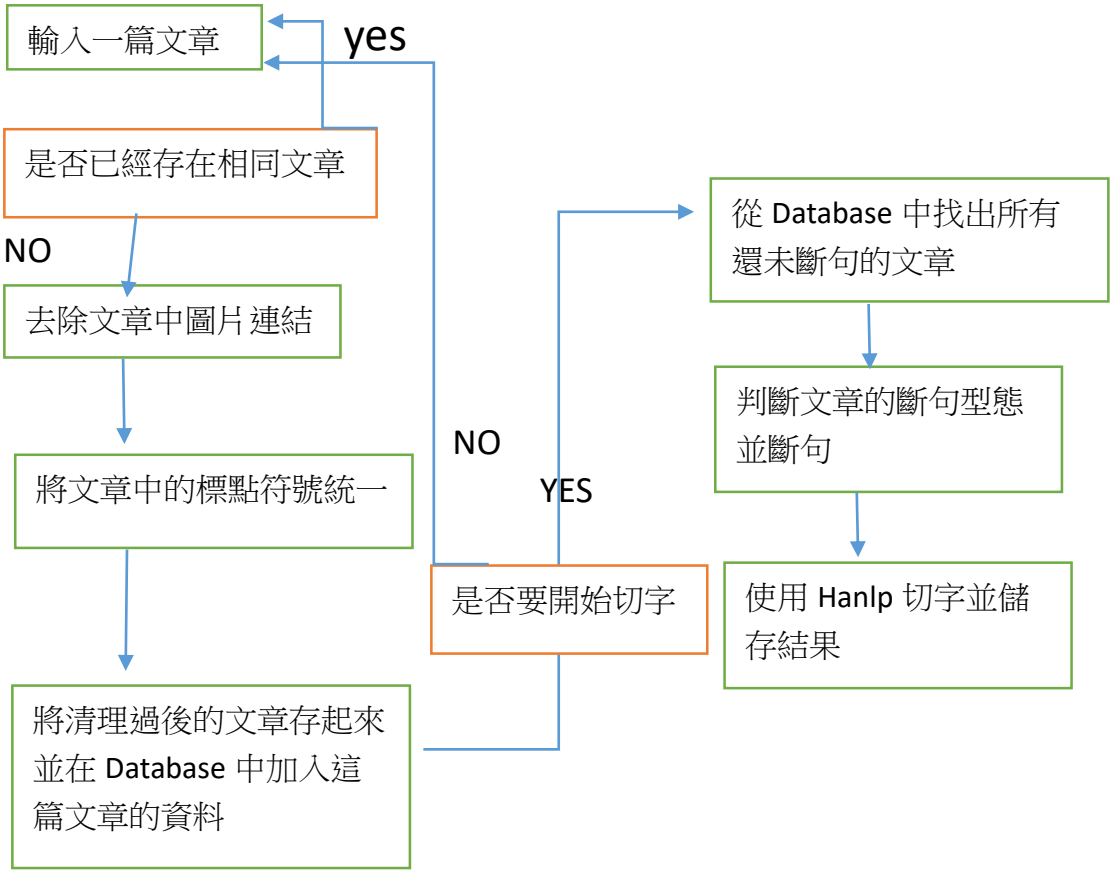
住宿：相鐵FRESA INN(上野御徒町)

Day2：
預計電動腳踏車一日遊輕井澤，使用東京廣域周遊券
去程:JR上野站(07：57)-->輕井澤(08：59)
回程:JR輕井澤站(18:05)-->上野站(19：06)
晚上打包簡易行李明天前往河口湖

住宿：相鐵FRESA INN(上野御徒町)

# 五、 資料清理:

(Package:data_collection_and_process)




## 細項:

首先由文章 Url 來判斷是否為重複文章，現存文章會以

```
sql = "create table if not exists ori (url unique,title,id,sep_or_not); ";
```

url ,標題,id(檔名),sep_or_not(是否以切過字)存在 database

而文章本體則另外存放。

原始文章中有許多圖片連結(字串)，這是不需要的

所以用 regular expression String.replaceAll 清除

```
"\\s*http(|s)[a-z A-Z . % : /  0-9 ? ! = \\+ \\- \\[ \\] \\s \\xa0 _]+[\r \n ,]+"
```

為了避免接下來判斷斷句時混淆，先將標點符號統一

```
article.replaceAll("#|//.|//,|//!|//?|，|。|\r",",");
```

在將文章丟入切字套件前，先將文章一句話一句話分開

需判斷不同文章寫法進行斷句

```
想問一下板上的大大們
目前只有兩個人
現在才準備要買7/28-8/3的機票跟住宿
目前查到
華航
去程0850-1315
回程2005-2245
機票價格兩人一共27808

住宿
用booking查
秋葉原apa酒店
6晚一共19712

這樣算起來一個人平均23760元

請問這樣的價格可以嗎？
```

(以換行斷句)

以標點符號斷句

```
每年度的詳細運行時間和交通可以從官網上確認，前往嵐山的交通十分多元，電車就有JR、嵐電和阪急等系統，主要要看是否有要搭小火車以及出發的地點去評估。由於我已經手持JR的關西周遊券，又是從其他城市玩回京都再前往嵐山，住宿的點也選在京都的JR站附近，所以首選是搭到JR嵯峨嵐山站。但以花燈路活動來說JR相對是最遙遠的選擇，實際上我覺得還是可以雖然要多花一點時間移動到主會場，但不趕時間晃完也是兩個小時內可行的。如果沒有來過嵐山的人可以傍晚在到就可以一次看到白天和夜晚不同的面貌。
```

Hanlp 是最短距離分詞，先行斷句並按句切字不影響結果

還可以避免一次丟入整篇文章使記憶體爆掉。

(要使用 Hanlp 的繁體切字，不然效果很差)

```
import com.hankcs.hanlp.tokenizer.TraditionalChineseTokenizer;
```

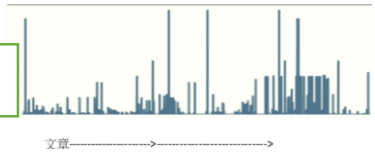
```
TraditionalChineseTokenizer.segment(sentence[i]);
```

# 六、 找出適合用來篩選文章的方法(使用 python)

用所有文章建立一個 dictionary "basedict" { key:詞 value:出現頻率}

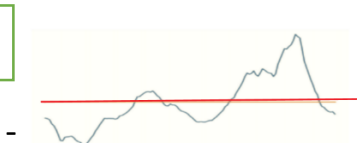
決定一個主題(如:食物)
並選擇相關關鍵字(如:食記)

篩選出標題含有 食記 的文章 need:
==>建立一個 dictionary “needcdict




用 needdict basedict 對文章打分

將文章每一句話的得分模糊化



重複 2-3 次

然後取得分前 n%的句子
重建 dictionary "need_dict"

讀取標題不包含關鍵字的文章”noneed”

調整
n 值使
保留比
例最大
化

選取一得分門檻使:當使用”needdict”給
noneed 打分時，有 95%的句子被篩除

使用同一門檻篩選 need 獲取保留比例

將 needdict 標準化使得分門檻為 1.0

紀錄 needdict

# 細項:

basedict needdict: { key:詞  value:出現頻率}

用 needdict basedict  對句子打分的 function  是:

(needdict(word)/(basedict(word)+1))/(len(sentence)**0.7)

將文章每一句話的得分模糊化的方法是:

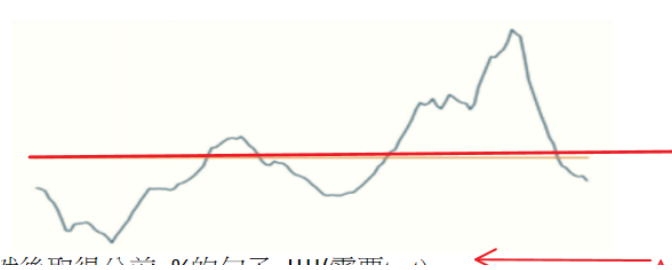
```
newarray=0
for i in range(-4,5,1):
    newarray+=np.roll(p,i)*0.7**abs(i)
```

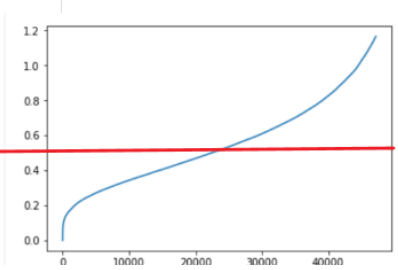
p 是原先文章得分分布(numpy array  每個 value 表一句話得分)

newarray 新的文章得分分布

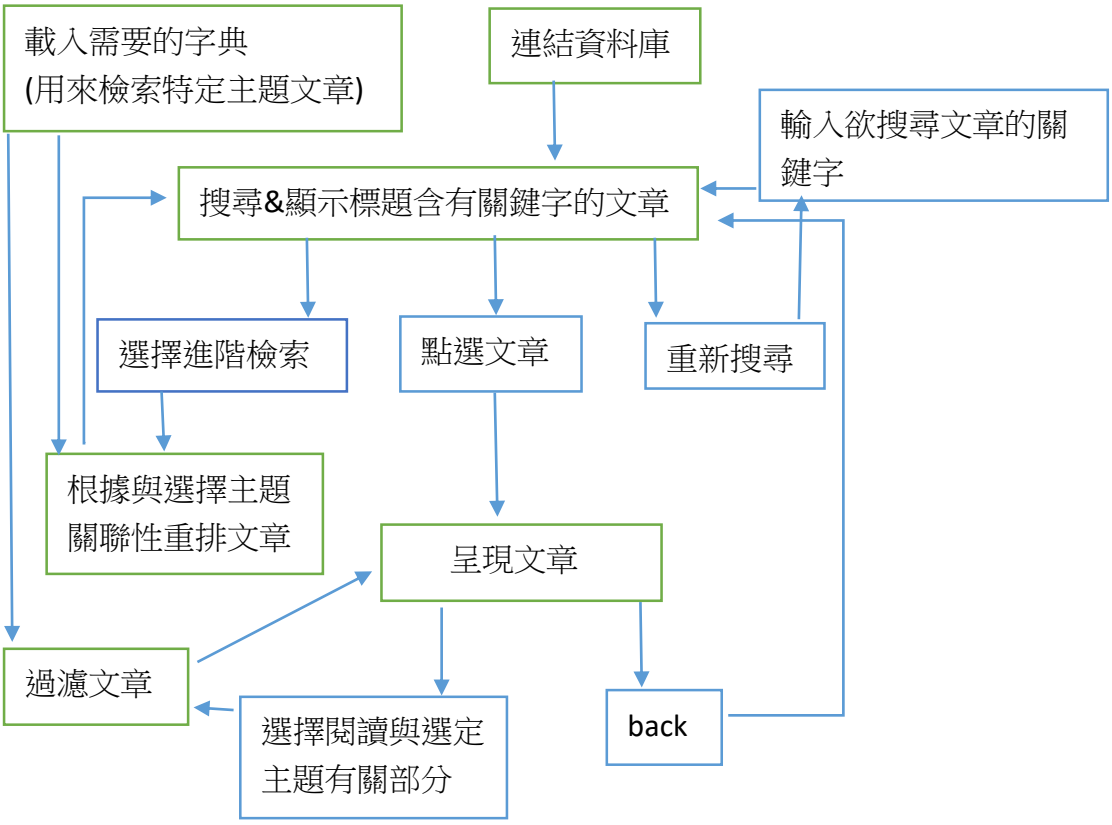
取得得分前 n%的句子的方法是:

sort  原文章得分分布  以第(len(array)*n%)  個值為門檻

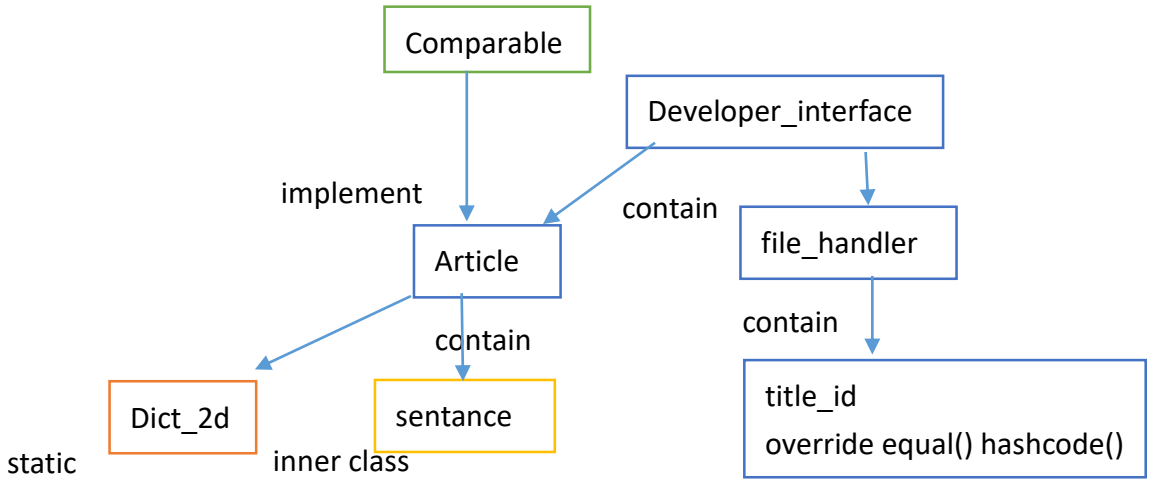




## 七、 專案核心結構(Package: travel article reader)




## 八、 Package 規劃(class diagram)



細項:

Dict_2d 是一個用兩個 hashmap 和一個 2dArray 組成的資料結構一個維度代表:主題，另一維度代表:詞，value:是得分。可以用這個結構查詢一個詞在一主題下的得分(此得分來自先前製作的 needdict)Article implement Comparable 是為了 Sort title_id override equal() hashcode() 是為了用 hashset 取交集

## Exception:

如果 DIct_2d 讀取檔案失敗，會丟出 `Dict 2d operate Exception`

而 error message 分為:只定主題不存在 and 指定 index 錯誤

# 九、資料庫設計

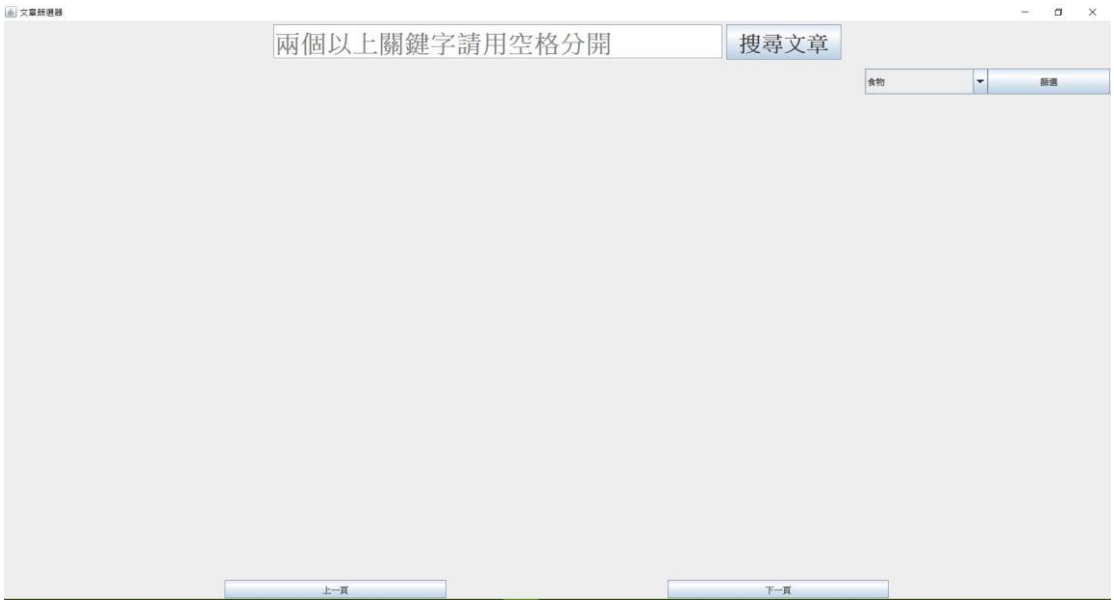
**Table:** ori (未切字)

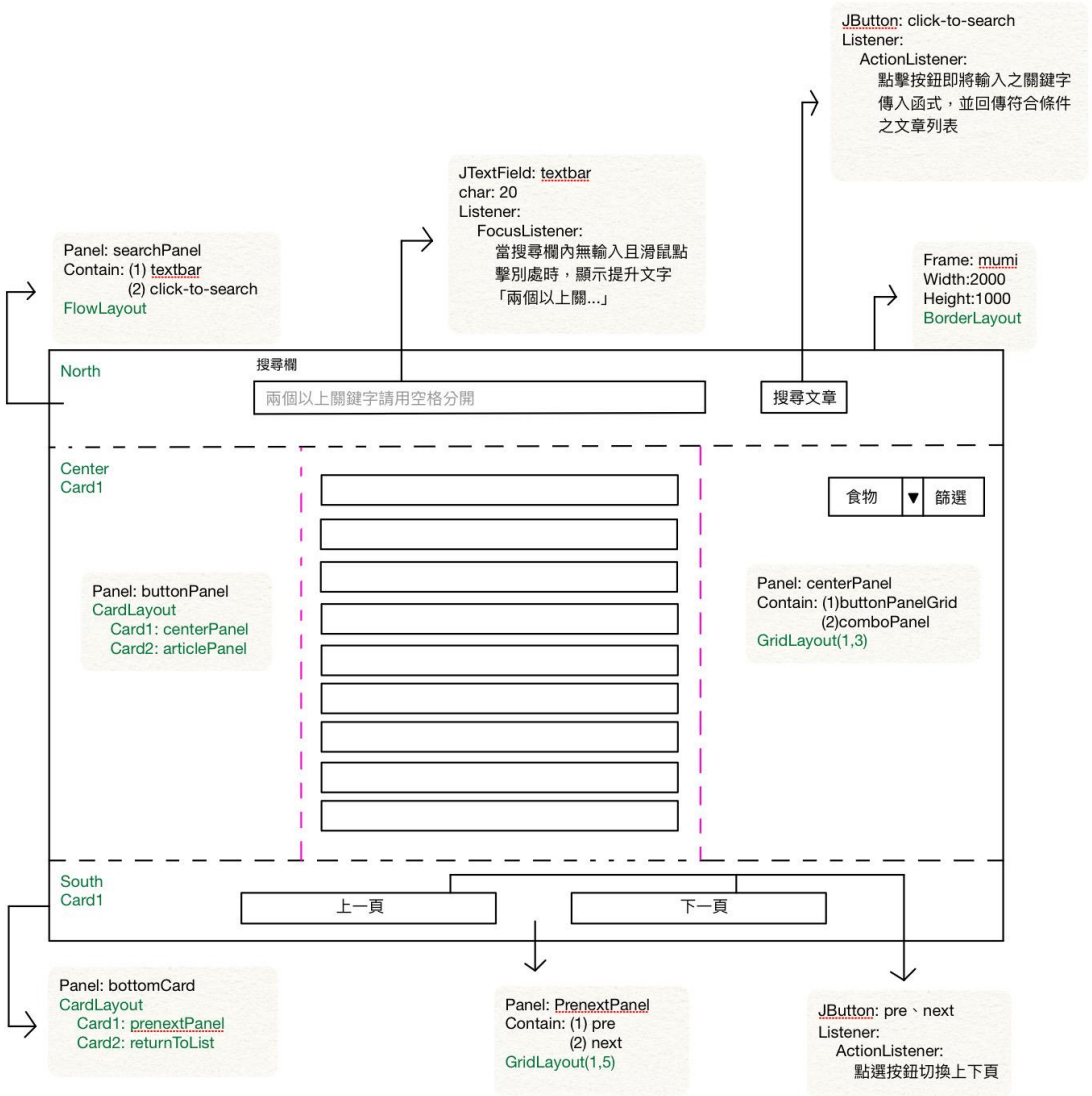
| url (unique) | title | id | sep or not |
|---|---|---|---|
| 文章網址 | 文章標題 | 文章檔名 | 文章是否以切字 |

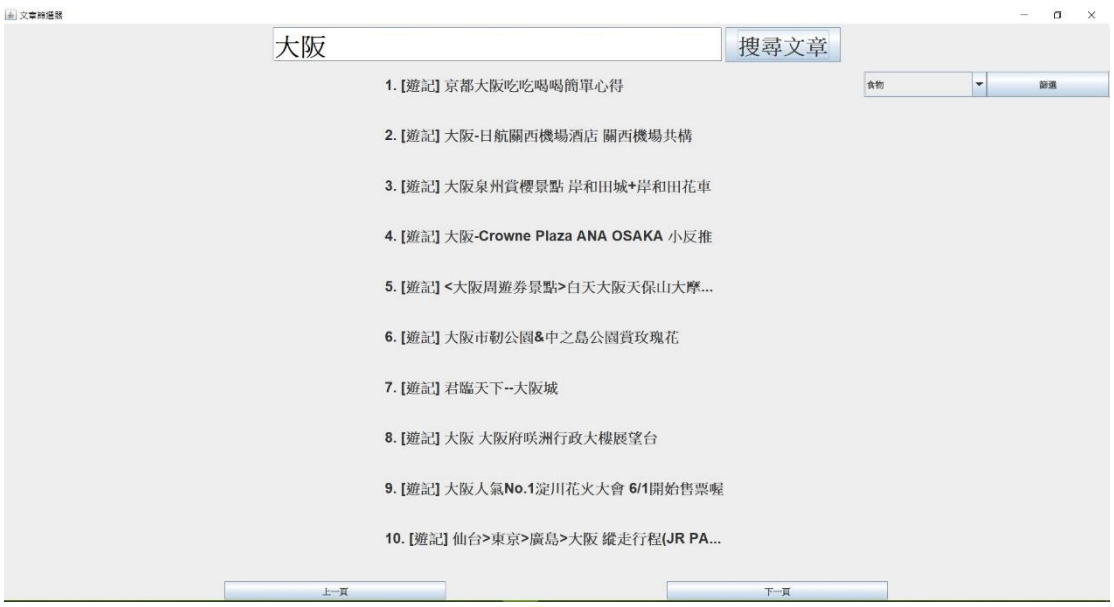
**Table: sep** (以切字)

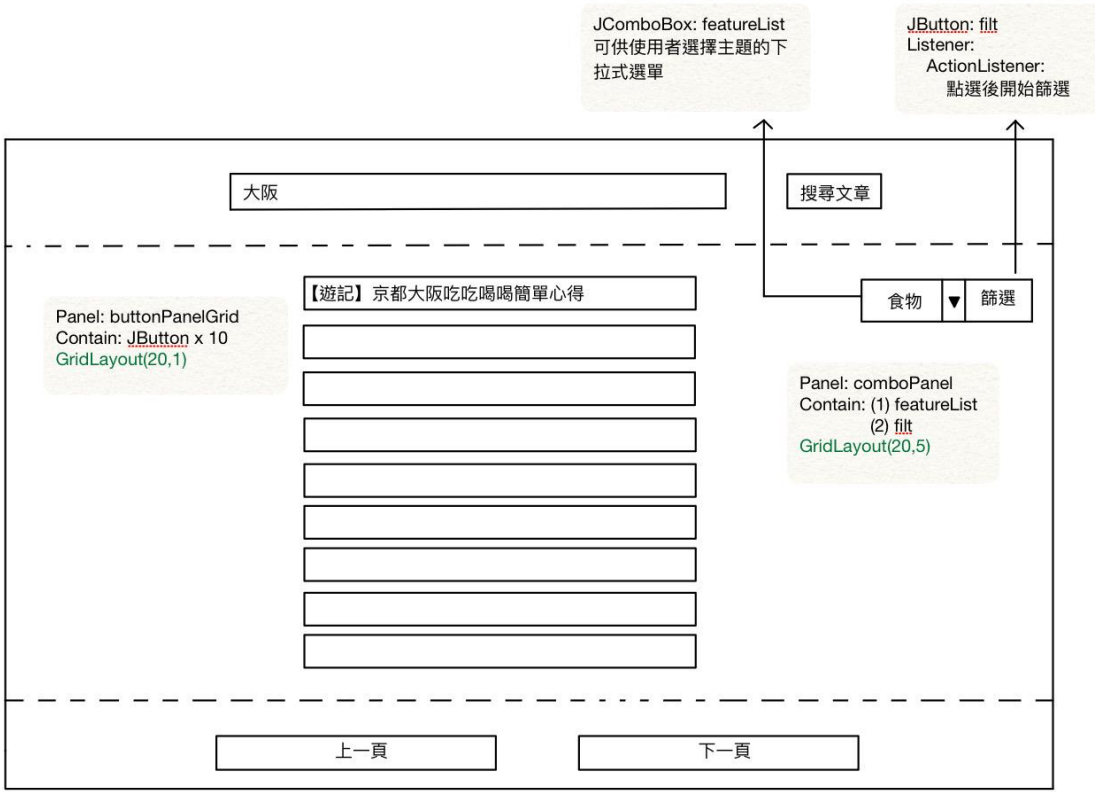
| title | id | length |
|---|---|---|
| 文章標題 | 文章檔名 | 文章長 |

# 十、GUI 介面截圖/配置簡介

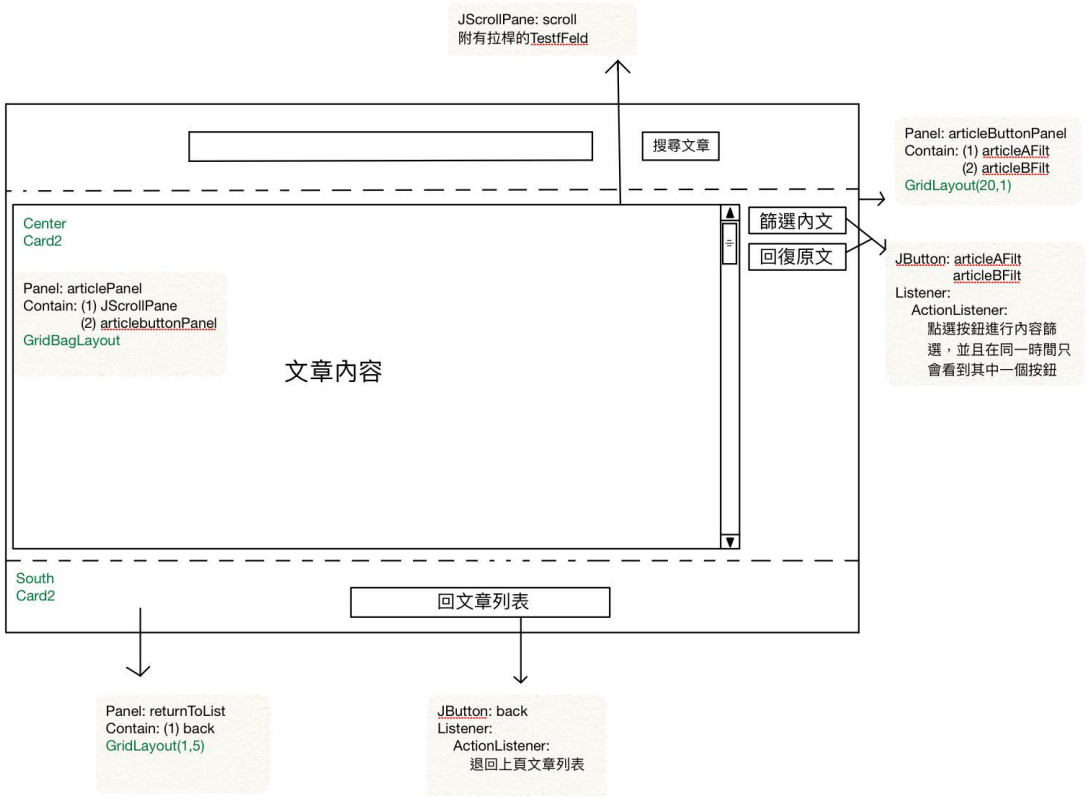



大阪
搜尋文章
食物
篩選
1. [遊記] 京都大阪吃吃喝喝簡單心得
2. [遊記] 大阪-日航關西機場酒店 關西機場共構
3. [遊記] 大阪泉州賞櫻景點 岸和田城+岸和田花車
4. [遊記] 大阪-Crowne Plaza ANA OSAKA 小反推
5. [遊記] <大阪周遊券景點>白天大阪天保山大摩...
6. [遊記] 大阪市勒公園&中之島公園賞玫瑰花
7. [遊記] 君臨天下--大阪城
8. [遊記] 大阪 大阪府咲洲行政大樓展望台
9. [遊記] 大阪人氣No.1淀川花火大會 6/1開始售票喔
10. [遊記] 仙台>東京>廣島>大阪 縱走行程(JR PA...
上一頁
下一頁


JComboBox: featureList
可供使用者選擇主題的下拉式選單
JButton: filt
Listener:
ActionListener:
點選後開始篩選
大阪
搜尋文章
【遊記】京都大阪吃吃喝喝簡單心得
食物
篩選
Panel: buttonPanelGrid
Contain: JButton x 10
GridLayout(20,1)
Panel: comboPanel
Contain: (1) featureList
(2) filt
GridLayout(20,5)
上一頁
下一頁


文章篩選器
大阪
搜尋文章
來到大阪除了去大阪難波周遭商圈逛街吃飯
附近的黑門市場也蠻值得前往
以下為黑門市場逛街+黑門市場美食心得整理
黑門市場短短600公尺的市集卻有數百家的店舖
大阪黑門市場美食相當多
主要販售海鮮、高級肉品烤肉、熟食...等
以下為石橋關東煮
たこ燒道樂
黑門三平
まぐろや黒銀
高級鮮魚魚福(奶油扁貝)....等用餐心得分享
詳細圖文版->【旅遊資訊】【旅遊時間】2015
23<20171018更新>【地點】黑門市場建議停留時間為1個半小時(含用餐)【地址】黑門市場
日本〒542-0073大阪府大阪市中央区日本橋２丁目４番１号【營業時間】8：00～18：00左右各店家不一→【用餐店鋪】2015年黑門拜訪1.石橋關東煮、大判燒、千成屋買草莓+水果、『た
心理就想著來到大阪一定也要到「黑門市場」晃晃
京都錦市場
請參考下面文章
黑門市場逛街心得黑門市場有「大阪の台所」的美稱的黑門市場
也有不少的好吃好逛的店家
店鋪的官網查詢網址如下
黑門店家檢索網站:請點我1.店家多以生猛海鮮居多
還有很多的蔬果店或者伴手禮店2.除了生猛海鮮外
也有大阪燒或道地的日本美食小吃不少肉舖都有提供現點現吃的高級牛肉
讓大家大飽口福3.有部分衣服店以及2間藥妝店家可以逛4.超級市場也有賣海鮮或者相關定食與炸物
價格會便宜許多
建議大家也可以去尋寶<下方為有逛到的店家googlemap標註地點>5.黑門市場的周遭也有不少的商圈可以逛大阪難波周遭逛街路線整理&心得分享(千日前通
難波本通商圈
黑門市場
宗右衛門町
道頓堀一蘭拉麵
國產牛燒肉放題)黑門市場交通手段▼黑門市場離難波地鐵或者南海店鐵還蠻近的
沿著南海通接著走千日前通
走路約5-10分鐘就可抵達黑門市場南海通這邊有很多店家可以逛或吃千日前通商店街這邊有許多的餐廳區
篩選內文
回文章列表


JScrollPane: scroll
附有拉桿的TestfFeld
搜尋文章
Panel: articleButtonPanel
Contain: (1) articleAFilt
(2) articleBFilt
GridLayout(20,1)
Center
Card2
篩選內文
回復原文
JButton: articleAFilt
articleBFilt
Listener:
ActionListener:
點選按鈕進行內容篩選，並且在同一時間只會看到其中一個按鈕
Panel: articlePanel
Contain: (1) JScrollPane
(2) articlebuttonPanel
GridBagLayout
文章內容
South
Card2
回文章列表
Panel: returnToList
Contain: (1) back
GridLayout(1,5)
JButton: back
Listener:
ActionListener:
退回上頁文章列表

## 十一、分工表:

張在然 B06505002:爬蟲 GUI 35%

莊博翰 B06505004:資料清理，算法研究，核心結構 55%

陳銘杰 b06505047 測試與意見反饋 10%

## 十二、心得：

莊博翰:

在本次自選題中我擔任撰寫 Project 核心部分的工作，

雖然我已經學了一兩年的 coding 這次卻是我第一次撰寫比較大的 project，這次的經驗讓我感受到，在開始撰寫 project 前的事前規劃非常重要，不然極有可能寫出一堆用不了或重複的的 code。

張在然:

在本次自選題中擔任資料收集和最後 GUI 呈現的工作，在爬蟲方面，不得不說比起 JAVA 幫我省去了許多與編碼奮戰的時間，爬下來是日文也能無壓力直接顯示真是令人舒適。而在 GUI 方面，關於如何配置各種 Layout 我認為實在是一門學問，我覺得理想的狀況應該要在做之前就先規劃好所有版面，才不至於做到一半時發現當下使用的 Layout 不符合需求而要一再更改，甚是浪費時間，也算是上到一課了。