B0928002 林力

In [1]:

```
import json
with open("movie_full.json") as fp:
    movies = json.load(fp)
fp.close()
print (len(movies))
```

```
6000
```

In [2]:

```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /Users/linli/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[2]:

```
True
```

In [3]:

```
import jieba
import re
from zhon.hanzi import punctuation
from nltk.corpus import stopwords

punctuations = list(punctuation)
stopword = [line.strip().replace('\n', '') for line in open('./needs/stopwor
punctuations = [line.strip().replace('\n', '') for line in open('./needs/pun
jieba.load_userdict('./needs/userDict.txt')
stop_words = stopwords.words("chinese")
```

```
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/hd/b9ptqs9s7wgd9115fq_x_b100000gn/T/ji
eba.cache
Loading model cost 0.520 seconds.
Prefix dict has been built successfully.
```

In [4]:

```
DATA =[]

for movie in movies:
    movie['intro'] = re.sub('[^\u4e00-\u9fa5]','',movie['intro'])
    info = movie['intro'].replace('\n','').replace('\t','').replace('\u3000'
    info = jieba.lcut(info)
    Info = [i for i in list(info) if i not in stopword]
    Info = [i for i in list(info) if i not in stop_words]

    info = [" ".join(Info)]
    info.insert(0, movie['class_label'][0] if len(movie['class_label'])>0 el

    DATA.append(info)
```

In [5]:

```
print (DATA[1])
```

```
['愛情', '奧斯卡 影帝 安東尼 霍普金斯 瑞秋懷茲 裘德洛 班佛 斯特 眾星 雲集 主演 無法無天
奧斯卡 大 導演 佛南度 梅 瑞爾 斯 執導 世紀 生動 充滿 懸念 感人至深 愛情 傳說 王冠 兩度
獲得奧斯卡 金像獎 提名 編劇 彼得 摩根 倫敦 影展 最佳影片 入選 改編 奧地利 劇作家 施尼 茨
勒 輪舞 現代社會 中 家庭 夫妻 的問題 美國 綜藝 報 影片 清楚地 表達 觀點 報導 一名 斯洛伐
克 女子 為 脫貧 當娼 揭開序幕 擔任 汽車公司 高層 主管 維也納 出差 想 背妻 偷吃 卻 慘遭
勒索 而其 妻子 出軌 搭上 巴西 攝影師 女兒 失蹤 多年 老父親 英國 飛到 美國 尋 女兒 途中
遇上 與 男友 分手 巴西 女子 風雪 阻擋 留在 洛杉磯 機場 回教 牙醫 愛上 有夫之婦 而她 俄羅
斯 黑幫 老公 黑吃黑 事件中 有了 一件 選擇 城市 男女 編織 出 簡單 卻又 令人 癡 醉的 多線
敘事 故事 人生 分岔 口 不確定性 令 生命 更有 期望 關於 電影 無法無天 奧斯卡 大 導演 佛南
度 梅 瑞爾 斯 執導 國際合作 戲劇 驚悚 電影 彼得 摩根 改編自 奧地利 劇作家 亞瑟 施尼 茨
勒 劇作 輪舞 導演 編劇 將原 著中 摩登 不停 變化 螺旋 敘事 透過 劇中 人物 串聯 來來 不同
的 國家 城市 講述 世紀 生動 充滿 懸念 感人至深 愛情 傳說 故事 起自 維也納 影片 美妙的 將
巴黎 倫敦 斯洛伐克 首都 巴西 里約 美國 丹佛 鳳凰城 城市的 人們 編織 出 一幅 簡單 卻又 令
人 癡 醉的 敘述 編織 出 令人 痴迷 命運 輪迴 色局 追 兇 電影 色局 追 兇 是一個 關係 交錯
愛情故事 有如 火線 交錯 與 蝴蝶效應 綜合體 各條 支線 看似 獨立 發展 卻又 莫名 交集 隱約
中 發現 個中 因果 循環 每個 主角 做的 決定 像是 蝴蝶效應 隨時 都可能 影響到 關係 值得一
提的是 全片 影帝 影后 主演 與 導演 第二次 合作的 瑞秋懷茲 信手拈來 隨便 都是 戲 演出 一
位 外遇 裡 掙扎 婦女 裘德洛 本片 演技 極為 出色 資深 奧斯卡 影帝 安東尼 霍普金斯 將 尋找
失蹤 女兒 父親 詮釋 相當 完美 全片 部 咀嚼 有味 作品 生命 就如 同一個 大 圓弧 浮世 男女
都 沉醉在 愛慾 沉淪 中 偶然之間 會 擦肩而過 看似 悲傷 卻 代表 著 一段 關係 結束 卻又 可
能是 一個人 正規的 省思 每個人 都會 愛情 圓圈 裡 找到 自己的 幸福']
```

In [6]:

```python
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
# X = [data[1] for data in DATA]
# y = [data[0] for data in DATA]
# train_X, test_X, train_y, test_y = train_test_split(X, y, test_size=0.0833
# TFIDF = TfidfVectorizer(token_pattern =r"(?u)\b\w\w+\b")
# train_X = TFIDF.fit_transform(train_x)
# train_X = TFIDF.transform(train_x)
# train_Y = TFIDF.fit_transform(train_y)
# train_Y = TFIDF.transform(train_y)

train_x = [data[1] for data in DATA]
train_y = [data[0] for data in DATA]
TFIDF = TfidfVectorizer(token_pattern =r"(?u)\b\w\w+\b")
train_X = TFIDF.fit_transform(train_x)
train_X = TFIDF.transform(train_x)
```

In [7]:

```python
train_X, test_X, train_y, test_y = train_test_split(train_X, train_y, test_s
```

In [8]:

```python
print (len(train_x))
print (len(train_y))
print (train_X.shape[0])
print (test_X.shape[0])
print (len(test_y))
```

```
6000
5500
5500
500
500
```

In [9]:

```python
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn import metrics
```

In [10]:

```python
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier(n_neighbors=30)
KNN.fit(train_X, train_y)
y_pred_KNN = KNN.predict(test_X)

test_y_predicted = KNN.predict(test_X)
```

```python
accuracy = metrics.accuracy_score(test_y, test_y_predicted)
print ('KNN')
print ('Predicttion Precision: ', (accuracy*100).round(), '%')
print ('')

print (classification_report(test_y, y_pred_KNN))
```

```
KNN
Predicttion Precision:  51.0 %

              precision    recall  f1-score   support

          冒險       0.00      0.00      0.00         6
          劇情       0.41      0.85      0.56       142
          動作       0.56      0.49      0.52        49
          動畫       0.76      0.58      0.66        43
          勵志       0.00      0.00      0.00         5
          喜劇       0.00      0.00      0.00        11
          奇幻       0.00      0.00      0.00         7
          影展       0.67      0.25      0.36         8
          恐怖       0.70      0.47      0.57        40
          愛情       0.63      0.60      0.62        93
       懸疑/驚悚       0.00      0.00      0.00        13
          戰爭       0.00      0.00      0.00         3
          武俠       0.00      0.00      0.00         1
       歷史/傳記       0.00      0.00      0.00         8
       溫馨/家庭       0.00      0.00      0.00        13
          犯罪       0.00      0.00      0.00         9
          科幻       0.00      0.00      0.00         9
         紀錄片       0.83      0.19      0.31        26
       音樂/歌舞       1.00      0.14      0.25        14

    accuracy                           0.51       500
   macro avg       0.29      0.19      0.20       500
weighted avg       0.49      0.51      0.45       500
```

```
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/neighbors/_classification.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 taken 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)
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In [11]:

```python
from sklearn import tree
clf = tree.DecisionTreeClassifier()
clf.fit(train_X, train_y)
y_pred_Dec = clf.predict(test_X)
```

```python
test_y_predicted = clf.predict(test_X)
accuracy = metrics.accuracy_score(test_y, test_y_predicted)
print ("DecisionTree")
print ('Predicttion Precision: ', (accuracy*100).round(), '%')
print ('')

print (classification_report(test_y, y_pred_Dec))
```

```
DecisionTree
Predicttion Precision:  42.0 %

              precision    recall  f1-score   support

          冒險       0.00      0.00      0.00         6
          劇情       0.39      0.50      0.44       142
          動作       0.38      0.35      0.36        49
          動畫       0.57      0.63      0.60        43
          勵志       0.00      0.00      0.00         5
          喜劇       0.29      0.18      0.22        11
          奇幻       0.12      0.14      0.13         7
          影展       0.50      0.25      0.33         8
          恐怖       0.56      0.47      0.51        40
          愛情       0.55      0.56      0.56        93
       懸疑/驚悚       0.11      0.08      0.09        13
          戰爭       0.00      0.00      0.00         3
          武俠       0.00      0.00      0.00         1
       歷史/傳記       0.00      0.00      0.00         8
       溫馨/家庭       0.00      0.00      0.00        13
          犯罪       0.00      0.00      0.00         9
          科幻       0.22      0.22      0.22         9
         紀錄片       0.42      0.38      0.40        26
       音樂/歌舞       0.50      0.36      0.42        14

    accuracy                           0.42       500
   macro avg       0.24      0.22      0.23       500
weighted avg       0.40      0.42      0.41       500
```

```
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In [12]:

```python
from sklearn import svm

linearSvcModel = svm.LinearSVC(C=1, max_iter=10000)
linearSvcModel.fit(train_X, train_y)
y_pred_SVM = linearSvcModel.predict(test_X)

test_y_predicted = linearSvcModel.predict(test_X)
accuracy = metrics.accuracy_score(test_y, test_y_predicted)
print ("SVM")
print ('Predicttion Precision: ', (accuracy*100).round(), '%')
```

```
print ('')

print (classification_report(test_y, y_pred_SVM))
```

```
SVM
Predicttion Precision:  59.0 %

              precision    recall  f1-score   support

          冒險       0.00      0.00      0.00         6
          劇情       0.49      0.80      0.61       142
          動作       0.59      0.55      0.57        49
          動畫       0.81      0.79      0.80        43
          勵志       0.00      0.00      0.00         5
          喜劇       1.00      0.09      0.17        11
          奇幻       1.00      0.14      0.25         7
          影展       0.75      0.38      0.50         8
          恐怖       0.78      0.78      0.78        40
          愛情       0.64      0.72      0.68        93
       懸疑/驚悚       0.00      0.00      0.00        13
          戰爭       0.50      0.33      0.40         3
          武俠       0.00      0.00      0.00         1
       歷史/傳記       1.00      0.12      0.22         8
       溫馨/家庭       0.00      0.00      0.00        13
          犯罪       0.00      0.00      0.00         9
          科幻       0.50      0.22      0.31         9
         紀錄片       0.62      0.38      0.48        26
       音樂/歌舞       0.83      0.36      0.50        14

    accuracy                           0.59       500
   macro avg       0.50      0.30      0.33       500
weighted avg       0.58      0.59      0.55       500
```

```
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_clas
sification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-de
fined and being set to 0.0 in labels with no predicted samples. Use `zero_di
vision` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_clas
sification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-de
fined and being set to 0.0 in labels with no predicted samples. Use `zero_di
vision` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_clas
sification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-de
fined and being set to 0.0 in labels with no predicted samples. Use `zero_di
vision` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In [13]:

```
from sklearn.naive_bayes import MultinomialNB
naive_bayes_classifier = MultinomialNB()
naive_bayes_classifier.fit(train_X, train_y)
y_pred_NB = naive_bayes_classifier.predict(test_X)

test_y_predicted = naive_bayes_classifier.predict(test_X)
accuracy = metrics.accuracy_score(test_y, test_y_predicted)
print ('MultinomialNB')
print ('Predicttion Precision: ', (accuracy*100).round(), '%')
print ('')

print (classification_report(test_y, y_pred_NB))
```

```
MultinomialNB
Predicttion Precision:  33.0 %

              precision    recall  f1-score   support

          冒險       0.00      0.00      0.00         6
          劇情       0.30      0.99      0.46       142
          動作       1.00      0.02      0.04        49
          動畫       0.00      0.00      0.00        43
          勵志       0.00      0.00      0.00         5
          喜劇       0.00      0.00      0.00        11
          奇幻       0.00      0.00      0.00         7
          影展       0.00      0.00      0.00         8
          恐怖       0.00      0.00      0.00        40
          愛情       0.85      0.24      0.37        93
       懸疑/驚悚       0.00      0.00      0.00        13
          戰爭       0.00      0.00      0.00         3
          武俠       0.00      0.00      0.00         1
       歷史/傳記       0.00      0.00      0.00         8
       溫馨/家庭       0.00      0.00      0.00        13
          犯罪       0.00      0.00      0.00         9
          科幻       0.00      0.00      0.00         9
         紀錄片       0.00      0.00      0.00        26
       音樂/歌舞       0.00      0.00      0.00        14

    accuracy                           0.33       500
   macro avg       0.11      0.07      0.05       500
weighted avg       0.34      0.33      0.20       500
```

```
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_clas
sification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-de
fined and being set to 0.0 in labels with no predicted samples. Use `zero_di
vision` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_clas
sification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-de
fined and being set to 0.0 in labels with no predicted samples. Use `zero_di
vision` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/linli/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_clas
sification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-de
fined and being set to 0.0 in labels with no predicted samples. Use `zero_di
vision` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

```
In [ ]:
```