匯入電影資料

In [1]:

```
import json

# 開啟並讀取 JSON 檔案
with open("movies_data.json", "r", encoding="utf-8") as file:
    movies_data = json.load(file)

# 只輸出前三筆資料
for i, movie in enumerate(movies_data[:3]):
    print(f"資料 {i + 1}:")
    print(movie)
    print()
```

資料 1:
{'doc_id': 1, 'cname': '一世狂野', 'ename': 'Blow', 'pagerank': 1.3232355756297015e-05, 'label': ['劇情', '犯罪', '歷史/傳記'], 'intro': '喬治戎格一生都在追求所謂的美國夢，也就是享受美好富裕的生活，但是他卻不願像他父親那樣一輩子都只是個出賣勞力的建築工人。於是他搬到陽光明媚的加州，靠著販賣大麻賺錢，起初，他販毒只是為了享受自由自在的生活，但是當他野心越來越大，他的勢力也日益坐大之際，卻在此時被捕入獄。他在牢裡認識一個能言善道，自稱熟識哥倫比亞販毒集團的牢友狄亞哥，他出獄後果真把當時勢力最大的毒梟艾斯科巴介紹給喬治認識，艾斯科巴計畫將古柯鹼大量引進美國的迪斯可舞廳，希望能引領一股吸毒狂歡的風潮。除了毒品供應商之外，狄亞哥也介紹了一個美艷又狂野的女人瑪莎給喬治，他們瘋狂相愛，之後馬莎還替他生下一個可愛的女兒克莉絲汀娜，也是喬治一生的最愛。喬治很快就靠著販毒發大財，他還得買一棟大房子專門存放每天賺進來的大把鈔票，但是日進斗金卻整天提心吊膽的生活卻讓喬治開始省思，到底他要繼續過著揮霍富裕的生活，還是為了自己心愛的女兒應該轉性投資正當的事業？可是這時聯邦調查局的探員，也開始盯上毒源禍首的喬治……', 'released_date': '2001-10-12', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/1'}

資料 2:
{'doc_id': 2, 'cname': '玩命關頭', 'ename': 'The Fast and the Furious', 'pagerank': 1.3232355756297015e-05, 'label': ['動作', '劇情', '犯罪', '懸疑/驚悚'], 'intro': '唐米尼杜洛托是洛城街頭賽車界的老大哥，他身邊有一群忠心耿耿的手下，他白天忙著組裝高性能跑車，晚上則是開著他的愛車，動輒以一次一萬美元的賭注和別人軋車。布萊恩也渴望接受極速的挑戰，他對自己的駕駛技術很有信心，但是在旁觀者的眼中他只是一個菜鳥，他開了一輛超炫的跑車想和唐老大一較高下，也希望得到他的青睞，當比賽結束，布萊恩輸得一塌塗地之後，警方接獲風聲前來取締，布萊恩在無意間從一名心狠手辣的幫派份子強尼手中救了唐老大一命，於是他就被納入唐老大的權力核心，唐老大的妹妹蜜雅也對布萊恩產生好感，但是他們都不知道布萊恩其實是一名臥底警探。布萊恩滲入賽車圈的目的是調查一連串的卡車搶案，嫌犯都是開著跑車的蒙面人，警方和聯邦調查局希望能儘早逮到搶匪，以免卡車司機採取激烈的手段對這些搶匪進行報復行動，其中最有嫌疑的就是唐老大和強尼。正當唐老大和強尼形成水火不相容的情勢。布萊恩和唐老大兄妹的關係卻越來越深，他不但和唐老大結為好友，更忍不住對蜜雅產生好感，但是他也同時承受來自警方和ＦＢＩ的壓力，必須儘快查出誰才是真正的搶匪，他在天人交戰之際，在法律和友情之間，必須做出困難的決定。', 'released_date': '2001-10-13', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/2'}

資料 3:
{'doc_id': 3, 'cname': '戰雲密佈', 'ename': 'Storm Catcher', 'pagerank': 1.3232355756297015e-05, 'label': ['動作', '犯罪', '懸疑/驚悚', '戰爭'], 'intro': '美國空軍最高機密的隱形戰機驚傳失蹤！祕密訓練的飛行軍官傑克，被誣陷勾結恐怖組織，參與竊取戰機陰謀，以叛國罪名被捕入獄。他為洗刷冤屈，想盡辦法逃獄，希望重獲自由後，直搗虎穴親自討還清白。越獄後，他透過好友帕克中校的幫助暗中調查，發現這不僅是一個單純的陷阱，背後更隱藏著更大的、更難掌握的恐怖主義陰謀網，而且更令人出乎意料的是──幕後黑手竟然是五角大廈的高層官員。美國政府相關人員在傑克逃獄後，緊鎖定其妻女作為人質，逼使他必須出面投案。傑克不但必須衝破重重危機拯救妻女，又得在時限之內取得機密資料，以挽回他的名譽。然而，就在他循線直搗恐怖組織大本營，即將揭發真相之際，更迫切的是──恐怖組織已發動攻勢，企圖以最先進的隱形戰機，攻擊美國政府的權力核心白宮，一場美國有史以來最大的征戰，將一觸即發…', 'released_date': '2001-10-13', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/3'}

In [2]:

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('chinese'))
import jieba
from collections import defaultdict
import re
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\user\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

分詞

In [3]:

```
# 分詞並過濾中文停用詞
for movie in movies_data:
    # 只對 cname, label, intro 欄位進行分詞
    for field in ['cname', 'intro']:
        if field == 'label':
            words = jieba.cut(' '.join(movie[field]), cut_all=False)
        else:
            words = jieba.cut(movie[field], cut_all=False)
        # 過濾中文停用詞
        words = [w for w in words if w not in {stopwords,'、','\t', '\n', ' ', '？', '，'
        # 將分詞結果合併起來
        movie[field + '_words'] = ' '.join(words)
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\user\AppData\Local\Temp\jieba.cache
Loading model cost 0.625 seconds.
Prefix dict has been built successfully.
```

In [10]:

```
print(movies_data[1])
print(movies_data[2])
print(movies_data[3])
```

{'doc_id': 2, 'cname': '玩命關頭', 'ename': 'The Fast and the Furious', 'pagerank': 1.3232355756297015e-05, 'label': ['動作', '劇情', '犯罪', '懸疑/驚悚'], 'intro': '唐米尼杜洛托是洛城街頭賽車界的老大哥，他身邊有一群忠心耿耿的手下，他白天忙著組裝高性能跑車，晚上則是開著他的愛車，動輒以一次一萬美元的賭注和別人軋車。布萊恩也渴望接受極速的挑戰，他對自己的駕駛技術很有信心，但是在旁觀者的眼中他只是一個菜鳥，他開了一輛超炫的跑車想和唐老大一較高下，也希望得到他的青睞，當比賽結束，布萊恩輸得一塌塗地之後，警方接獲風聲前來取締，布萊恩在無意間從一名心狠手辣的幫派份子強尼手中救了唐老大一命，於是他就被納入唐老大的權力核心，唐老大的妹妹蜜雅也對布萊恩產生好感，但是他們都不知道布萊恩其實是一名臥底警探。布萊恩滲入賽車圈的目的是調查一連串的卡車搶案，嫌犯都是開著跑車的蒙面人，警方和聯邦調查局希望能儘早逮到搶匪，以免卡車司機採取激烈的手段對這些搶匪進行報復行動，其中最有嫌疑的就是唐老大和強尼。正當唐老大和強尼形成水火不相容的情勢。布萊恩和唐老大兄妹的關係卻越來越深，他不但和唐老大結為好友，更忍不住對蜜雅產生好感，但是他也同時承受來自警方和ＦＢＩ的壓力，必須儘快查出誰才是真正的搶匪，他在天人交戰之際，在法律和友情之間，必須做出困難的決定。', 'released_date': '2001-10-13', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/2', 'cname_words': '玩命 關頭', 'intro_words': '唐米尼 杜洛托 是 洛城 街頭賽 車界 老大哥 他 身邊 有 一群 忠心耿耿 手下 他 白天 忙 著 組裝 高性能 跑車 晚上 則是 開著 他 愛車 動輒 以 一次 一萬 美元 賭注 和 別人 軋車 布萊恩 也 渴望 接受 極速 挑戰 他 對 自己 駕駛技術 很 有 信心 但是 在 旁 觀者 眼中 他 只是 一個 菜 鳥 他開 了 一輛 超炫 跑 車想 和 唐 老大 一較 高下 也 希望 得到 他 青睞 當比賽 結束 布萊恩輸 得 一塌塗 地 之 後 警方 接獲 風聲 前來 取締 布萊恩 在 無意間 從 一名 心狠手辣 幫派 份子 強尼 手中 救 了 唐 老大 一命 於 是 他 就 被 納入 唐 老大 權力 核心 唐 老大 妹妹 蜜雅 也 對 布萊恩產生 好感 但是 他們 都 不 知道 布萊恩 其實 是 一名 臥底 警探 布萊恩 滲入 賽車 圈 目的 是 調查 一連串 卡車 搶案 嫌犯 都 是 開著 跑車 蒙面人 警方 和 聯邦 調查局 希望 能 儘 早 逮到 搶匪 以免 卡車 司機 採取 激烈 手段 對 這些 搶匪 進行 報 復 行動 其中 最有 嫌疑 就是 唐 老大 和 強尼 正當 唐 老大 和 強尼 形成 水火不相容 情勢 布萊恩 和 唐 老大 兄妹 的關 係 卻 越來 越深 他 不但 和 唐 老大 結為 好友 更 忍不住 對 蜜雅產生 好感 但是 他 也 同時 承受 來 自 警方 和 Ｆ Ｂ Ｉ 壓力 必 須 儘 快 查出 誰 才 是 真正 搶 匪 他 在 天人 交戰 之際 在 法律 和 友情 之間 必須 做出 困難 決定'}
{'doc_id': 3, 'cname': '戰雲密佈', 'ename': 'Storm Catcher', 'pagerank': 1.3232355756297015e-05, 'label': ['動作', '犯罪', '懸疑/驚悚', '戰爭'], 'intro': '美國空軍最高機密的隱形戰機驚傳失蹤！祕密訓練的飛行軍官傑克，被誣陷勾結恐怖組織，參與竊取戰機陰謀，以叛國罪名被捕入獄。他為洗刷冤屈，想盡辦法逃獄，希望重獲自由後，直搗虎穴親自討還清白。越獄後，他透過好友帕克中校的幫助暗中調查，發現這不僅是一個單純的陷阱，背後更隱藏著更大的、更難掌握的恐怖主義陰謀網，而且更令人出乎意料的是──幕後黑手竟然是五角大廈的高層官員。美國政府相關人員在傑克逃獄後，緊鎖定其妻女作為人質，逼使他必須出面投案。傑克不但必須衝破重重危機拯救妻女，又得在時限之內取得機密資料，以挽回他的名譽。然而，就在他循線直搗恐怖組織大本營，即將揭發真相之際，更迫切的是──恐怖組織已發動攻勢，企圖以最先進的隱形戰機，攻擊美國政府的權力核心白宮，一場美國有史以來最大的征戰，將一觸即發…', 'released_date': '2001-10-13', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/3', 'cname_words': '戰雲密佈', 'intro_words': '美國空 軍 最高 機密 隱形 戰機 驚傳 失 蹤 祕 密訓練 飛行 軍官 傑克 被 誣陷 勾結 恐怖 組織 參與 竊取 戰機 陰謀 以 叛國 罪名 被捕 入獄 他 為 洗刷 冤屈 想 盡 辦法 逃獄 希望 重獲 自由 後 直 搗 虎穴 親自討還 清白 越獄 後 他 透過 好友 帕克中 校 幫助 暗中 調查 發現 這不僅 是 一個 單純 陷阱 背後更 隱藏著 更 大 更難 掌握 恐怖 主義陰謀網 而且 更 令人 出乎意料 是 －－ 幕 後 黑手 竟然 是 五角 大廈 高層 官員 美國 政府 相關 人員 在 傑克 逃獄 後 緊鎖定 其 妻女 作為 人質 逼使 他 必須 出面 投案 傑克 不但 必須 衝破 重重 危機 拯救 妻女 又 得 在 時限之內 取得 機密 資料 以 挽回 他 名譽 然而 就 在 他 循線 直 搗 恐怖 組織 大本營 即將 揭發 真相 之際 更 迫切 是 －－ 恐怖 組織 已 發動 攻勢 企圖 以 最先 進 隱形 戰機 攻擊 美國 政府 權力 核心 白宮 一場 美國 有史 以來 最大 征戰 將一觸 即發'}
{'doc_id': 4, 'cname': '騎士風雲錄', 'ename': "A Knight's Tale", 'pagerank': 1.3232355756297015e-05, 'label': ['動作', '冒險', '喜劇'], 'intro': '14世紀中古時期的社會階級分明，出身卑微的平民不論如何努力和奮鬥，都無法跨越階級制度而翻身致富，當時正興起一種只有貴族騎士才能參加的運動「長槍比武大賽」。一位出身卑賤的年輕人威廉(希斯萊傑飾），從小就想成為一位騎士，雖然這個夢想對於貧苦的威廉是個難以實現的願望，但在父親和朋友小魏(艾倫圖克飾）、老洛(馬克艾迪飾）的鼓勵下，威廉苦練劍術和馬術。有一天威廉跟隨的貴族騎士在一次長槍比武比賽中身亡，威廉抓住機會冒充貴族的身份，展開一段勇敢追求夢想的冒險之旅。旅程中他巧遇當時身無分文的名作家喬塞(保羅貝特尼飾），喬塞為他捏造假的貴族證書，參加各地的「長槍比武大賽」，威廉憑著本身精湛

的劍術以及命運巧合的安排，一次一次的過關斬將，還邂逅了一位美麗的貴族女子喬絲琳(夏儂索莎蒙飾) 。威廉歷經種種考驗，最後不但贏得騎士比武大賽，也贏得喬絲琳的心，成了明星級的騎士。', 'released_date': '2001-10-19', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/4', 'cname_words': '騎士 風雲錄', 'intro_words': '14 世紀 中古 時期 社會 階級 分明 出身 卑微 平民 不論 如何 努力 和 奮鬥 都 無法 跨越 階級 制度 而 翻身 致富 當時 正興起 一種 只有 貴族 騎士 才能 參加 運動 「 長槍 比武 大賽 」 一位 出身 卑賤 年 輕人 威廉 希斯 萊傑飾) 從 小 就 想 成為 一位騎士 雖然 這個 夢 想 對 於 貧苦 威廉 是 個 難以 實現 願望 但 在 父親 和 朋友 小魏 艾倫 圖克飾) 老洛 馬 克 艾迪 飾) 鼓勵 下 威廉 苦練劍術 和馬術 有 一天 威廉 跟 隨的 貴族 騎士 在 一次 長 槍 比武 比賽 中 身亡 威廉 抓住 機會 冒充 貴族 身份 展開 一段 勇敢 追求 夢想 冒險 之 旅 旅程 中 他 巧遇 當時身 無 分文 名作家 喬塞 保羅貝 特尼 飾) 喬塞為 他 捏造 假 貴族 證書 參加 各地 「 長 槍 比武 大賽 」 威廉 憑著 本身 精湛 劍術 以及 命運 巧合 安排 一次 一次 過關 斬將 還 邂逅 了 一位 美麗 貴族 女子 喬絲琳 夏儂 索莎蒙飾) 威廉 歷經 種種 考驗 最 後 不但 贏得 騎士 比武 大賽 也 贏得 喬絲琳 心 成 了 明星 級 騎士'}

計算TF-IDF 並存起來

In [5]:

```
import collections
import pandas as pd
import math

# 初始化變量
record_set = set()
idf_count_dict = collections.defaultdict(int)
num_articles = 0

# 計算 IDF 值
for d in movies_data:
    if len(d["label"]) != 0:
        tokens = d["intro_words"]
        # 計算每個單詞的 IDF
        for token in set(tokens):
            record_set.add(token)
            idf_count_dict[token] += 1
        num_articles += 1
    if num_articles == 6000:
        break

x_data, y_data = [], []
# 計算每個單詞的 TF-IDF 值
for d in movies_data:
    if len(d["label"]) != 0:
        # 計算單詞出現次數
        chart = collections.Counter(d["intro_words"])
        temp_dict = {}
        # 計算 TF-IDF
        for w, n in chart.items():
            tf = round(n / sum(chart.values()), 4)
            idf = round(num_articles / idf_count_dict[w], 4)
            temp_dict[w] = round(tf * math.log(idf, 10), 4)

        x_data.append(temp_dict)
        y_data.append({"label": d["label"][0]})

        if len(y_data) % 100 == 0:
            if len(y_data) == 6000:
                break

print("Creating x...")
# 將 x_data 轉換為 DataFrame
x_df = pd.DataFrame(x_data, columns=list(record_set))
x_df = x_df.fillna(0)
print("Creating y...")
# 將 y_data 轉換為 DataFrame
y_df = pd.DataFrame(y_data, columns=["label"])
```

```
Creating x...
Creating y...
```

訓練模型KNN SVM RF

In [6]:

```
x_train, x_test = x_df[: -500], x_df[-500: ]
y_train, y_test = y_df[: -500], y_df[-500: ]
```

In [7]:

```
from sklearn.neighbors import KNeighborsClassifier

KNN = KNeighborsClassifier()
print("Training KNN")
KNN.fit(x_train, y_train)
print("Predicting KNN")
result1 = KNN.predict(x_test)

from sklearn import ensemble

RF = ensemble.RandomForestClassifier(n_estimators = 100)
print("Training RF")
RF.fit(x_train, y_train)
print("Predicting RF")
result2 = RF.predict(x_test)

from sklearn import svm
SVM = svm.SVC()
print("Training SVM")
SVM.fit(x_train, y_train)
print("Predicting SVM")
result3 = SVM.predict(x_test)
```

```
Training KNN
Predicting KNN

C:\Users\user\anaconda3\lib\site-packages\sklearn\neighbors\_classificatio
n.py:198: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples,), for exampl
e using ravel().
  return self._fit(X, y)
C:\Users\user\anaconda3\lib\site-packages\sklearn\neighbors\_classificatio
n.py:228: FutureWarning: Unlike other reduction functions (e.g. `skew`, `k
urtosis`), the default behavior of `mode` typically preserves the axis it
acts along. In SciPy 1.11.0, this behavior will change: the default value
of `keepdims` will become False, the `axis` over which the statistic is ta
ken will be eliminated, and the value None will no longer be accepted. Set
`keepdims` to True or False to avoid this warning.
  mode, _ = stats.mode(_y[neigh_ind, k], axis=1)
C:\Users\user\AppData\Local\Temp\ipykernel_5792\1509355256.py:13: DataConv
ersionWarning: A column-vector y was passed when a 1d array was expected.
Please change the shape of y to (n_samples,), for example using ravel().
  RF.fit(x_train, y_train)

Training RF
Predicting RF
Training SVM

C:\Users\user\anaconda3\lib\site-packages\sklearn\utils\validation.py:993:
DataConversionWarning: A column-vector y was passed when a 1d array was ex
pected. Please change the shape of y to (n_samples, ), for example using r
avel().
  y = column_or_1d(y, warn=True)

Predicting SVM
```

In [8]:

```
def evaluate_score(prediction, answer):
    n = 0
    for i in range(len(prediction)):
        if (prediction[i] == answer._get_value(5500 + i, "label")):
            n += 1

    return n / len(prediction)
```

結果

In [9]:

```
print(f"KNN score: {evaluate_score(result1, y_test)}")
print(f"RF score: {evaluate_score(result2, y_test)}")
print(f"SVM score: {evaluate_score(result3, y_test)}")
```

```
KNN score: 0.296
RF score: 0.496
SVM score: 0.538
```

有嘗試用9500筆訓練，發現準確率更低

KNN score: 0.29 RF score: 0.47 SVM score: 0.468 因此最後選擇用5500筆做訓練的模型來預測後500筆資料

In [ ]:

```

```