資料概述

全部資料有 10000 筆的評論，包含了 training data 的 7997 筆和 test data 的 2003 筆，每筆資料都含有以下資訊。

1. review_id (ID for the posted review)
2. business_id (ID of the business being reviewed)
3. user_id (User's id)
4. text (Review text)
5. date (Day the review was posted)
6. stars (1–5 rating for the business)

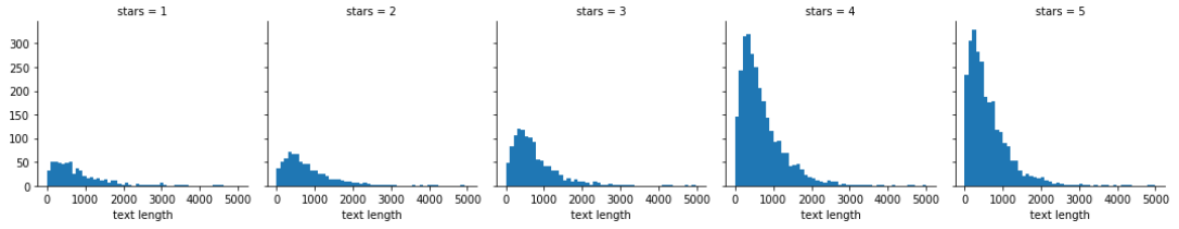
檢視 training data 的型態：

	review_id	business_id	user_id	text	date	stars
0	3223	2055	2533	Sometimes things happen, and when they do this...	2010-12-30	5
1	9938	4165	6371	I know Kerrie through my networking and we ben...	2011-04-26	5
2	7123	869	4929	Love their pizza!!!\nVery fresh. Their cannoli...	2012-09-28	5
3	3601	1603	2789	Being from NJ I am always on the prowl for my ...	2009-06-07	4
4	3948	2347	1245	We have tried this spot a few times and each v...	2011-02-20	4

為了更加了解 text 故我們多加入了一行 text length 儲存每個評論的字數，如下圖所示：

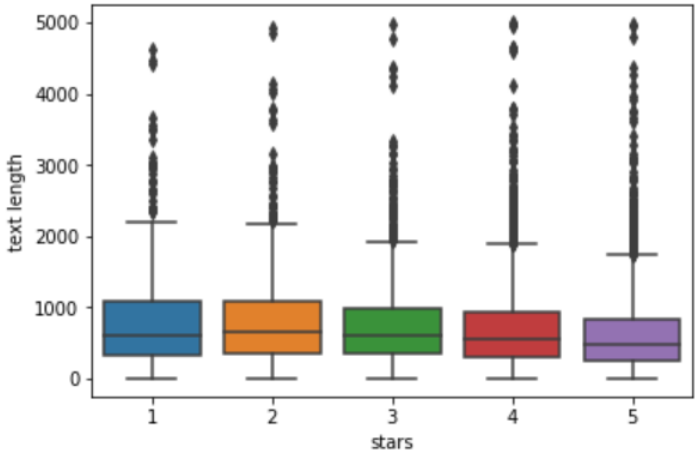
	review_id	business_id	user_id	text	date	stars	text length
0	3223	2055	2533	Sometimes things happen, and when they do this...	2010-12-30	5	211
1	9938	4165	6371	I know Kerrie through my networking and we ben...	2011-04-26	5	426
2	7123	869	4929	Love their pizza!!!\nVery fresh. Their cannoli...	2012-09-28	5	110
3	3601	1603	2789	Being from NJ I am always on the prowl for my ...	2009-06-07	4	442
4	3948	2347	1245	We have tried this spot a few times and each v...	2011-02-20	4	332

接著我們透過圖來進一步了解資料資訊，首先利用直方圖來觀察 text length 分別在五個 stars 裡的分布狀況，如下圖所示:



可以觀察 text length 的分布情形在五個星等中大致類似，但是明顯地可以看出在四星和五星中的評論數量偏多。

接著我們畫出 text length 和星等的箱型圖，如下所示:



.

可以看出一星和二星的 text length 較其他三個星等略長，不過也有許多 outliers，所以觀察完圖形後可以發現 text length 對於預測星等來說可能不是一個很好的特徵。

資料預處理

因為一些字詞對於分析上無實質幫助，故我們必須先將之做文章清理的動作，例如停頓詞（Stop Words）和標點符號，結果如下所示:

範例:

```
Sometimes things happen, and when they do this is the place where y
ou want them taken care of. The orthopedics department and trauma d
epartment have some of the best, and nicest Doctors you'll find in
the state.
```

清理後結果:

```
'Sometimes', 'things', 'happen', 'place', 'want', 'taken', 'care',
 'orthopedics', 'department', 'trauma', 'department', 'best', 'nic
est', 'Doctors', 'youll', 'find', 'state'
```

可以看出清理後的結果都只剩下重要的單詞，冗言贅字與標點符號皆以刪除，如此一來便有利後續分析。

向量化

我們先將 train data 所有的 text 做文章清理後，將所有 unique 的單字丟進詞袋(Bag-of-words)裡，發現總共有 39002 個單詞，接著以 train data 的第 25 則評論為例，可以轉換成以下結果:

Potbelly, you have been handed the torch. It was a good ride Jimmy Johns, but you have no locations close to my newly purchased home. Its not you, its me. I promise. I love italian-style subs. The Wreck at Potbelly is fantastic. The Wreck salad is also great (for a salad). This place is my new go to sub shack.

```
(0, 6713)     1
(0, 6741)     1
(0, 9587)     2
(0, 13049)    2
(0, 13851)    1
(0, 17398)    1
(0, 21296)    1
(0, 22813)    1
(0, 22869)    1
(0, 23085)    1
(0, 23463)    1
(0, 24019)    1
(0, 25209)    1
(0, 26294)    1
(0, 26424)    1
(0, 28050)    1
(0, 28061)    1
(0, 29985)    1
(0, 30770)    1
(0, 30967)    1
(0, 32127)    1
(0, 32495)    2
(0, 33240)    1
(0, 35082)    1
(0, 35100)    1
(0, 36508)    1
```

左邊的圖代表第 25 則評論裡有 26 個唯一的字元，有三個單字出現了兩次，其餘皆只出現一次。

分別是 Potbelly
Wreck
salad

接著我們就可以將全部的 train data 轉成類似下圖的形式:

	Review 1	Review 2	...	Review N
Word 1 Count	0	1	...	0
Word 2 Count	0	0	...	0
...	1	2	...	0
Word N Count	0	1	...	1

Multinomial Naive Bayes

因為上一步在做向量化時，我們將資料去除停頓詞（Stop Words）和標點符號後轉換成詞頻的形式，故在這邊我們選擇使用 sklearn 裡的 Multinomial Naive Bayes 模型去做分類，其中我們將資料再分割成 80%的訓練集和 20%的測試集，接著我們用測試集測試 MODEL 做出混淆矩陣和分類報告，如下圖所示。

```
[[ 22  11  11  51  20]
 [  4  11  15  94  16]
 [  0   5  15 205  21]
 [  3   3  11 412 129]
 [  2   2   5 236 296]]
```

	precision	recall	f1-score	support
1	0.71	0.19	0.30	115
2	0.34	0.08	0.13	140
3	0.26	0.06	0.10	246
4	0.41	0.74	0.53	558
5	0.61	0.55	0.58	541
avg / total	0.47	0.47	0.43	1600

從混淆矩陣可以看出除了第一星等外，我們的模型很容易將資料分到上下的兩個星等，例如第四星等便有 20%分到第三星等 23%分到第五星等，所以從分類報告可看出大約只有 47%的正確率，我們猜測可能是資料量太少，或是使用者的評論雖然是好的，但給的星數不一定如同評論預期的那麼高所致。

已註解 [M帳1]:

接著我們用 train data 實際測試一下我們的模型，先使用五星的正評當例子，如下圖所示:

```
star: 5


Huge fan of Bikram yoga! It's a 90 minute, torturous-at-times, sweat-filled  workout but when you are done, you feel insanely a
mazing. The feeling continues throughout the day leaving you in complete zen....Road rage? Gone. Annoying food cravings? Zero.
Jeans? Fitting awesome. You really must try it, it can completely change your body.

There are a few places in the valley that offer Bikram, most of them exclusively rather than in combo with other methods of yog
a. The Bikram Institute on Miller and Indian School is where I first tried it, and I like it the best so far. They offer classe
s throughout the day and later in the evening as well. Like other studios, this place offers first timers a week-long pass for
around $20. What I really like about this studio is their instructors and the fact that they keep their room darkened throughou
t the entire practice. I've have been going to another studio lately (for their first-time student deals) and their instructors
bark out instructions non-stop plus they keep the lights on the whole time. So when you're staring up lying in your relaxed sta
te, your eyes meet the glare of the fluorescent.

I'll very likely return here after I've exhausted the new student deals at the studios around the valley. Yoga can be expensiv
e!


predict star: 5
```

可看出我們的模型預測正確，那麼接下來我們來測試看看 1 星的評論

```
star: 1

$8 for a tiny sandwich with one egg, two greasy strips of bacon, and a stale muffin. They say it's pricy because they get it lo
cal? It has to have substance to be considered anything. So whack.

predict star: 4
```

可看出我們的模型將一星的分類成了四星，至於為什麼會如此，我們猜測可能是資料多偏向四星和五星，故訓練出來的模型也會更傾向將資料分類至四星及五星。

Github

https://github.com/shakewang11/CP2.git

Reference

1.https://medium.com/tensorist/classifying-yelp-reviews-using-nltk-and-scikit-learn-c58e71e962d9