In [1]:

```
import json
import jieba
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split

# 讀取json文件
with open('C:/nlp2023/hw2/movies_data_with_page_ranks.json', 'r', encoding='utf-8') as f
    data = json.load(f)

# 分詞處理
jieba.setLogLevel(20)  # 防止jieba輸出警告信息
for movie in data:
    movie['intro'] = ' '.join(jieba.cut(movie['intro'], cut_all=False))

# 提取特徵向量
tfidf = TfidfVectorizer()
X = tfidf.fit_transform([movie['intro'] for movie in data])
y = [movie['label'][0] if movie['label'] else 'NA' for movie in data]

# 劃分訓練集和測試集
test_size = 500/len(data)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size, random_st
```

In [6]:

```
knn = KNeighborsClassifier(n_neighbors=30)
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
precision_knn = sum(y_pred_knn[i] == y_test[i] for i in range(len(y_test))) / len(y_test
print('KNN：', precision_knn)
```

```
KNN： 0.456

C:\Users\iwin4\anaconda3\lib\site-packages\sklearn\neighbors\_classificati
on.py:228: FutureWarning: Unlike other reduction functions (e.g. `skew`, `
kurtosis`), 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)
```

In [5]:

```
svm = SVC(C=3.0)
svm.fit(X_train, y_train)
y_pred_svm = svm.predict(X_test)
precision_svm = sum(y_pred_svm[i] == y_test[i] for i in range(len(y_test))) / len(y_test
print('SVM：', precision_svm)
```

```
SVM： 0.508
```

In [ ]:

```
```