這幾天我都在提升程式的效率，尤其是增加 PDF 文件時

以新增 5 份 PDF 為例

Before:

```
folder = '..\\product infomation'

import os
files = [f for f in os.listdir(folder) if f.endswith('.pdf')]

for pdf_file in files:
    db.addPDF(os.path.join(folder, pdf_file))
    tfidf.addPDF(os.path.join(folder, pdf_file))
```

✓ 4m 45.5s

原本要將近 5 分鐘

After:

```
import os
path = '..\\product infomation'
files = [file for file in os.listdir(path) if file.endswith('.pdf')]

for file in files:
    chromaAndTFIDF.addPDF(os.path.join(path, file))
```

✓ 29.5s

現在只要 30 秒

另外還有改寫相似分數的運算方式

```
chroma result:
世界上最透明的故事（日本出版界話題作，只有紙本書可以體驗的感動）.pdf / 0.14689918109712044
DE-291-1 DE-293 工作桌.pdf / 0.07183448398840939
SADES DIABLO 暗黑鬥狼RGB REALTEK 電競耳麥 7.1 (USB) SA-916.pdf / 0.06785434930668828
[折疊收納]懶人折疊桌.pdf / 0.06334593949736794
W202 人體工學椅.pdf / 0.06050530943644959
羅技 Logitech H340 USB耳機麥克風.pdf / 0.060396035834790814

tfidf result:
世界上最透明的故事（日本出版界話題作，只有紙本書可以體驗的感動）.pdf / 0.44134082769879823
羅技 Logitech H340 USB耳機麥克風.pdf / 0.0
[折疊收納]懶人折疊桌.pdf / 0.0
W202 人體工學椅.pdf / 0.0
SADES DIABLO 暗黑鬥狼RGB REALTEK 電競耳麥 7.1 (USB) SA-916.pdf / 0.0
DE-291-1 DE-293 工作桌.pdf / 0.0
```

```
hybrid result:
世界上最透明的故事（日本出版界話題作，只有紙本書可以體驗的感動）.pdf / 0.32356416905812707
DE-291-1 DE-293 工作桌.pdf / 0.028733793595363755
SADES DIABLO 暗黑鬥狼RGB REALTEK 電競耳麥 7.1 (USB) SA-916.pdf / 0.027141739722675313
[折疊收納]懶人折疊桌.pdf / 0.02533837579894718
W202 人體工學椅.pdf / 0.024202123774579837
羅技 Logitech H340 USB耳機麥克風.pdf / 0.024158414333916328
```

讓兩種分數都越高表示越相似，所以使用 chroma 和 tfidf 合併查詢時更方便，也只要由大到小排序

```python
def delPDF(self, file_name: str) -> None:
    if file_name not in self.__fileList:
        print(f'{file_name} does not exist')
        return

    # delete from chroma
    self.__collection.delete(where={"file_name": file_name})
    # self.__cromaClient.persist()

    # delete from TF-IDF
    vectorizer, matrix, fileIndex = self.__tfidfLoadData()
    del fileIndex[file_name]
    document = list(fileIndex.values())
    vectorizer = TfidfVectorizer()
    matrix = vectorizer.fit_transform(document)
    self.__tfidfStoreData(vectorizer, matrix, fileIndex)

    print(f'Deleted {file_name}')
    self.__fileCount -= 1
    del self.__fileList[self.__fileList.index(file_name)]
```

再新增刪除函式，刪除 chroma 和 tfidf 矩陣中的文件，方便管理資料