自動文件管理系統

HSIEH,WU-CHAO
Engineering Science and Ocean Engineering
National Taiwan University
Taipei, Taiwan
b12505023@ntu.edu.tw

LIANG,CHIA-YU
Engineering Science and Ocean Engineering
National Taiwan University
Taipei, Taiwan
b12505006@ntu.edu.tw

WANG,YU-CHIEH
Engineering Science and Ocean Engineering
National Taiwan University
Taipei, Taiwan
b12505007@ntu.edu.tw

CHEN, KE-YING
Business Administration
National Taiwan University
Taipei, Taiwan
b13701222@ntu.edu.tw

Ni,Yu-Qin
Department of Philosophy
National Taiwan University
Taipei, Taiwan
kingni950417@gmail.com

***Abstract*—**本專案旨在開發一個自動化的個人數位檔案管理系統，以解決大量檔案導致的管理不易，並提升檔案整理的效率和便利性。主要問題包含未妥善歸類、重複檔案佔用大量空間以及缺乏備份。此系統的目標是根據檔案特徵進行分類，識別重複檔案並提供刪除建議，同時自動備份至雲端。而專案將使用**PyPDF2**、**python-docx**、**openpyxl** 等**Python** 套件來處理不同類型的檔案，並透過**Pandas** 及**Pillow** 進行數據和影像處理。預期成效包括檔案分類、重複檔案標示、自動刪除建議和定時同步備份。

***Index Terms*—**自動化檔案管理、自動分類、重複檔案清除

I. 要解決的問題

此專案主要目標是讓使用者能更高效且方便的管理檔案，解決以下幾個管理時易發生的問題。

A. 分類雜亂和效率不佳

a) 問題*:* 隨著數位檔案增加，許多人會發現資料夾內充滿了不同類型的檔案，如文件、圖片、影片、音樂、壓縮檔等，且常常分佈於不同資料夾中，沒有統一的分類方式。這導致搜尋特定檔案相當耗時，讓使用者在日常工作時為此困擾

b) 目標*:* 設計一個自動化的系統，根據檔案的類型、創建時間、地點等特徵自動進行分類，以提升管理效率。

B. 重複檔案佔用儲存空間

a) 問題*:* 使用者經常在不同資料夾中儲存重複檔案，尤其是下載或備份時。這些重複檔案會佔用大量儲存空間，甚至在需要使用時令人混淆。

b) 目標*:* 開發一種方式來識別重複檔案，提供清理建議或自動刪除，以節省儲存空間並保持資料夾的清晰簡潔。

C. 檔案備份和數據丟失風險

a) 問題*:* 使用者容易忘記備份重要資料，因此當電腦故障或意外刪除時，檔案丟失可能造成重大損失。

b) 目標*:* 設計自動化的備份機制，將檔案同步至雲端儲存服務，確保其不會因意外而丟失，並且無論何時何地都可以取用備份資料。

D. 專案的核心挑戰

a) 精確分類*:* 如何讓系統準確區分不同檔案的類型及其使用場景？

b) 有效識別重複檔案*:* 能否設計一種高效且準確的重複文件識別機制，減少誤刪？

II. 預計會經手的資料形態

A. 檔案格式

PDF、DOCX、XLSX、TXT、JPG、PNG

*B. Python*物件類型

Dictionary、List、DataFrame、String、Datetime

III. 介紹專案可能會使用到的其他資源

- *PyPDF2* or *pdfplumber*：讀取PDF 文件內容
- *python-docx*：讀取DOCX 文件內容
- *openpyxl* or *pandas*
- *Pillow*：處理圖片格式
- *pytesseract*：進行圖像文字的辨識

IV. 期望結果

我們期望在此專案的幫助下，提醒使用者哪些檔案長時間沒有開啓，提供一鍵刪除、識別重複下載的照片與文件並標示。同時，若使用者有私有雲端，也可以設定將特定資料夾內的資料定時同步至雲端。

V. 專案開發時程表與分工

第一週(11/13～11/17)	確認計畫細節與執行細節（是否使用版本控制工具等），並確認檔案格式需求（PDF、DOC、JPG 等）。
第二、三週(11/18～12/01)	程式碼實行階段，每週確認進度，確保與計劃書目標需求相符。
第四週(12/02～12/08)	測試執行狀況，並根據使用者的評價進行修正與優化。
第五週(12/09～12/15)	將專案修整為完成品並撰寫完整書面報告。

REFERENCES

[1] “Welcome to PyPDF2 — PyPDF2 documentation,” pypdf2.readthedocs.io. [Online]. Available: https://pypdf2.readthedocs.io/en/3.x/

[2] J. Singer-Vine, “jsvine/pdfplumber,” GitHub, Dec. 23, 2020. [Online]. Available: https://github.com/jsvine/pdfplumber

[3] “openpyxl - A Python library to read/write Excel 2010 xlsx/xlsm files — openpyxl 3.0.3 documentation,” openpyxl.readthedocs.io. [Online]. Available: https://openpyxl.readthedocs.io/en/stable/

[4] “API reference — pandas 1.3.3 documentation,” pandas.pydata.org. [Online]. Available: https://pandas.pydata.org/docs/reference/index.html#api

[5] S. Canny, “python-docx: Create, read, and update Microsoft Word .docx files.,” PyPI. [Online]. Available: https://pypi.org/project/python-docx/

[6] A. C. (PIL F. Author), “Pillow: Python Imaging Library (Fork),” PyPI. [Online]. Available: https://pypi.org/project/Pillow/

[7] M. Lee, “pytesseract: Python-tesseract is a python wrapper for Google’s Tesseract-OCR,” PyPI, Aug. 17, 2022. [Online]. Available: https://pypi.org/project/pytesseract/