資訊檢索與文字探勘導論期末專題

專題名稱：Automated Resume Recommendations

第十二組

資管碩二	資管碩二	資管碩一	資管碩一	資管碩一
R11725048	R11725061	R12725026	R12725048	R12725054
洪立曄	孫合均	秦孝媛	楊喻妃	李婷穎

PROJECT PURPOSE

當今的求職市場競爭日益激烈，初踏入職場的新鮮人在撰寫第一份履歷時常感到迷茫，不確定如何有效地展現自己的經驗和技能，我們希望能開發一套基於文字探勘的系統，以協助求職者在撰寫履歷時有更明確的方向。

此系統能透過分析使用者輸入的職位名稱、工作內容和過去經驗，從資料庫中匹配並提供最相關的描述範例。這不僅能幫助求職者獲得靈感，還能讓他們了解特定職位所需的關鍵技能和經驗。此專案將為求職者提供寶貴的資源，幫助求職過程更加順利，有效提升求職者找到理想工作的機會。

我們選定跟資管相關度較大的職業類別（包括：Software Engineer、Project Manager、Information Technology、Data Analyst 等等）來當作我們這次期末專案的目標。

ABSTRACT

本組設計一個適用於資訊科技領域的英文履歷內容產生器，解決台灣學生撰寫英文履歷的困擾。本系統的製作過程可以分成三個階段：

一、資料蒐集：透過爬蟲整理出網路上的資訊科技領域相關的英文履歷資料。

二、前處理：將職位名稱及職位敘述做斷字、正規化前處理並轉換成 TF-IDF 向量保存在資料庫。

三、實作：第一步驟先讓使用者輸入希望撰寫的職位名稱當作 Query Term, 本系統會先推薦五個最相關的職位名稱供使用者選擇。第二步驟會根據使用者在第一步驟所選擇的職位名稱產生該職位名稱常用字的文字雲，再請使用者參考文字雲輸入想要新增的工作內容關鍵字當作額外的 Query Term。最後本系統會產生五個職位內容的描述供使用者在撰寫英文履歷時參考。

KEYWORDS

Information Retrieval,

Content Generator

CSS CONCEPTS

- Information Systems
- Information Retrieval
- Query Processing

LITERATURE REVIEW

詞頻 - 逆向檔案頻率（TF-IDF）最早於 1970 年代由資訊檢索領域的研究者提出。最初用於改善檢索結果的相關性評估，隨著時間的推移，它已被廣泛應用於各種文本相關的應用，如文檔分類、主題建模等。其中詞頻（TF）為衡量詞語在文檔中出現的頻率，逆向檔案頻率（IDF）是一個詞語普遍重要性的度量。某一特定詞語的逆向檔案頻率，可以由總檔案數目除以包含該詞語之檔案的數目，再將得到的商取以 2 為底的對數得到。詞頻 - 逆向檔案頻率是一種在資訊檢索和文本挖掘中廣泛使用的統計方法。它用於評估一個詞語在一個文檔集合或資料庫中的重要性。TF-IDF 數值會隨著詞語出現的頻率增加而增加，但會隨著詞語在語料庫中出現的文檔數目增加而下降，有助於過濾掉常見的詞語，保留重要的詞語。

DATA COLLECTION AND DATA CLEANING

我們爬取 Live Career 上的資訊科技產業相關履歷，總共 1,391 份。再透過人工篩選的方式刪除與資訊科技產業無關的職業（例如：Truck Driver、Chef 等等）並合併相似的職位名稱（例如：Information Technology Manager 與 Information Technology Director 合併為 Information Technology Manager）以及對職位名稱做正規化（例如：Software Engineering 會改成 SoftwareEngineer）。經過以上前處理後，資料庫總計有 230 個職位名稱及 10,337 個職位內容敘述。

資料舉例如下：


```
    "Cloud Big Data Architect": {
        "9362": "Devised and lead architecture and implementation of Real world evidence (RWE) Platform for more efficient data ingestion and
processing and ML.",
        "9363": "Expertise in understanding data and designing/Implementing enterprise platforms like Hadoop Data lake •Analyzed system
bottlenecks and proposed solutions to eliminate them.",
        "9364": "Fine-tuned several complex ETL Reporting applications with goal of providing faster and more efficient BI platform for business
users.",
        "9365": "Developed proof-of-concepts to reduce engineering churn.",
        "9366": "Provide review and feedback for existing physical architecture, data architecture and individual code.",
        "9367": "Debug and solve issues with Hadoop as subject matter expert.",
        "9368": "This could include things from patching components to post-mortem analysis of errors.",
        "9369": "Provide mentorship and guidance to other architects to help them become independent.",
        "9370": "Focused on issues around data science and Data processing at scale."
    },
```


圖一、職位名稱與職位內容敘述

- Information Technology Manager
- Network Support Engineer
- Information Technology Technician
- Software Engineer
- Software Developer
- Information Technology, Network Administrator
- Information Technology Intern
- Information Technology Instructor
- Information Technology support
- Senior Project Manager
- Information Technology Professional
- Desktop Engineering Supervisor
- Desktop Engineer
- Senior Information Technology Service Manager
- Information Technnology Director
- Information Technology Supervisor
- Information Technology Specialist
- Senior Software Engineer
- IT Data Analyst & Computer Programmer
- Information Technology Project Manager
- Technical Support Engineer
- Information Technology Director
- Head, Information Technology and Information Center

圖二、職位名稱

DATA PREPROCESSING AND DATA TRANSFORMATION

在使用者 Query 之前，我們會針對職位名稱及職位內容敘述做前處理並以 TF-IDF 向量的形式保存在資料庫裡。

以下會敘述此階段前處理的詳細步驟。

a. 只保留英文文本內容。
b. 將內容敘述中出現 “ IT “ 轉換為 “ Information Technology ”，才能夠對應到 TF-IDF 向量空間中。
c. 將內容全部轉換為英文小寫。
d. 刪除文本前後的空白處。
e. 移除 stopwords。
f. 將出現超過兩個空白處的地方改為一個空白處。
g. 利用 Porter’s Algorithm 進行 stemming。

完成上述前處理後，計算職位名稱及職位內容敘述的 TF-IDF 向量並保存在兩張表格中以供後續 Query 使用。

表格舉例如下：Jobtitle_tfidf.csv

存放所有職位名稱中出現的 terms 在每個職位名稱（每一個職位名稱都視為一個 Document）中的 TF-IDF 值。

docID/term	admin	administr	contractor	...
0	0	0.2556	0	
1	0	0	0	
2	0	0	0.362	
...				

圖三、職位名稱 TF-IDF 表格

表格舉例如下：Jobline_tfidf.csv

存放所有職位內容敘述中出現的 terms 在每個職位內容敘述（每一個職位內容敘述都視為一個 Document）中的 TF-IDF 值。

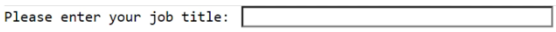
docID/term	applic	inform	transact	...
0	0	0.2556	0	
1	0	0	0	
2	0	0	0.362	
...				

圖四、職位內容敘述 TF-IDF 表格

PROJECT IMPLEMENTATION AND SYSTEM OUTCOMES

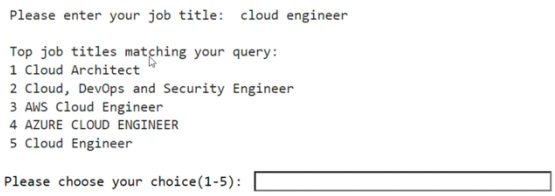
我們的檢索流程大致可分為五個步驟。

- 第一步驟：請使用者輸入想找的職位名稱當作系統的 Query Term。（如圖五所示）



圖五

- 第二步驟：系統會針對使用者所下的 Query Term （例如：Cloud Engineer）去做前處理（處理流程跟上述職位名稱跟職位內容敘述一樣），之後計算每個職位名稱的 Query Term 總 TF-IDF 分數，之後回傳 TF-IDF 分數最高的五個相關職位名稱提供使用者選擇。（如圖六所示）



圖六

- 第三步驟：當使用者選擇最想要的相關職位名稱以後（以 5 Cloud Engineer 為例），系統會產生一個文字雲，文字雲的內容是基於使用者所選擇的職位名稱當中的職位內容敘述中出現頻率最高的 term。

（如圖七所示）

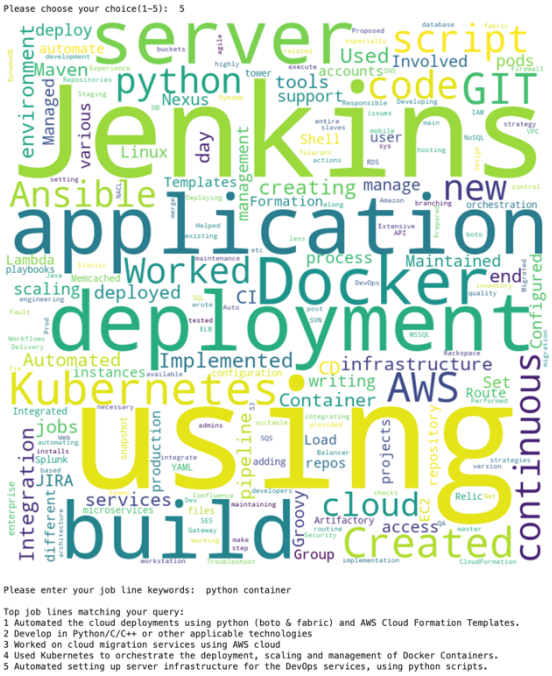
圖七

- 第四步驟：使用者可以參考本系統所產生的文字雲去新增 Query Term（以 Python、Container 為例）提供本系統去搜尋更符合需求的職位內容敘述。

（如圖七所示）

- 第五步驟：系統會結合使用者所選擇的相關職位名稱及使用者新增的 Query Term 去做 TF-IDF 排序，列出五則 TF-IDF 分數最高的職位內容敘述供使用者參考。

（如圖八所示）



圖八

CONCLUSIONS 、LIMITATIONS

本組所設計之英文履歷產生器是建立在 TF-IDF 上的一個應用。當我們事先算好 TF-IDF 值的時候，系統的執行效能十分迅速，缺點就是必須佔據記憶體空間來儲存。未來如果我們英文履歷產生器必須涵蓋資訊科技以外的其他職業，那勢必得佔用更大的儲存空間，這將是未來擴建系統的一個挑戰。而我們本

次也只使用 TF-IDF 排序來排序履歷內容，未來也可以應用其他上課所學（例如：Clustering）及文字探勘領域的其他技術（例如：Word2Vec 等）來完善我們的英文履歷推薦系統。

REFERENCES

[1] English Resume, https://www.livecareer.com/resume-search/

[2] English Resume, https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

[3] TF-IDF, SPA¨RCK JONES, K. 1972. Exhaustivity and specificity. *J. Document. 28*, 11–21.