In [1]:

```
pip install jieba
```

```
Requirement already satisfied: jieba in c:\users\user\anaconda3\lib\site-p
ackages (0.42.1)
Note: you may need to restart the kernel to use updated packages.
```

In [6]:

```
class Movie:
    def __init__(self, name, intro, label) -> None:
        self.name = name
        self.intro = intro
        self.label = label
```

In [7]:

```
import json
path = "hw3.json"
movie_data = []
movies = []

with open(path) as f:
    movie_data = json.load(f)

for movie in movie_data:
    try:
        label = movie['label'][0]
    except:
        label = 'NA'
    m = Movie(movie['cname'], movie['intro'], label)
    movies.append(m)

print(movie_data[0])
print(movies[0].label)
```

```
{'doc_id': 1, 'cname': '一世狂野', 'ename': 'Blow', 'pagerank': 1.14899966
5350786e-08, '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': '2001-10-1
2', 'links': [14848, 14208, 14690, 14850, 14849, 13733, 14822, 14245, 1482
4, 14944, 14301, 13749, 14557, 14618, 14299, 14652, 14653, 14558]}
劇情
```

In [8]:

```
with open("inverted.json") as f:
    inverted_index = json.load(f)

key_map = {}
x = 0

for key in inverted_index.keys():
    if not key in key_map.keys():
        key_map[key] = x
        x += 1
```

In [9]:

```
import pandas as pd
import jieba

dataframe = pd.DataFrame()
i = 0
for movie in movies:
    seg_list = []
    for seg in jieba.cut_for_search(movie.name):
        seg_list.append(key_map[seg])
    for seg in jieba.cut_for_search(movie.intro):
        seg_list.append(key_map[seg])
    df = pd.DataFrame({'name':[movie.name], 'label':[movie.label]})
    df = pd.concat([df, pd.DataFrame([seg_list])], axis=1)
    dataframe = pd.concat([dataframe, df], ignore_index=True)
    i += 1
    if i == 6000:
        break

dataframe.fillna(-1, inplace=True)

dataframe
```

Out[9]:

	name	label	0	1	2	3	4	5	6	7	...
0	一世狂野	劇情	0	1	2	3.0	4.0	5.0	6.0	7.0	...
1	玩命關頭	動作	169	170	171	172.0	35.0	173.0	174.0	175.0	...
2	戰雲密佈	動作	339	340	341	342.0	343.0	344.0	9.0	345.0	...
3	騎士風雲錄	動作	448	449	450	451.0	452.0	453.0	9.0	454.0	...
4	金法尤物	喜劇	571	572	6	573.0	574.0	575.0	389.0	573.0	...
...	...	...	...	...	...	...	...	...	...	...	...
5995	少年(2017)	動作	1397	488	89504	840.0	4910.0	693.0	7550.0	396.0	...
5996	兒子的完美告別	劇情	1345	9	4638	8504.0	4910.0	19747.0	8760.0	2185.0	...
5997	絕愛日本：篠田正浩的大和浮世繪	NA	171791	4497	964	59980.0	148605.0	9.0	136.0	203.0	...
5998	夏日，戀愛-市總圖藝術電影院	劇情	4584	12	9023	2948.0	167773.0	77444.0	10838.0	3268.0	...
5999	萌牛費迪南	動畫	171973	102707	4910	693.0	1488.0	21809.0	26185.0	1489.0	...

6000 rows × 8082 columns

In [10]:

```
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

dataframe.columns = dataframe.columns.astype(str)

X_data = dataframe.drop(columns=["name", "label"])
y_data = dataframe["label"]

X_data

X_train, X_test, y_train, y_test = train_test_split(X_data, y_data, test_size=0.017)
clf = KNeighborsClassifier()
clf.fit(X_train,y_train)
clf.score(X_test,y_test)
```

Out[10]:

```
0.30097087378640774
```

In [11]:

```
from sklearn import tree

clf = tree.DecisionTreeClassifier()
clf.fit(X_train,y_train)
clf.score(X_test,y_test)
```

Out[11]:

```
0.20388349514563106
```

In [ ]:

```

```