PoemClassifier

2018 年 7 月 5 日

1 單純貝氏 (詩詞分類)

1.1 介紹

我們上一章節談到了單純貝氏，以及歸類文章，現在我們再試試看

但是這次我們拿的資料是比較抽象的詩詞資料，我們看看我的單純貝氏是否能研究出三位詩人的常用語

對無名詩詞做出分類

1.2 資料集

自行收集的詩詞資料集

https://drive.google.com/open?id=1KuH3QyTaD7yrqGv8uTPPVgDQD_XW8amC

poem_train.csv: 供你訓練模型

poem_test.csv: 供你測試模型

1.3 目標

標籤總共有三種:

1.3.1 李白

風格 (節錄自 wiki):

李詩富個性，有強烈的主觀抒情色彩，內容表現出蔑視庸俗，反抗和不媚權貴的叛逆精神，歌頌遊俠和仙道，被譽為「詩俠」、「詩仙」，後世亦以詩仙李白稱之。

1.3.2 杜甫

風格 (節錄自 wiki):

杜詩主要風格是沉鬱頓挫，氣魄闊大雄偉，詩歌意象鮮明強烈。風格多樣，豐富多姿，或雄渾奔放，或清新細膩，或沉鬱悲涼、或辭藻富麗、或平易質樸、或通俗自然。杜詩融冶吸收前人藝術技巧，發展成一種獨特的新風格。

1.3.3 白居易

風格 (節錄自 wiki):

開頭破題，在結尾時凸顯全詩要旨；用辭淺顯，使人容易明瞭；用語直接而銳利，使人警惕；敘事可靠可信；體例流暢而可以傳唱於歌曲之中。

1.4 資料預處理

1.4.1 資料讀取

csv 的讀取就直接使用 pandas read_csv 即可

```
In [15]: import pandas as pd

         # 為了顯示的漂亮，我刻意的把印出來的 row 只顯示 15 個和 column 只顯示十個
         # 大家練習的時候可以去掉下面兩行
         pd.set_option('display.max_rows', 15)
         pd.set_option('display.max_columns', 10)

         # scikit-learn 會有些 deprecation warning，為了顯示漂亮，我刻意地忽略掉
         import warnings
         warnings.filterwarnings('ignore')
```

```
df = pd.read_csv("poem_train.csv", encoding = "utf-8")
df
```

Out[15]:

	作者	詩名 \
0	李白	菩薩蠻 · 平林漠漠煙如織
1	李白	把酒問月
2	李白	春思
3	李白	春夜洛城聞笛
4	李白	古風 其十九
5	李白	關山月
6	李白	將進酒 · 君不見黃河之水天上來
...	...	...
2724	白居易	秋蝶
2725	白居易	三年為刺史二首
2726	白居易	彆元九後詠所懷
2727	白居易	早秋曲江感懷
2728	白居易	東墟晚歌　時退居渭村。
2729	白居易	南秦雪
2730	白居易	寄蘄州簟與元九因題六韻　時元九鰥居。

	內容
0	平林漠漠煙如織，寒山一帶傷心碧。\r\n 暝色入高樓，有人樓上愁。玉階空佇立，宿鳥歸飛急
1	青天有月來幾時，我今停杯一問之：人攀明月不可得，月行卻與人相隨？皎如飛鏡臨丹闕，綠煙
2	燕草如碧絲，秦桑低綠枝。當君懷歸日，是妾斷腸時。春風不相識，何事入羅幃
3	誰家玉笛暗飛聲，散入春風滿洛城。此夜曲中聞折柳，何人不起故園情。
4	西上蓮花山，迢迢見明星。(西上 一作：西嶽) 素手把芙蓉，虛步躡太清。霓裳曳廣帶，飄拂升
5	明月出天山，蒼茫雲海間。長風幾萬裡，吹度玉門關。漢下白登道，胡窺青海灣。由來征戰地，
6	君不見黃河之水天上來，奔流到海不複回。君不見高堂明鏡悲白發，朝如青絲暮成雪。人生得意
...	...
2724	秋花紫蒙蒙，秋蝶黃茸茸。花低蝶新小，飛戲叢西東。日暮涼風來，紛紛花落叢。夜深白露冷，
2725	三年為刺史，無政在人口。唯向郡城中，題詩十餘首。慚非甘棠詠，豈有思人不？三年為刺史，
2726	零落桐葉雨，蕭條槿花風。悠悠早秋意，生此幽閒中。況與故人彆，中懷正無悰。勿雲不相送，
2727	離離暑雲散，嫋嫋涼風起。池上秋又來，荷花半成子。朱顏易銷歇，白日無窮已。人壽不如山，
2728	涼風冷露蕭索天，黃蒿紫菊荒涼田。繞塚秋花少顏色，細蟲小蝶飛翻翻。中有騰騰獨行者，手扌
2729	往歲曾為西邑吏，慣從駱口到南秦。\r\n 三時雲冷多飛雪，二月山寒少有春。\r\n 我思舊事猶
2730	笛竹出蘄春，霜刀劈翠筠。織成雙鎖簟，寄與獨眠人。卷作筒中信，舒為席上珍。滑如鋪薤葉，

```
[2731 rows x 3 columns]
```

1.4.2 類別處理

由於 scikit-learn 不接受字串，所以我們一定要把類別轉換成整數
你可以使用 cat.categories 得到所有類別
再使用 cat.codes 轉換成整數

```
In [16]: df['作者'] = df['作者'].astype('category')
         saved_map = { cat:df['作者'].cat.categories.get_loc(cat) for cat in df['作者'].cat.cat
         saved_map

Out[16]: {' 李白': 0, ' 杜甫': 1, ' 白居易': 2}
```

1.4.3 分詞處理

我們直接把分詞和簡易的內容處理定義成一個函式

```
In [17]: # 定義好等等我們要對所有內容做的 split
         def split_poem(poem):
             return " ".join(jieba.cut(poem))

         def process_poems(df):
             df['內容'] = df['內容'].apply(split_poem)
             df['內容'] = df['內容'].str.replace('\r', '')
             df['內容'] = df['內容'].str.replace('\n', '')
             # 詩名我們今天沒用到，先 drop 掉
             df = df.drop(["詩名"], axis = 1)
             df['作者'] = df['作者'].astype('category')
             return df
```

1.4.4 建立詞向量

我們可以選用 CountVectorizer 或者 TfidfVectorizer 建立你的詞向量
那建議使用 sklearn 內建的轉換來做，因為這樣她會自動地幫你把詞向量特徵建起來起來，不需要自己建立

```
In [18]: import jieba

         df = process_poems(df)
         df
```

Out[18]:

	作者	內容
0	李白	平林 漠漠 煙如織 ，寒山 一帶 傷心 碧 。 暝 色入 高樓 ，有人 樓上 愁 。 ...
1	李白	青天 有 月 來 幾時 ，我今 停杯 一問 之 ： 人攀 明月 不可 得 ，月行 卻 與...
2	李白	燕草 如碧絲 ，秦桑低 綠枝 。當君 懷歸日 ，是 妾 斷腸時 。春風 不 相識 ， ...
3	李白	誰 家玉笛 暗飛聲 ，散 入春 風滿 洛城 。此 夜曲 中聞折 柳 ，何人 不起 故園情
4	李白	西上 蓮 花山 ，迢迢 見 明星 。 （ 西 上 一作 ： 西 嶽 ） 素 手把 芙蓉...
5	李白	明月 出 天山 ，蒼茫 雲海間 。長 風 幾萬裡 ，吹度 玉門關 。漢下 白登道 ， ...
6	李白	君不見 黃 河之水 天上 來 ，奔流 到海 不 複 回 。君不見 高堂 明鏡 悲白發 ， .
...	...	...
2724	白居易	秋花 紫 蒙蒙 ，秋蝶 黃 茸茸 。花低 蝶 新 小 ，飛戲 叢西東 。日暮 涼風來...
2725	白居易	三年 為 刺史 ，無政 在 人口 。唯向 郡 城中 ，題 詩十餘首 。慚非 甘棠 詠...
2726	白居易	零落 桐葉雨 ，蕭條 槿 花風 。悠悠 早 秋意 ，生此 幽閒 中 。況 與 故人 ...
2727	白居易	離離 暑雲散 ，嫋嫋 涼風起 。池上 秋 又 來 ，荷花 半成 子 。朱顏易銷 歇 ...
2728	白居易	涼風 冷露 蕭索天 ，黃 蒿 紫菊荒 涼田 。繞 塚 秋花少 顏色 ，細蟲 小蝶飛 翻..
2729	白居易	往歲 曾 為 西邑吏 ，慣 從 駱口 到 南秦 。 三時 雲冷多 飛雪 ，二月 山寒 ..
2730	白居易	笛 竹出 蘄 春 ，霜刀 劈翠筠 。織成 雙鎖 簟 ，寄 與 獨眠 人 。卷作筒 中...

[2731 rows x 2 columns]

In [5]:
```
df['作者'] = df['作者'].cat.codes
df
```

Out[5]:

	作者	內容
0	0	平林 漠漠 煙如織 ，寒山 一帶 傷心 碧 。 暝 色入 高樓 ，有人 樓上 愁 。 ...
1	0	青天 有 月 來 幾時 ，我今 停杯 一問 之 ： 人攀 明月 不可 得 ，月行 卻 與...
2	0	燕草 如碧絲 ，秦桑低 綠枝 。當君 懷歸日 ，是 妾 斷腸時 。春風 不 相識 ， ...
3	0	誰 家玉笛 暗飛聲 ，散 入春 風滿 洛城 。此 夜曲 中聞折 柳 ，何人 不起 故園情 。
4	0	西上 蓮 花山 ，迢迢 見 明星 。 （ 西 上 一作 ： 西 嶽 ） 素 手把 芙蓉...
5	0	明月 出 天山 ，蒼茫 雲海間 。長 風 幾萬裡 ，吹度 玉門關 。漢下 白登道 ， ...
6	0	君不見 黃 河之水 天上 來 ，奔流 到海 不 複 回 。君不見 高堂 明鏡 悲白發 ， ...
...	..	...
2724	2	秋花 紫 蒙蒙 ，秋蝶 黃 茸茸 。花低 蝶 新 小 ，飛戲 叢西東 。日暮 涼風來...
2725	2	三年 為 刺史 ，無政 在 人口 。唯向 郡 城中 ，題 詩十餘首 。慚非 甘棠 詠...
2726	2	零落 桐葉雨 ，蕭條 槿 花風 。悠悠 早 秋意 ，生此 幽閒 中 。況 與 故人 ...
2727	2	離離 暑雲散 ，嫋嫋 涼風起 。池上 秋 又 來 ，荷花 半成 子 。朱顏易銷 歇 ...
2728	2	涼風 冷露 蕭索天 ，黃 蒿 紫菊荒 涼田 。繞 塚 秋花少 顏色 ，細蟲 小蝶飛 翻...
2729	2	往歲 曾 為 西邑吏 ，慣 從 駱口 到 南秦 。 三時 雲冷多 飛雪 ，二月 山寒 ...
2730	2	笛 竹出 蘄 春 ，霜刀 劈翠筠 。織成 雙鎖 簟 ，寄 與 獨眠 人 。卷作筒 中...

```
[2731 rows x 2 columns]
```

```
In [6]: test_df = pd.read_csv("poem_test.csv")
        test_df = process_poems(test_df)
        test_df
```

Out[6]:

	作者	內容
0	李白	日照 香爐生 紫煙 ， 遙看 瀑布 掛 前川 。飛流 直下 三千尺 ， 疑是 銀河 落九天 。
1	李白	朝辭 白帝 彩雲間 ， 千裡 江陵 一日 還 。兩岸 猿聲 啼 不住 ， 輕舟 已過 萬 ...
2	李白	李白 乘舟 將欲行 ， 忽聞 岸上 踏歌 聲 。桃花潭水 深 千尺 ， 不及 汪倫送 我情 。
3	李白	故人 西辭黃鶴樓 ， 煙花 三月 下揚州 。孤帆 遠影 碧空 儘 ， 唯見長 江天 際流 。
4	李白	危樓 高 百尺 ， 手可摘 星辰 。不敢 高聲語 ， 恐驚 天上 人 。
5	李白	床前 明月光 ， 疑是 地上 霜 。舉頭 望明月 ， 低頭思 故鄉 。
6	李白	天門 中斷 楚江 開 ， 碧水 東流 至此 回 。兩岸 青山 相對 出 ， 孤帆 一片 日...
..	...	...
23	白居易	雨 砌 長 寒蕪 ， 風庭 落秋果 。窗間 有 閒叟 ， 儘 日 看 書 坐 。書中見 ...
24	白居易	睡足 肢體 暢 ， 晨起 開 中堂 。初旭泛 簾幕 ， 微風 拂 衣裳 。二婢 扶 盥櫛...
25	白居易	履道 西門 有 弊居 ， 池塘 竹樹繞 君廬 。豪華肥壯 雖無分 ， 飽暖安閒 即 有 餘...
26	白居易	昨日 複 今辰 ， 悠悠 七十 春 。所經 多 故處 ， 卻 想 似 前身 。散 秩優遊...
27	白居易	不 與 老為期 ， 因何 兩鬢絲 ？才 應免 夭促 ， 便 已 及 衰 羸 。昨夜 夢 ...
28	白居易	暖床 斜 臥日 曛 腰 ， 一覺 閒眠 百病 銷 。儘 日 一 飧 茶 兩 碗 ， 更無所...
29	白居易	選石 鋪 新路 ， 安橋 壓古堤 。似 從 銀漢下 ， 落傍玉 川西 。影定 欄杆 倒 ...

```
[30 rows x 2 columns]
```

1.4.5 替換測試類別

這邊必須使用剛剛存起來的字典來替換
因為如果直接使用 code 可能會發生沒對照到的事故

```
In [7]: test_df['作者'] = test_df['作者'].replace(saved_map)
        test_df
```

Out[7]:

	作者	內容
0	0	日照 香爐生 紫煙 ， 遙看 瀑布 掛 前川 。飛流 直下 三千尺 ， 疑是 銀河 落九天 。
1	0	朝辭 白帝 彩雲間 ， 千裡 江陵 一日 還 。兩岸 猿聲 啼 不住 ， 輕舟 已過 萬 ...
2	0	李白 乘舟 將欲行 ， 忽聞 岸上 踏歌 聲 。桃花潭水 深 千尺 ， 不及 汪倫送 我情 。
3	0	故人 西辭黃鶴樓 ， 煙花 三月 下揚州 。孤帆 遠影 碧空 儘 ， 唯見長 江天 際流 。

```
4    0                危樓 高 百尺 ， 手可摘 星辰 。不敢 高聲語 ， 恐驚 天上 人 。
5    0                 床前 明月光 ， 疑是 地上 霜 。舉頭 望明月 ， 低頭思 故鄉 。
6    0  天門 中斷 楚江 開 ， 碧水 東流 至此 回 。兩岸 青山 相對 出 ， 孤帆 一片 日...
..  ..                                                  ...
23   2  雨 砌 長 寒蕪 ， 風庭 落秋果 。窗間 有 閒叟 ， 儘 日 看 書 坐 。書中見 ...
24   2  睡足 肢體 暢 ， 晨起 開 中堂 。初旭泛 簾幕 ， 微風 拂 衣裳 。二婢 扶 盥櫛...
25   2  履道 西門 有 弊居 ， 池塘 竹樹繞 君廬 。豪華肥壯 雖無分 ， 飽暖安閒 即 有 餘...
26   2  昨日 複 今辰 ， 悠悠 七十 春 。所經 多 故處 ， 卻 想 似 前身 。散 秩優遊...
27   2  不 與 老為期 ， 因何 兩鬢絲 ？才 應免 夭促 ， 便 已 及 衰 羸 。昨夜 夢 ...
28   2  暖床 斜 臥日 曛 腰 ， 一覺 閒眠 百病 銷 。儘 日 一 飧 茶 兩 碗 ， 更無所...
29   2  選石 鋪 新路 ， 安橋 壓古堤 。似 從 銀漢下 ， 落傍玉 川西 。影定 欄杆 倒 ...

[30 rows x 2 columns]
```

1.5 開始訓練

1.5.1 建立詞向量並預測

我們可以選用 CountVectorizer 或者 TfidfVectorizer 建立你的詞向量

那建議使用 sklearn 內建的轉換來做，因為這樣她會自動地幫你把詞向量特徵建起來起來，不需要自己建立

```
In [8]: from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.naive_bayes import MultinomialNB
        vec = TfidfVectorizer()
        bag = vec.fit_transform(df['內容'])
        print("維度:", len(vec.get_feature_names()))
        clf = MultinomialNB(alpha = 0.001)
        clf.fit(bag, df['作者'])
```

```
維度: 52294
```

```
Out[8]: MultinomialNB(alpha=0.001, class_prior=None, fit_prior=True)
```

```
In [9]: from sklearn.metrics import accuracy_score

        test_bag = vec.transform(test_df['內容'])
        predict = clf.predict(test_bag)
```

```
        print("預測:", list(predict))
        print("正確標籤:", list(test_df['作者']))
        print("Naive-Bayes 正確率: ", accuracy_score(test_df['作者'], predict) * 100, "%")
```

```
預測: [0, 0, 0, 0, 0, 2, 0, 0, 0, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 0, 2, 2, 2, 2, 2, 2, 2, 2, 1
正確標籤: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2,
Naive-Bayes 正確率:  80.0 %
```

```
In [10]: from sklearn.metrics import confusion_matrix
         cm = confusion_matrix(test_df['作者'], predict)
         pd.DataFrame(cm)
```

Out[10]:

	0	1	2
0	8	1	1
1	0	8	2
2	1	1	8

1.5.2 跟 kNN 比較一下

我們試試看 kNN 在這裡表現的如何

```
In [11]: from sklearn import neighbors

         clf = neighbors.KNeighborsClassifier(n_neighbors=8)
         clf = clf.fit(bag, df['作者'])
         predict = clf.predict(test_bag)

         print("預測:", list(predict))
         print("正確標籤:", list(test_df['作者']))
         print("kNN 正確率: ", accuracy_score(test_df['作者'], predict) * 100, "%")
```

```
預測: [0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 2, 0, 1, 1, 2, 2, 0, 0, 2, 0, 2
正確標籤: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2,
kNN 正確率:  53.333333333333336 %
```

1.6 結論

你最後發現，對於文字這種資料，還是單純貝氏來的比較合適

那為什麼我們的訓練結果跟上一次我們使用新聞的結果差距比較大呢?

因為我們的詩詞是一首比較短的向量，所以訓練資料相對的更少了，而在訓練資料少的時候我們更可以看到機率算法的好處 (對於資料的多少稍微不敏感一點)

1.6.1 優點

1. 啟動資料少的時候，單純貝氏還是可以達到一個不錯的結果
2. 對於文字這種稀疏的特徵，是一個非常好的分類器

1.6.2 缺點

1. 解釋性稍微低了一點
2. 用在普通非稀疏的模型，效果甚至可能不如 kNN 或者決策樹類型的演算法來得好