# IR - Project Proposal

## 題目: 瀏覽器書籤全文搜索器

## 欲解決之問題

上網查資料時，常常會看到一些覺得不錯的網頁，就順手加入書籤中。但隨著書籤越來越多(以我自己的案例是將近1800筆)，要找到當初加入的書籤難度越來越高，很多時候因為害怕找不到這篇文章而將之加入書籤，但實際上加進去之後還是找不到，等同於做白工。
雖然現行瀏覽器具備書籤搜尋功能，但大多**只是根據標題搜尋**而已，如果你下的關鍵字跟標題無關，就找不到資料。

## 解決方案

我們想要做一個可以搜索各個書籤**全文**的搜尋引擎，使用者只要下關鍵字就可以全文搜尋之前加入過的書籤。讓書籤這個功能可以更好的被利用。

## 實作

### 資料取得方法

透過瀏覽器匯出書籤的方式，取得各個書籤的網頁title以及其網址，然後透過**爬蟲**將每個網頁的全文抓取下來，存成一個資料表。

### 階段一

第一步我們打算先做個雛型出來，利用**傳統建立索引**的方法，為每個網頁標註索引，並利用基本的文字命令介面做出一個可以用的方案。

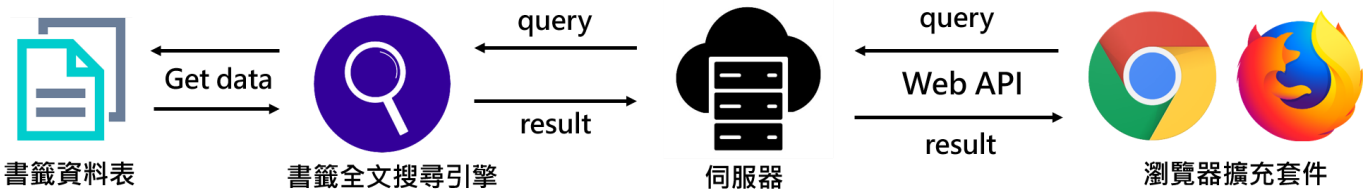
### 階段二

第二步我們透過建立詞向量模型的方式來做我們的搜尋引擎，可能使用word2vec、bert或是其他的模型，視情況挑選一個來製作。
接著我們針對傳統索引方法以及詞向量方法進行評估，了解兩者差異，以及何種方法較適用於我們這個問題，挑選出效果最好的方法。

### 階段三

第三步將做好的搜尋引擎，包裝成一個容易使用的介面，目前規劃是做成chrome以及firefox的擴充套件。
因瀏覽器套件使用JavaScript較為容易撰寫，但我們的搜尋引擎使用python製作會比較容易，因此可能會將python的部分做成Web API，透過這樣的方式讓擴充套件的JavaScript可以使用。

## 程式架構圖




## 使用技術

- 程式語言
  - Python: 搜尋引擎製作、資料爬取
  - JavaScript: 瀏覽器擴充套件製作
- 套件
  - Gensim: 用於建立索引
  - Request: 用於向伺服器發出請求，取得HTML資料
  - Beautiful Soup: 處理HTML資料
  - TensorFlow: 用於模型訓練，以及將query轉換為向量，與資料進行比對
  - Extension API: 用於製作瀏覽器擴充套件
- 資料表
  - 利用JSON格式保存