第十屆巨量資料分析就業養成班

BDSE10

專題發表展示會

專題報告時程表

第十屆巨量資料分析
就業養成班
BDSE10

Yelp 餐廳美食推薦	09 : 30 – 09 : 55
日本旅遊行程推薦及路線規劃	10 : 00 – 10 : 25
摸摸紐思 More² News	10 : 30 – 10 : 55
中場休息	11 : 00 – 11 : 10
台灣美食論壇文字探勘	11 : 10 – 11 : 35
電商平台熱銷新品預測推薦引擎	11 : 40 – 12 : 05

Yelp 餐廳美食推薦

Group 1 趙上涵 林庭宇 薛正暐 吳岱凌

AGENDA

01 團隊介紹

02 核心能力

03 專案價值

04 專案流程

05 結論

01 團隊介紹

趙上涵

資料清理
叢集架設
深度/機器學習

薛正暐

資料清理
叢集架設
深度/機器學習

林庭宇

資料清理
文本分析
深度/機器學習

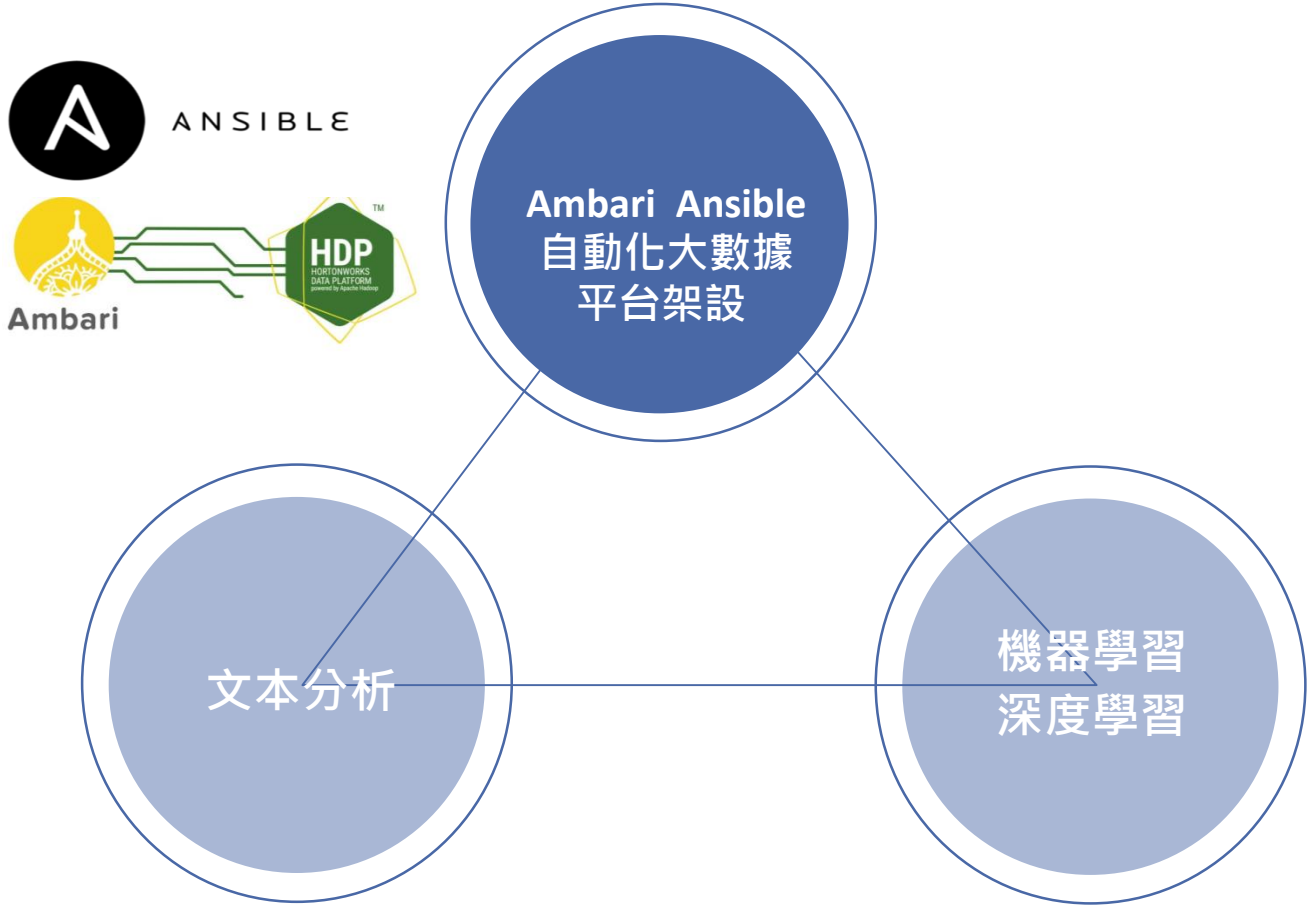
吳岱凌

資料清理
文本分析
Collaborative filtering
簡報製作

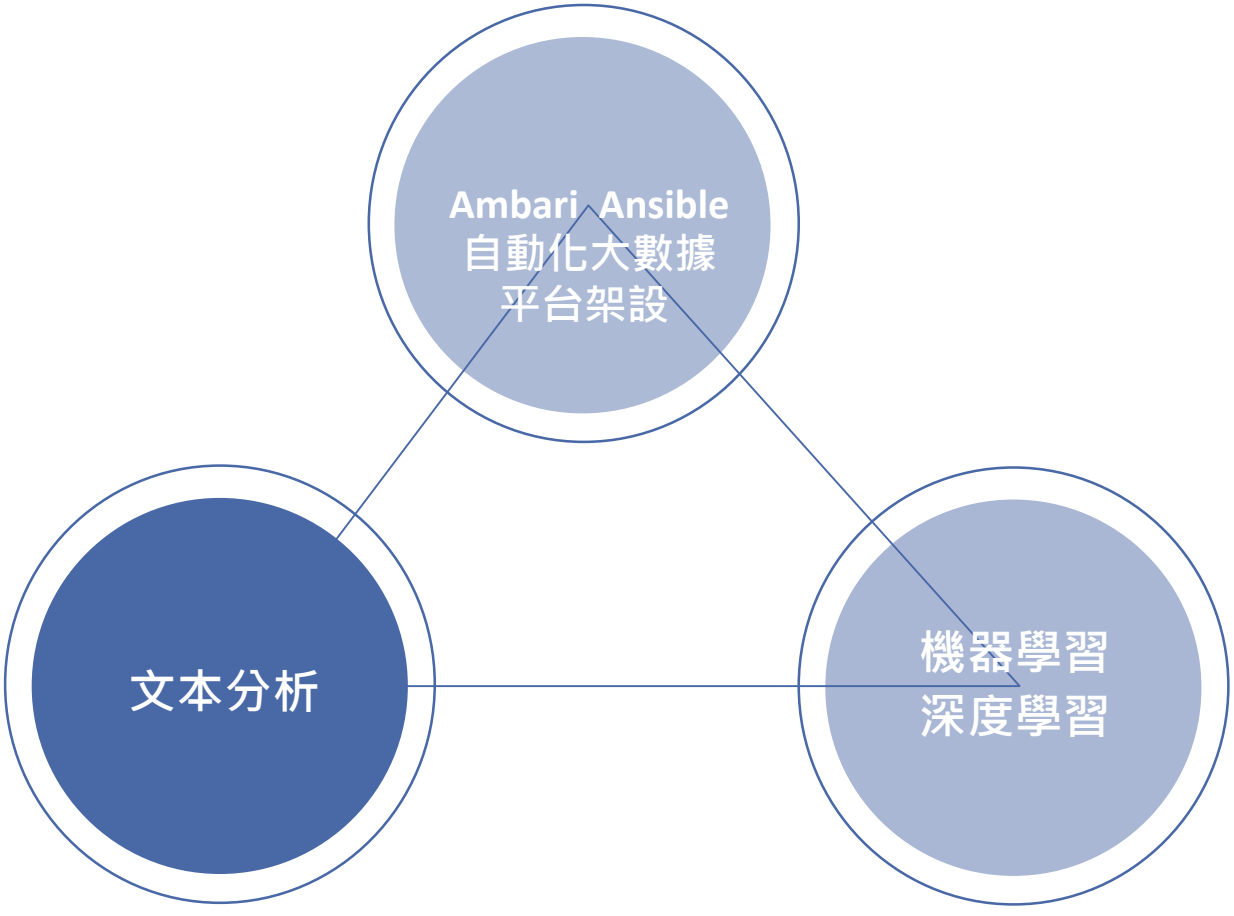
02

核心能力

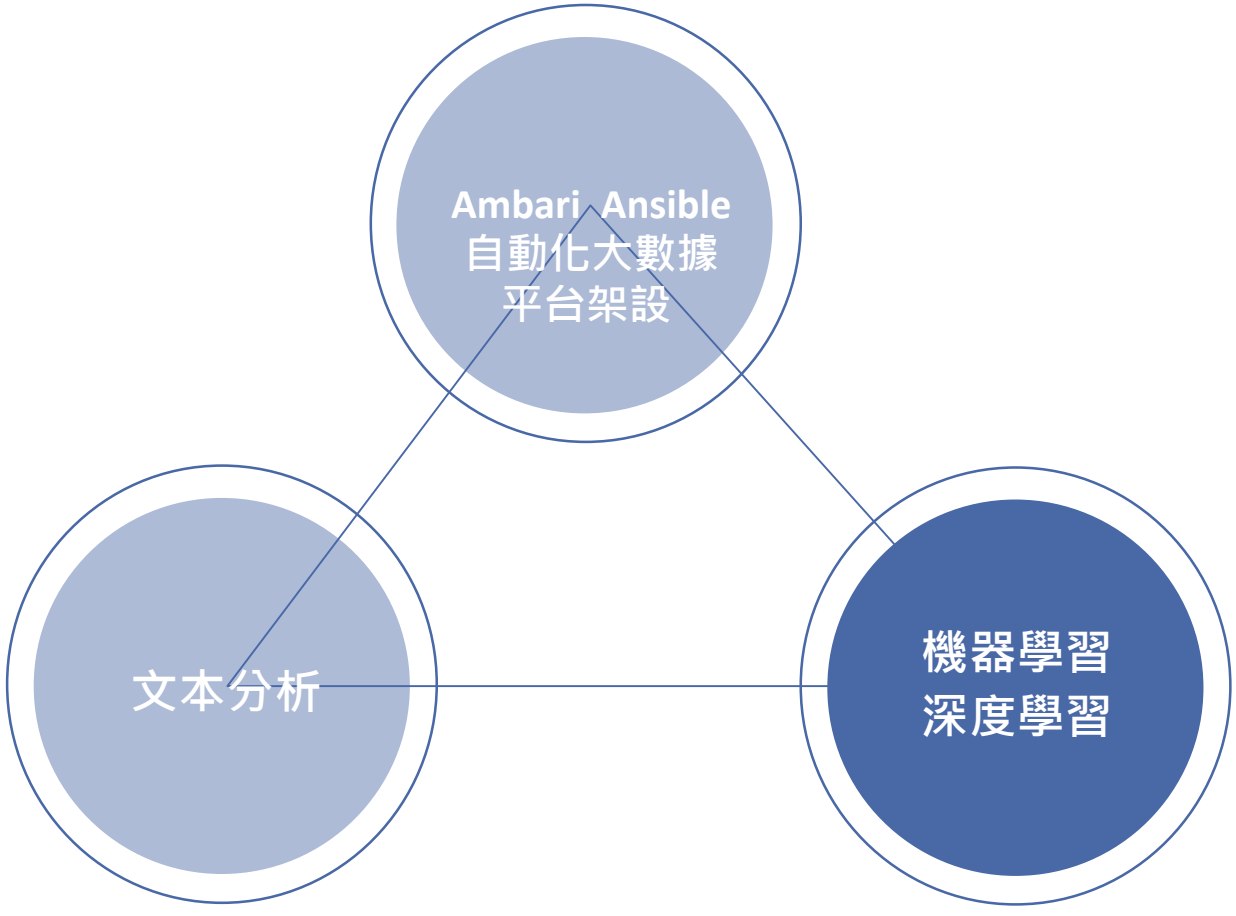
趙上涵


ANSIBLE
Ambari
HDP
HORTONWORKS DATA PLATFORM
Ambari Ansible
自動化大數據
平台架設
文本分析
機器學習
深度學習

趙上涵



趙上涵



dmlc XGBoost

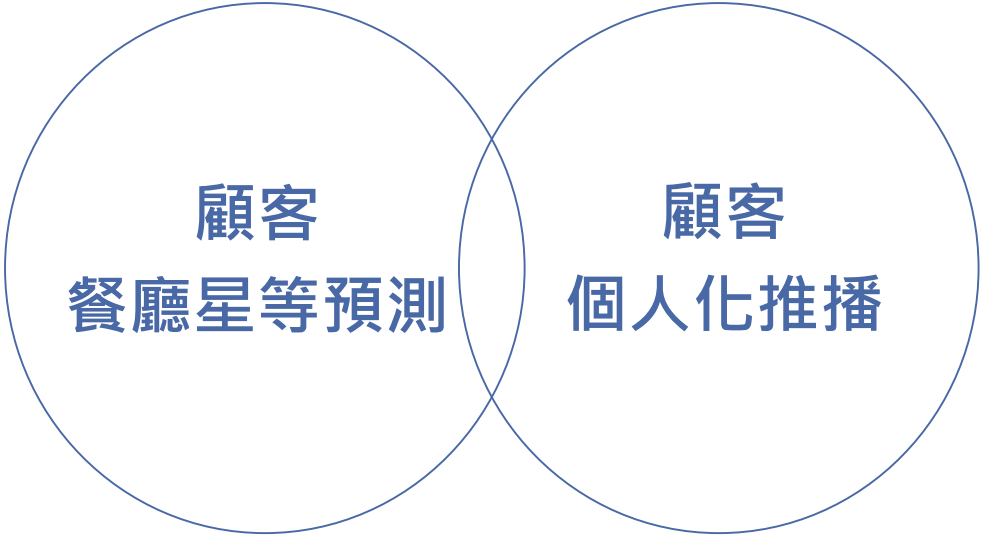
LUDWIG

[:::]

03

專案價值

趙上涵



趙上涵

趙上涵

趙上涵

五特質分數
打卡時間
周間分布
餐廳類別

+

使用者特質
過去評論特質
評論次數
平均給予星等

趙上涵

專案目標

進行用戶評分星等預測

機器學習

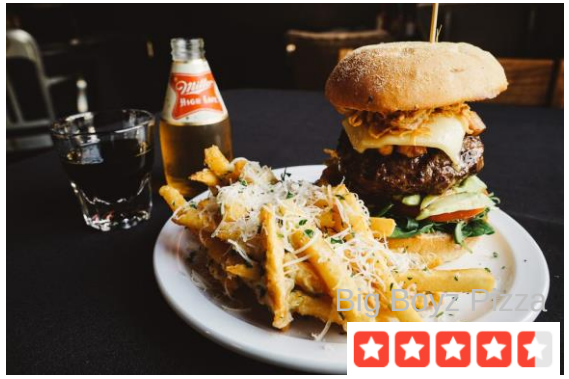
深度學習

Collaborative Filtering

yelp

Breakfast & Brunch | Pizza | Seafood

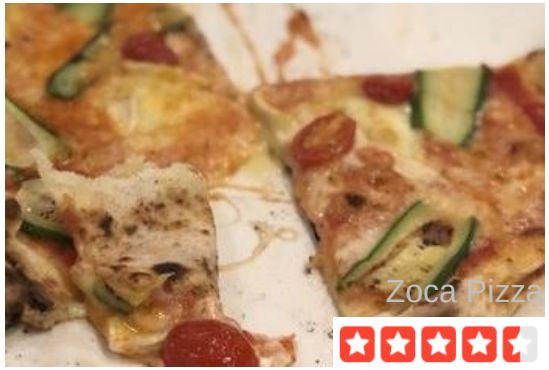
Coffee & Tea | Burgers | Bars



yelp

Breakfast & Brunch | **Pizza** | Seafood

Coffee & Tea | Burgers | Bars



04

專案流程

04 專案流程

非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析

非關聯式資料庫

資料預處理

匯入關聯式資料庫

大數據平台架設

資料分析

mongoDB®

非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析

非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析

Microsoft® SQL Server®

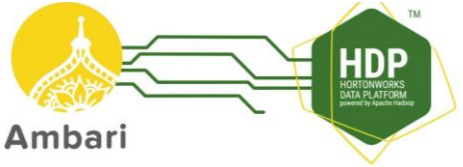
非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析



04 專案流程

非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析

非關聯式資料庫

資料預處理

匯入關聯式資料庫

大數據平台架設

資料分析

kaggle yelp

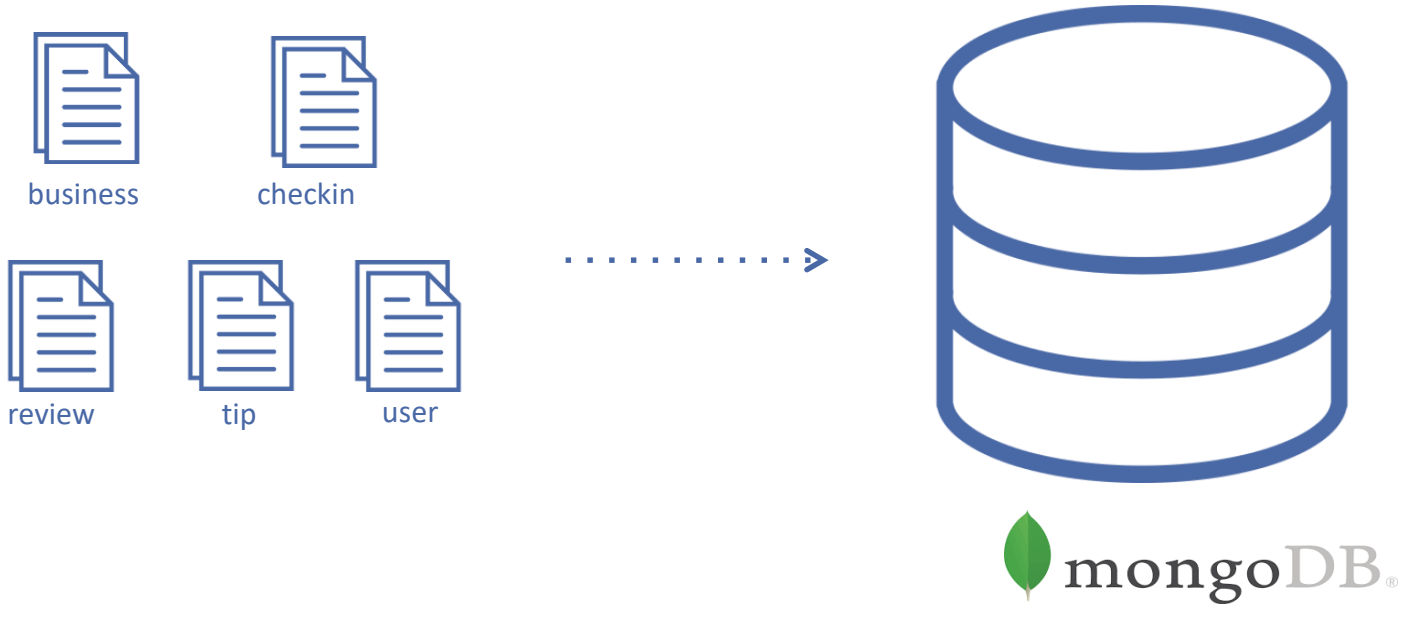
business checkin

review tip user

資料筆數 10,474,809

檔案大小 8.6G

檔案格式 JSON


business
checkin
review
tip
user
mongoDB.

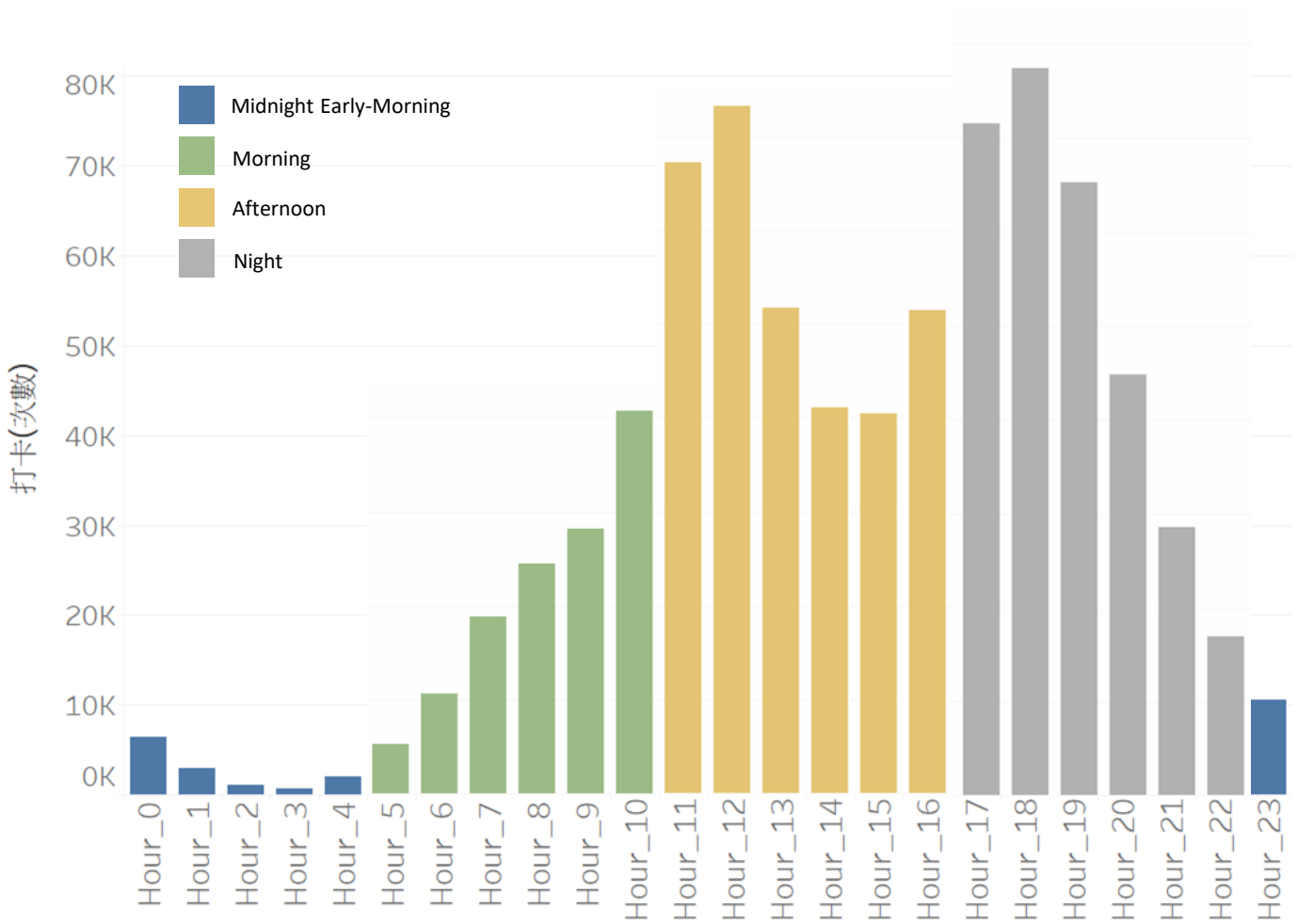
非關聯式
資料庫

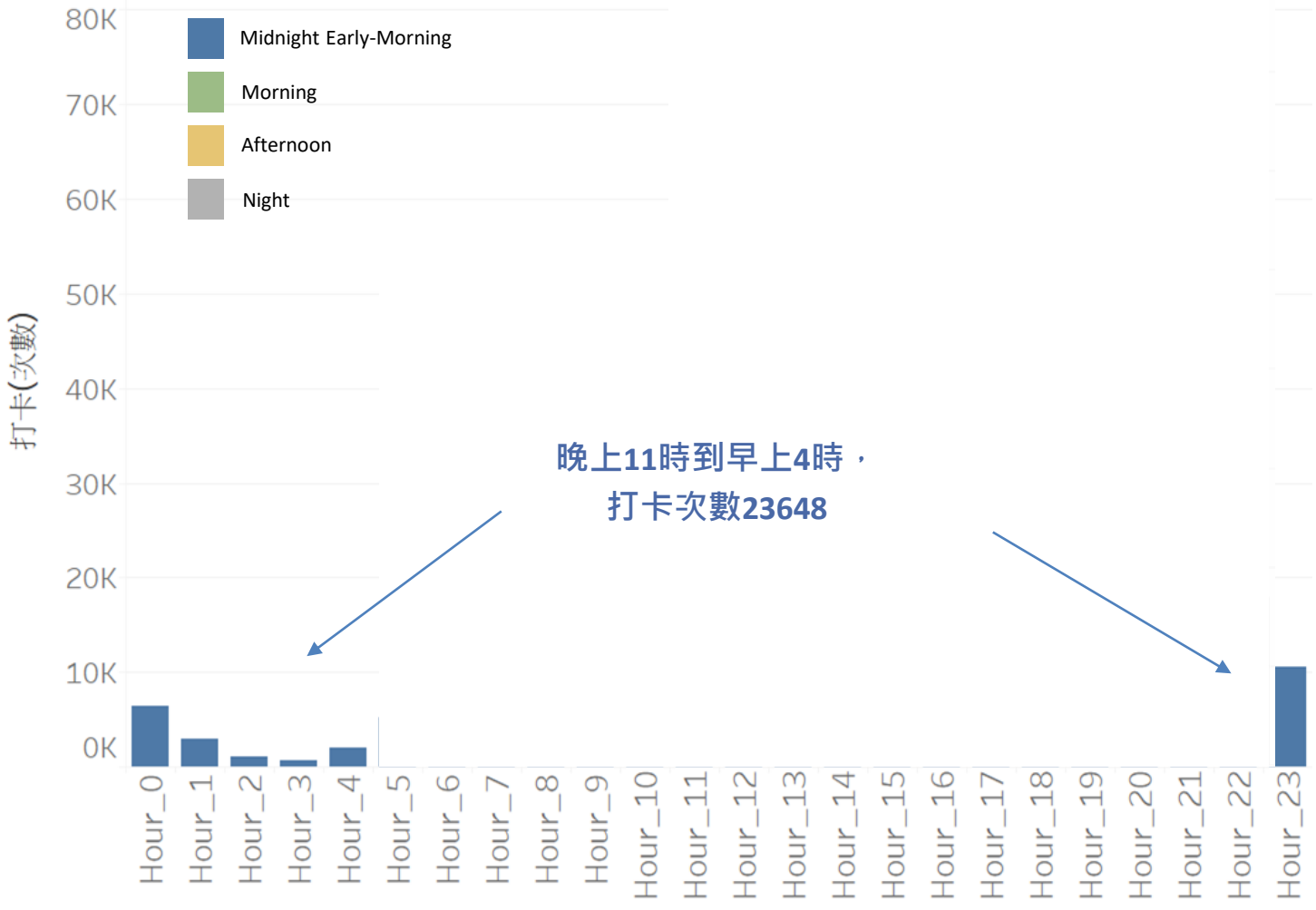
資料預處理

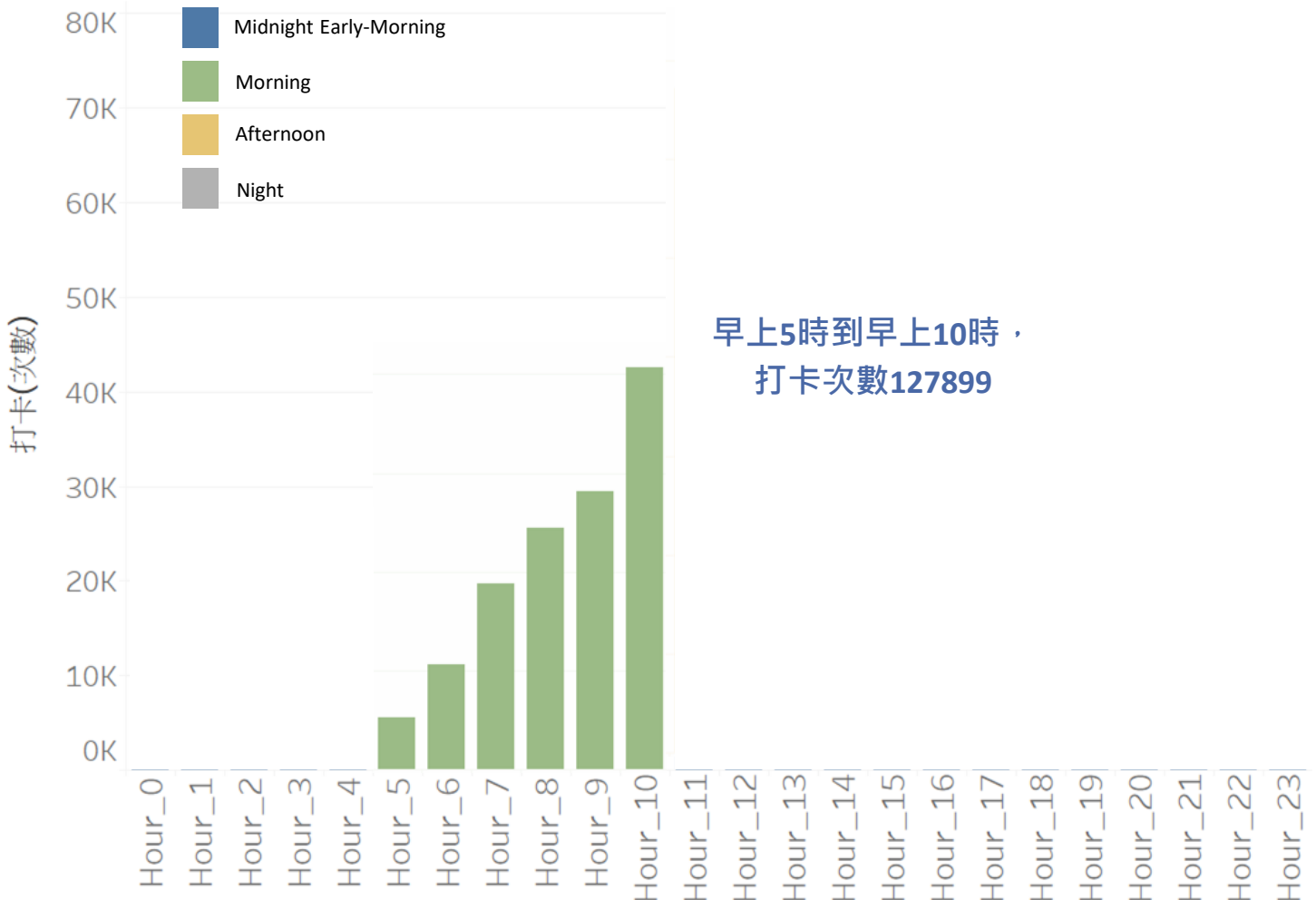
匯入
關聯式資料庫

大數據
平台架設

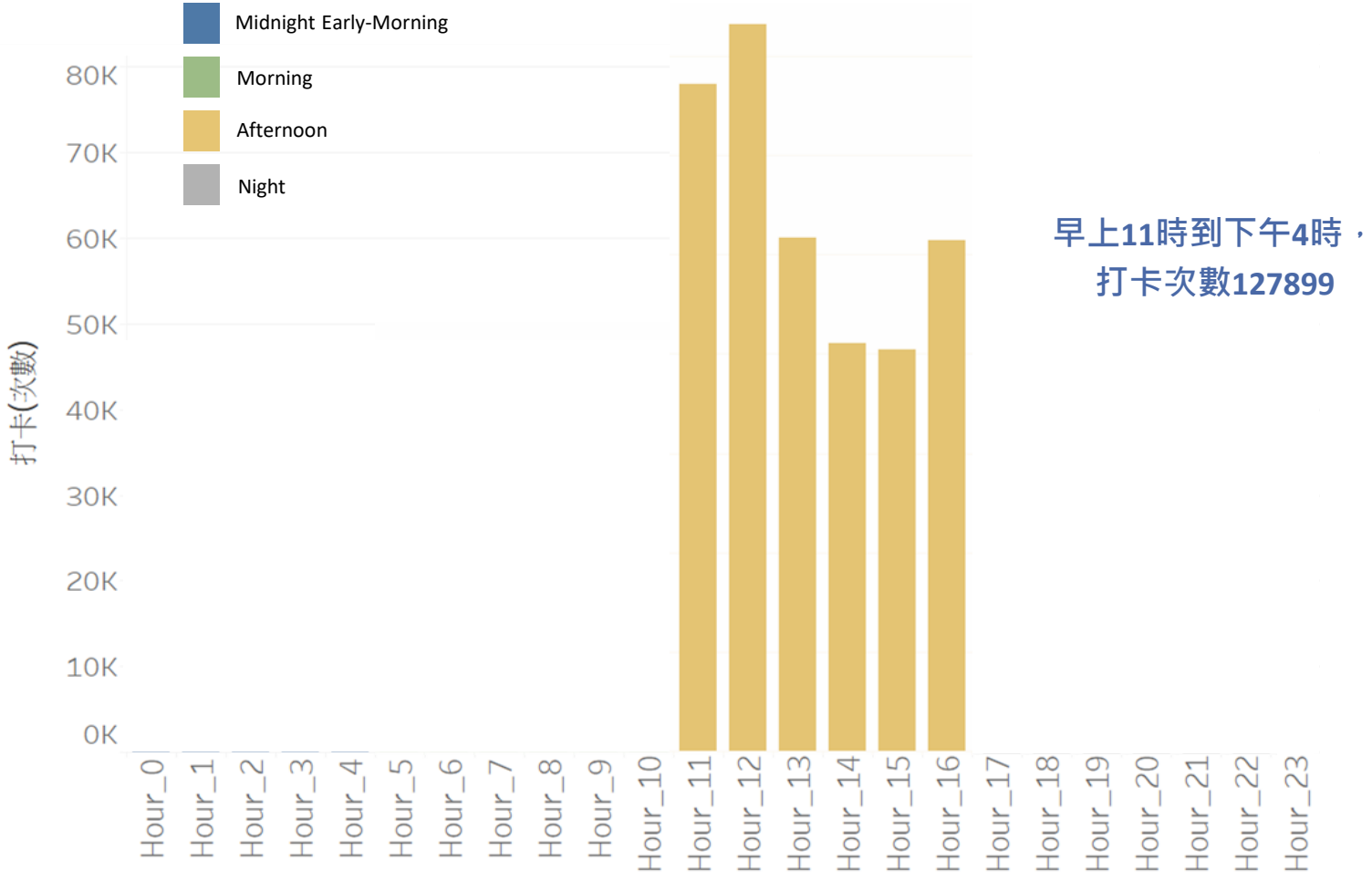
資料分析


Midnight Early-Morning
Morning
Afternoon
Night
打卡(次數)
80K
70K
60K
50K
40K
30K
20K
10K
0K
Hour_0
Hour_1
Hour_2
Hour_3
Hour_4
Hour_5
Hour_6
Hour_7
Hour_8
Hour_9
Hour_10
Hour_11
Hour_12
Hour_13
Hour_14
Hour_15
Hour_16
Hour_17
Hour_18
Hour_19
Hour_20
Hour_21
Hour_22
Hour_23

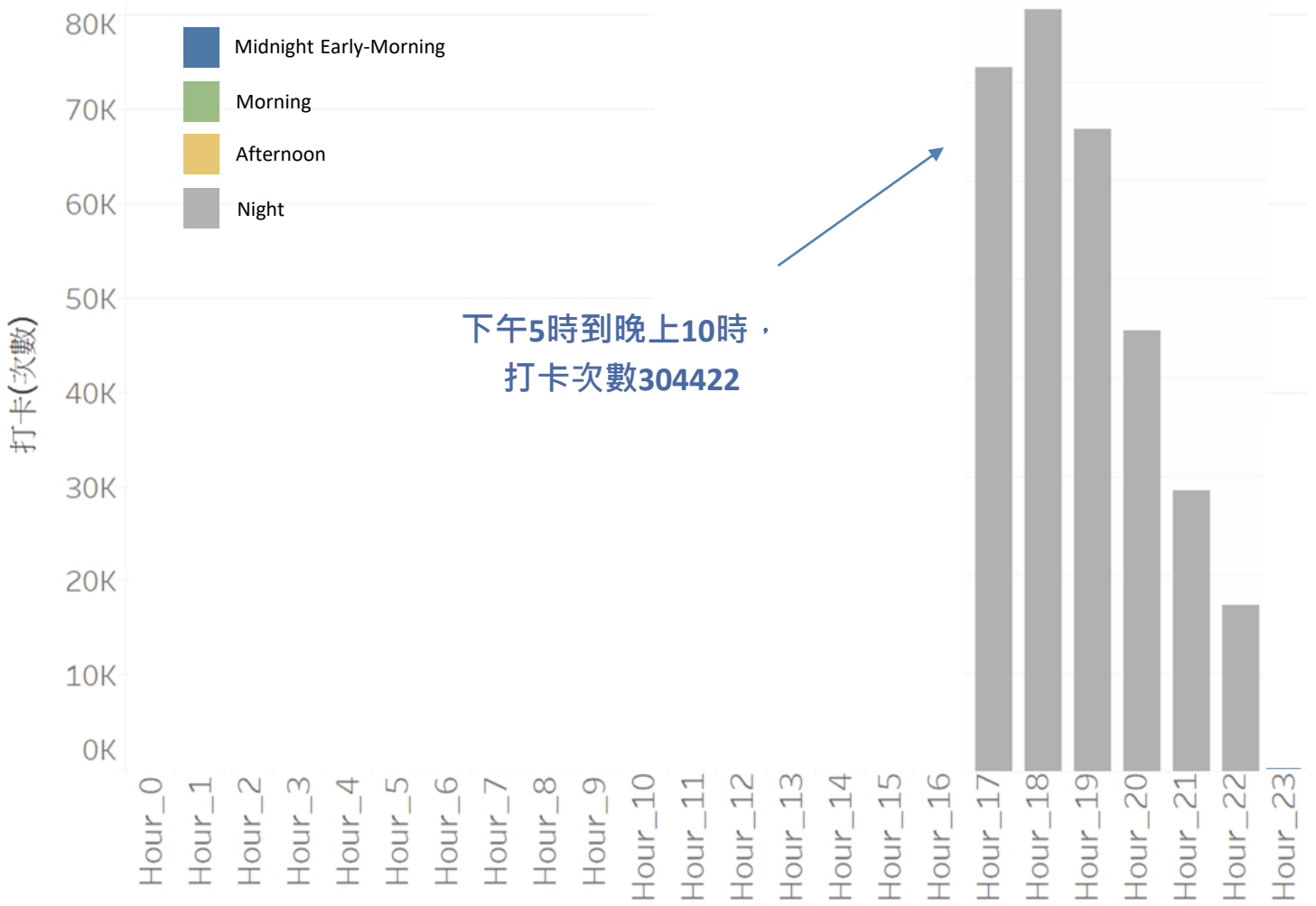

Midnight Early-Morning
Morning
Afternoon
Night
打卡(次數)
80K
70K
60K
50K
40K
30K
20K
10K
0K
晚上11時到早上4時，
打卡次數23648
Hour_0
Hour_1
Hour_2
Hour_3
Hour_4
Hour_5
Hour_6
Hour_7
Hour_8
Hour_9
Hour_10
Hour_11
Hour_12
Hour_13
Hour_14
Hour_15
Hour_16
Hour_17
Hour_18
Hour_19
Hour_20
Hour_21
Hour_22
Hour_23



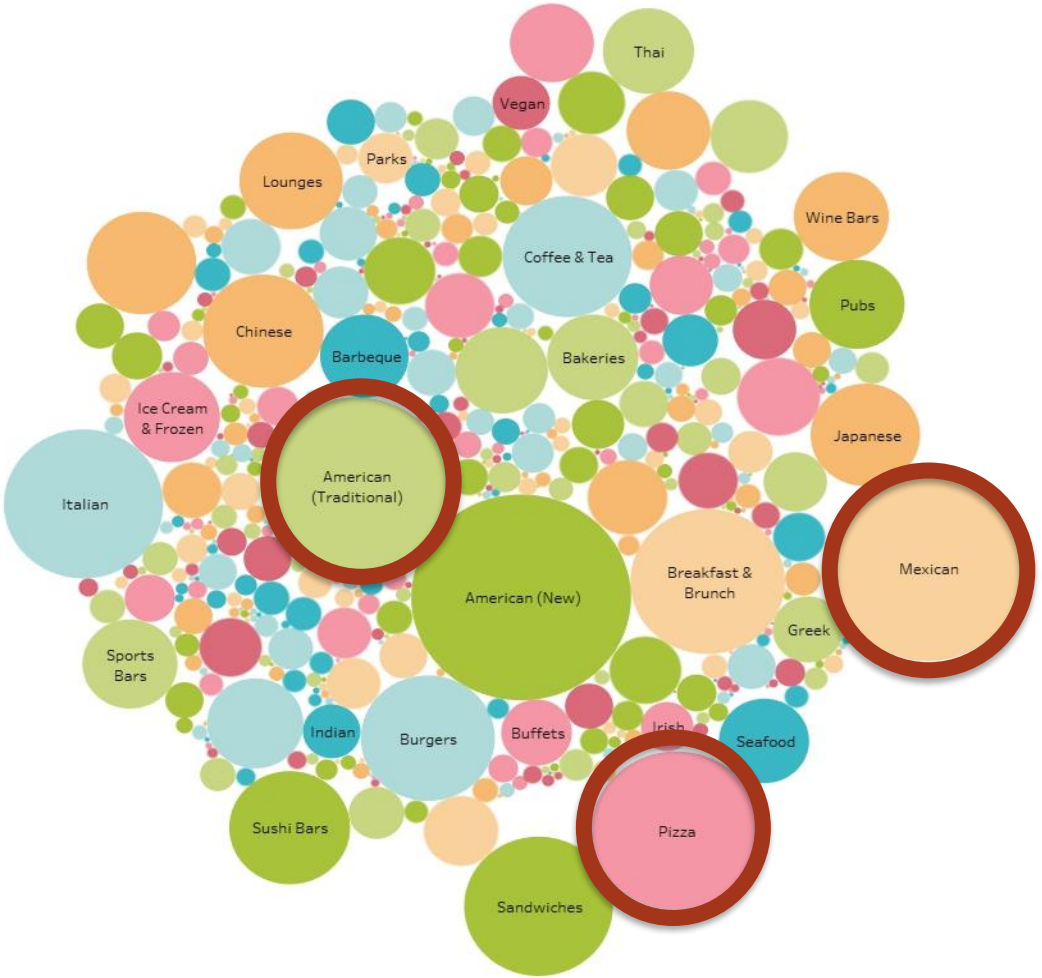
早上5時到早上10時，
打卡次數127899



早上11時到下午4時，
打卡次數127899

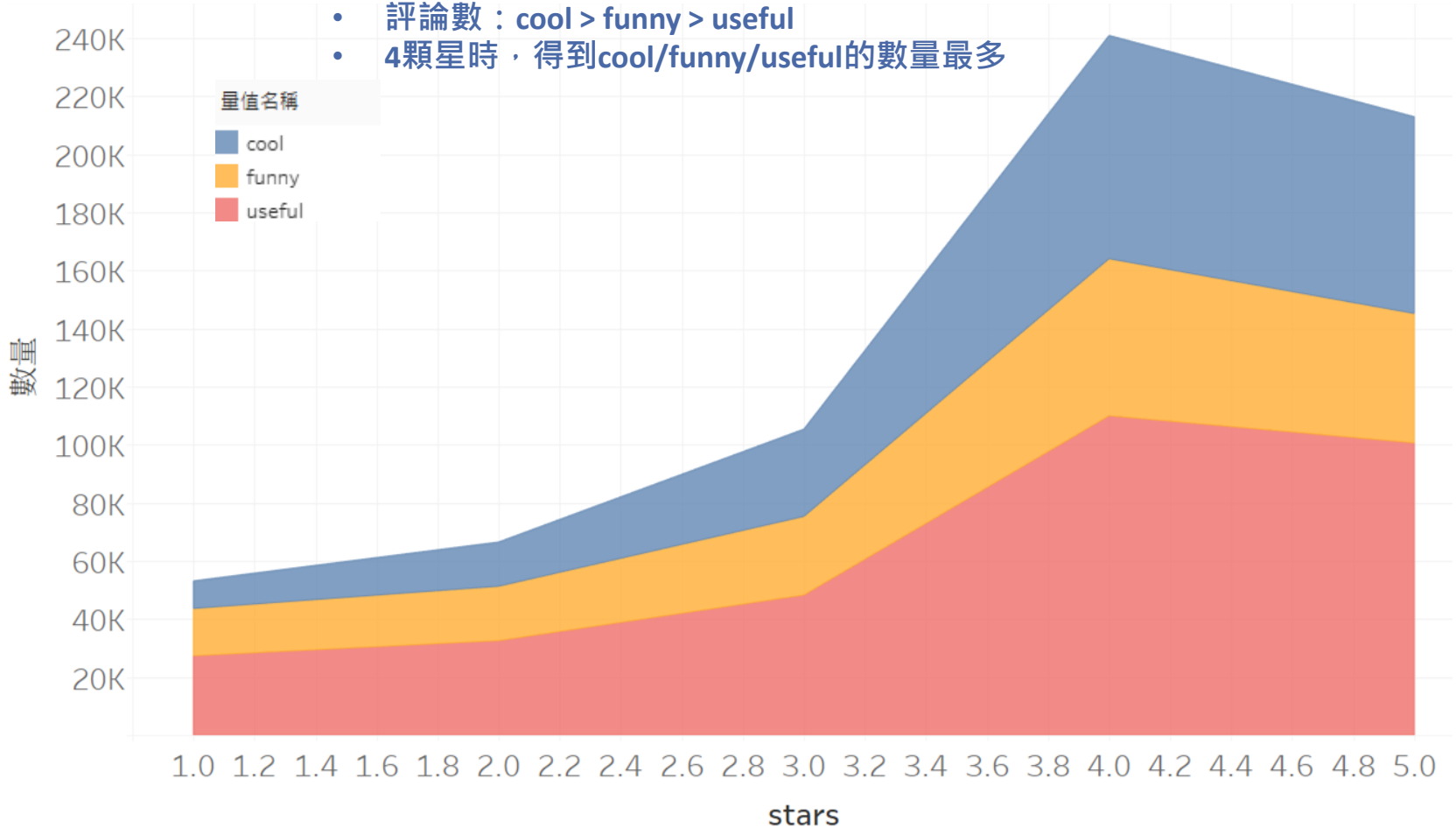

80K
70K
60K
50K
40K
30K
20K
10K
0K
打卡(次數)
Midnight Early-Morning
Morning
Afternoon
Night
下午5時到晚上10時，
打卡次數304422
Hour_0
Hour_1
Hour_2
Hour_3
Hour_4
Hour_5
Hour_6
Hour_7
Hour_8
Hour_9
Hour_10
Hour_11
Hour_12
Hour_13
Hour_14
Hour_15
Hour_16
Hour_17
Hour_18
Hour_19
Hour_20
Hour_21
Hour_22
Hour_23


Thai
Vegan
Parks
Lounges
Wine Bars
Coffee & Tea
Pubs
Chinese
Barbeque
Bakeries
Ice Cream & Frozen
Japanese
American (Traditional)
Italian
Mexican
American (New)
Breakfast & Brunch
Greek
Sports Bars
Indian
Burgers
Buffets
Irish
Seafood
Sushi Bars
Pizza
Sandwiches

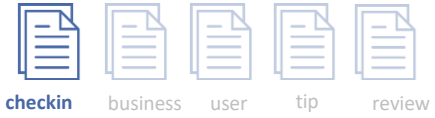


美式、墨西哥式、披薩餐廳
為美國民眾喜愛餐廳大宗

- 評論數：cool > funny > useful
- 4顆星時，得到cool/funny/useful的數量最多



資料清理
文本分析




```
check_in.head()
```

	business_id	checkin_info	type
0	KO9CpaSPOoqm0iCWm5scmg	{'11-3': 17, '8-5': 1, '15-0': 2, '15-3': 2, '...	checkin
1	oRqBAYtcBYZHXA7G8FlPaA	{'0-5': 1, '2-6': 2, '2-5': 3, '3-6': 1, '3-5'...	checkin
2	6cy2C9aBXUwkrh4bY1DApw	{'13-4': 1, '7-4': 1, '15-3': 1, '18-5': 1, '2...	checkin
3	D0IB17N66FiyYDCzTIAI4A	{'13-5': 1, '17-6': 1, '15-1': 1, '20-0': 1, '...	checkin
4	HLQGo3EaYVvAv22bONGklw	{'16-2': 1, '14-5': 1, '12-5': 2, '15-4': 1, '...	checkin

資料清理
文本分析

checkin business user tip review

將時間訊息分為時刻與星期幾

checkin_info

下午四點 星期二

16 - 2

時刻 星期幾

	business_id	checkin_info	type
0	KO9CpaSPOoqm0iCWm5scmg	{'11-3': 17, '8-5': 1, '15-0': 2, '15-3': 2, '...	checkin
1	oRqBAYtcBY[illegible]GFlPaA	{'0-5': 1, '2-6': 2, '2-5': 3, '3-6': 1, '3-5'...	checkin
2	6cy2C9aBXI[illegible]4bY1DApw	{'13-4': 1, '7-4': 1, '15-3': 1, '18-5': 1, '2...	checkin
3	D0IB17N66FiyYDCzTIAI4A	{'13-5': 1, '17-6': 1, '15-1': 1, '20-0': 1, '...	checkin
4	HLQGo3EaYVvAv22bONGkIw	{'16-2': 1, '14-5': 1, '12-5': 2, '15-4': 1, '...	checkin

資料清理
文本分析

checkin business user tip review

時刻整合成四個時段
星期整合成平日與周末

下午四點　星期二

16 - 2

時刻　星期幾

midnight early morning
morning
afternoon
night

weekdays
weekends

資料清理
文本分析

checkin business user tip review

```
check_in.head()
```

X5_0	X6_0	...	Hour_21	Hour_22	Hour_23	midnight_earlyMorning	morning	afternoon	night	weekdays	weekends
0	0	...	0	0	0	2	28	192	1	164	59
0	0	...	0	0	1	14	1	4	10	15	14
0	0	...	2	0	1	2	17	16	10	32	13
0	0	...	0	0	0	0	4	13	4	16	5
0	0	...	0	0	0	0	0	6	0	6	0

將categories攤平

	business_id	categories	city	full_address	latitude	longitude	name	neighborhoods	open	review_count	stars	state
0	rncjoVoEFUJGCUoC1JgnUA	[Accountants, Professional Services, Tax Servi...	Peoria	8466 W Peoria Ave\nSte 6\nPeoria, AZ 85345	33.581867	-112.241596	Peoria Income Tax Service	[]	True	3	5.0	AZ
1	0FNFSzCFP_rGUoJx8W7tJg	[Sporting Goods, Bikes, Shopping]	Phoenix	2149 W Wood Dr\nPhoenix, AZ 85029	33.604054	-112.105933	Bike Doctor	[]	True	5	5.0	AZ
2	3f_lyB6vFK48ukH6ScvLHg	[]	Phoenix	1134 N Central Ave\nPhoenix, AZ 85004	33.460526	-112.073933	Valley Permaculture Alliance	[]	True	4	5.0	AZ
3	usAsSV36QmUej8--yvN-dg	[Food, Grocery]	Phoenix	845 W Southern Ave\nPhoenix, AZ 85041	33.392210	-112.085377	Food City	[]	True	5	3.5	AZ
4	PzOqRohWw7F7YEPBz6AubA	[Food, Bagels, Delis, Restaurants]	Glendale Az	6520 W Happy Valley Rd\nSte 101\nGlendale Az	33.712797	-112.200264	Hot Bagels & Deli	[]	True	14	3.5	AZ

	business_id	Food	Bagels	Delis	Restaurants	Sandwiches	Mexican	Pizza	Burgers	...	Live.Raw.Food	Fruits...Veggies
0	usAsSV36QmUej8--yvN-dg	1	0	0	0	0	0	0	0	...	0	0
1	PzOqRohWw7F7YEPBz6AubA	1	1	1	1	0	0	0	0	...	0	0
2	qarobAbxGSHI7ygf1f7a_Q	0	0	0	1	1	0	0	0	...	0	0
3	gA5CuBxF-0CnOpGnryWJdQ	0	0	0	1	0	1	0	0	...	0	0
4	acaBJcFEKPmmSDIO6c-ZGQ	1	0	0	0	0	0	0	0	...	0	0

新增性別欄位

	average_stars	name	review_count	type	user_id	votes	gender
0	5.00	Jim	6	user	CR2y7yEm4X035ZMzrTtN9Q	{'funny': 0, 'useful': 7, 'cool': 0}	Boy
1	1.00	Kelle	2	user	_9GXoHhdxc30ujPaQwh6Ew	{'funny': 0, 'useful': 1, 'cool': 0}	Boy
2	5.00	Stephanie	2	user	8mM-nqxjg6pT04kwcjMbsw	{'funny': 0, 'useful': 1, 'cool': 0}	Girl
3	5.00	T	2	user	Ch6CdTR2IVaVANr-RgIMOg	{'funny': 0, 'useful': 2, 'cool': 0}	Boy
4	1.00	Beth	1	user	NZrLmHRyiHmyT1JrfzkCOA	{'funny': 0, 'useful': 0, 'cool': 0}	Girl

資料清理
文本分析

checkin business user tip review

將votes攤平

	average_stars	name	review_count	type	user_id	votes	gender
0	5.00	Jim	6	user	CR2y7yEm4X035ZMzrTtN9Q	{'funny': 0, 'useful': 7, 'cool': 0}	Boy
1	1.00	Kelle	2	user	_9GXoHhdxc30ujPaQwh6Ew	{'funny': 0, 'useful': 1, 'cool': 0}	Boy
2	5.00	Stephanie	2	user	8mM-nqxjg6pT04kwcjMbsw	{'funny': 0, 'useful': 1, 'cool': 0}	Girl
3	5.00	T	2	user	Ch6CdTR2IVaVANr-RgIMOg	{'funny': 0, 'useful': 2, 'cool': 0}	Boy
4	1.00	Beth	1	user	NZrLmHRyiHmyT1JrfzkCOA	{'funny': 0, 'useful': 0, 'cool': 0}	Girl

資料清理
文本分析

checkin business user tip review

	average_stars	name	review_count	type	user_id	funny	useful	cool
0	5.0	Jim	6.0	review	CR2y7yEm4X035ZMzrTtN9Q	0.0	7.0	0.0
1	1.0	Kelle	2.0	review	_9GXoHhdxc30ujPaQwh6Ew	0.0	1.0	0.0
2	5.0	Stephanie	2.0	review	8mM-nqxjg6pT04kwcjMbsw	0.0	1.0	0.0
3	5.0	T	2.0	review	Ch6CdTR2IVaVANr-RglMOg	0.0	2.0	0.0
4	1.0	Beth	1.0	review	NZrLmHRyiHmyT1JrfzkCOA	0.0	0.0	0.0

將votes攤平

	business_id	date	review_id	stars	text	type	user_id	votes
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	{'funny': 0, 'useful': 5, 'cool': 2}
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbIuQ	{'funny': 0, 'useful': 0, 'cool': 0}
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	{'funny': 0, 'useful': 1, 'cool': 0}
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGaISbqqaMHlNnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	{'funny': 0, 'useful': 2, 'cool': 1}
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	{'funny': 0, 'useful': 0, 'cool': 0}

	business_id	date	review_id	stars	text	type	user_id	funny	useful	cool
0	9yKzy9PApeiPPOUJEtnvkg	2011/1/26	fWKvX83p0-ka4JS3dc6E5A	5.0	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	0	5	2
1	ZRJwVLyzEJq1VAihDhYiow	2011/7/27	IjZ33sJrzXqU-0X6U8NwyA	5.0	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbIuQ	0	0	0
2	6oRAC4uyJCsJl1X0WZpVSA	2012/6/14	IESLBzqUCLdSzSqm0eCSxQ	4.0	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	0	0	0
3	_1QQZuf4zZOyFCvXc0o6Vg	2010/5/27	G-WvGaISbqqaMHlNnByodA	5.0	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	0	2	1
4	6ozycU1RpktNG2-1BroVtw	2012/1/5	1uJFq2r5QfJG_6ExMRCaGw	5.0	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0

做文本分析

	business_id	date	review_id	stars	text	type	user_id	votes
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	{'funny': 0, 'useful': 5, 'cool': 2}
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbIuQ	{'funny': 0, 'useful': 0, 'cool': 0}
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	{'funny': 0, 'useful': 1, 'cool': 0}
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGaISbqqaMHlNnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	{'funny': 0, 'useful': 2, 'cool': 1}
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	{'funny': 0, 'useful': 0, 'cool': 0}

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習特徵欄位

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習特徵欄位

詞 頻

$$TF\text{-}IDF = TF \times IDF$$

$$\frac{\text{某詞出現的次數}}{\text{該篇review的總詞數}}$$

The food was flavorful and plenty of it. Eating with only your fingers is quite fun - make your momma proud. My girlfriend won the best dish of the night - lentils and chicken and cabage/carrots. It all had kind of sweetish/bitterish/spicy flavor. I had the lamb and jalapeno dish, which was kind of greasy. You eat with your fingers and can scoop up the food with the bread they give you - a tartish type crepe, which goes well with the food. I'd be happy to go back and try something different. If you are used to frozen pizzas and Burger King, you might want to stick with that as this experience might bit be as enjoyable as those who are more open to what they consume.

The food was flavorful and plenty of it. Eating with only your fingers is quite fun - make your momma proud. My girlfriend won the best dish of the night - lentils and chicken and cabage/carrots. It all had kind of sweetish/bitterish/spicy flavor. I had the lamb and jalapeno dish, which was kind of greasy. You eat with your fingers and can scoop up the food with the bread they give you - a tartish type crepe, which goes well with the food. I'd be happy to go back and try something different. If you are used to frozen pizzas and Burger King, you might want to stick with that as this experience might bit be as enjoyable as those who are more open to what they consume.

$$TF_{food} : \frac{3(出現次數)}{128(總字數)} = 0.023$$

The food was flavorful and plenty of it. Eating with only your fingers is quite fun - make your momma proud. My girlfriend won the best dish of the night - lentils and chicken and cabage/carrots. It all had kind of sweetish/bitterish/spicy flavor. I had the lamb and jalapeno dish, which was kind of greasy. You eat with your fingers and can scoop up the food with the bread they give you - a tartish type crepe, which goes well with the food. I'd be happy to go back and try something different. If you are used to frozen pizzas and Burger King, you might want to stick with that as this experience might bit be as enjoyable as those who are more open to what they consume.

The food was flavorful and plenty of it. Eating with only your fingers is quite fun - make your momma proud. My girlfriend won the best dish of the night - lentils and chicken and cabage/carrots. It all had kind of sweetish/bitterish/spicy flavor. I had the lamb and jalapeno dish, which was kind of greasy. You eat with your fingers and can scoop up the food with the bread they give you - a tartish type crepe, which goes well with the food. I'd be happy to go back and try something different. If you are used to frozen pizzas and Burger King, you might want to stick with that as this experience might bit be as enjoyable as those who are more open to what they consume.

$$TF_{\text{the}} : \frac{7(\text{出現次數})}{128(\text{總字數})} = 0.055$$

反向文件頻率

$$TF\text{-}IDF = TF \times IDF$$

$$\log\left(\frac{\text{全部評論的總篇數}}{\text{包含該詞的評論總數}}\right)$$

1000篇

包含 food 的文章

100

$IDF_{food} = 1$

960

包含the的文章

$IDF_{the} = 0.018$

$$TF_{food} \times IDF_{food} = 0.023$$

$$TF_{the} \times IDF_{the} = 0.001$$

TF x IDF(重要程度) : food > the

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習特徵欄位

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step1
計算各詞TF-IDF

建立目標字字典

目標字

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

Step1
計算各詞TF-IDF

建立目標字字典

目標字

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

order 訂單
staff 工作人員
service 服務

Step1
計算各詞TF-IDF

建立目標字字典

目標字

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

place 位置
location 地點
area 區域

Step1
計算各詞TF-IDF

建立目標字字典

目標字

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

atmosphere 氣氛

fun 樂趣

Step1
計算各詞TF-IDF

建立目標字字典

目標字

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

price 價格
discount 折扣

Step1
計算各詞TF-IDF

建立目標字字典

目標字

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

lunch 午餐
flavor 口味
dish 菜餚

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習特徵欄位

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習
特徵欄位

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

Nice location , good service but a little slow.

建立目標字字典

目標字

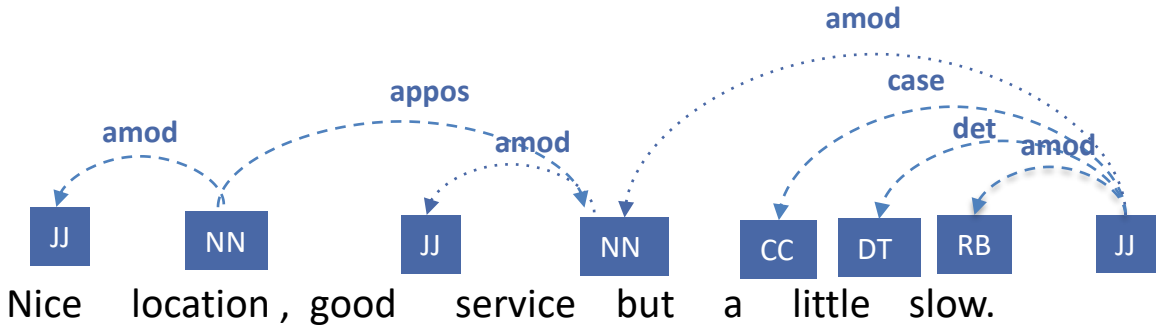
五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞



建立目標字字典

目標字

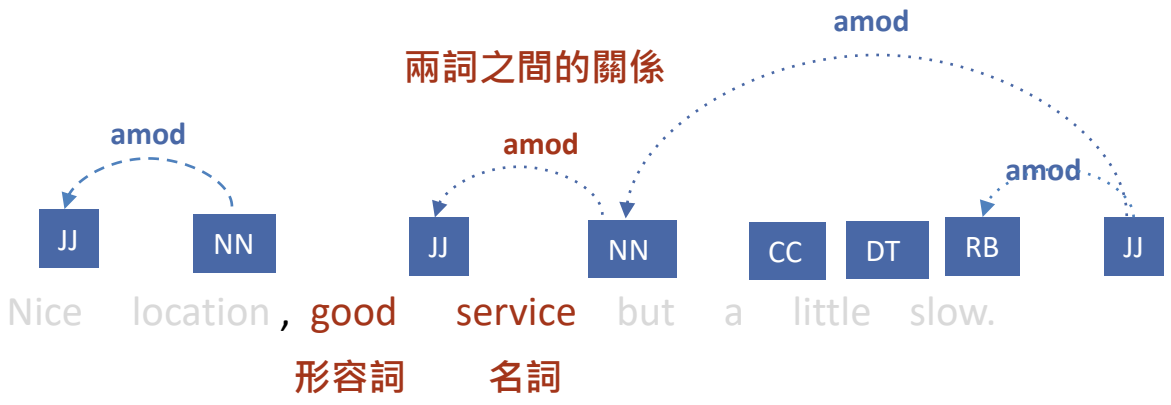
五大特質

- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞



建立目標字字典

目標字

五大特質

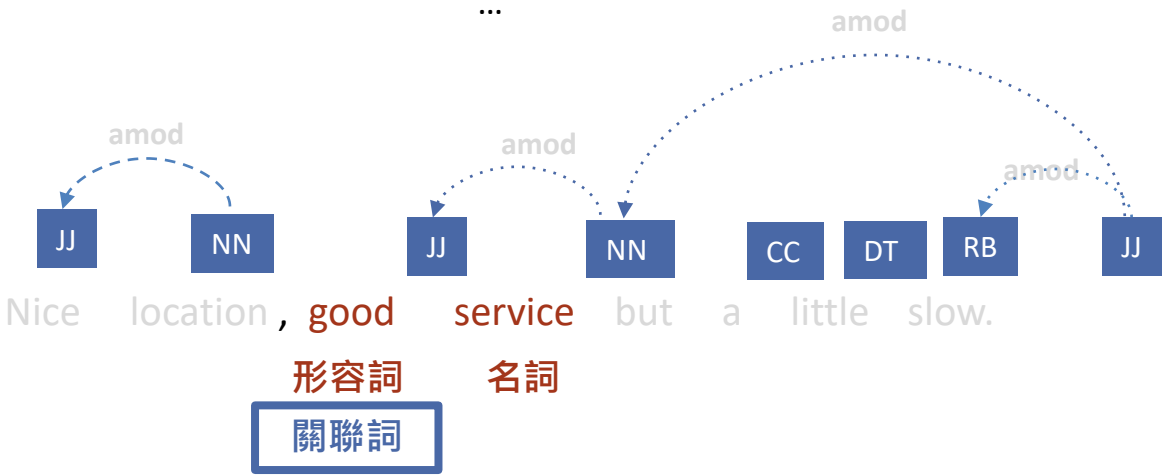
- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

awful
good
amazing
...



Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習
特徵欄位

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習
特徵欄位

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

awful
good
amazing
…

(0.2345556 , 0.6849237 , 0.11234555……)
(0.4345009 , 0.7309724 , 0.39986501……)
(0.2901128 , 0.1119273 , 0.88495060……)
…

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習
特徵欄位

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習
特徵欄位

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

詞向量化

向量化後的詞

awful	(0.2345556 , 0.6849237 , 0.11234555.....)
good	(0.4345009 , 0.7309724 , 0.39986501.....)
amazing	(0.2901128 , 0.1119273 , 0.88495060.....)

great	+1
awsome	+1
bad	-1
...	

情感分類

awful	-1
good	+1
amazing	+1

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

詞向量化

向量化後的詞

awful	(0.2345556 , 0.6849237 , 0.11234555.....)
good	(0.4345009 , 0.7309724 , 0.39986501.....)
amazing	(0.2901128 , 0.1119273 , 0.88495060.....)
...	...

great +1
awsome +1
bad -1
...

情感分類

awful	-1
good	+1
amazing	+1
...	

否定詞 * 關聯詞

not * good
not * bad
...

(-1) * good (+1)
(-1) * bad (-1)
...

not * good	-1
not * bad	+1
...	

Step1
計算各詞TF-IDF

建立目標字字典

目標字

Step2
Dependency Parsing

建立關聯詞字庫

關聯詞

Step 3
word2Vec

詞向量化

向量化後的詞

Step 4
丟入情緒字庫做機器學習
opinion-lexicon-English / Bing Liu

情感分類

Step5
針對五大特質計算分數

作為機器學習
特徵欄位

	business_id	atmosphere	food	service	price	place	overall
0	9yKzy9PApeiPPOUJEtnvkg	6.0	15.0	2.0	0.0	33.0	11.2
1	ZRJwVLyzEJq1VAihDhYiow	5.0	30.0	17.0	3.0	-1.0	10.8
2	6oRAC4uyJCsJl1X0WZpVSA	2.0	3.0	11.0	16.0	13.0	9.0
3	-yxfBYGB6SEqszmxJxd97A	-1.0	0.0	-1.0	0.0	0.0	-0.4
4	zp713qNhx8d9KCJJnrw1xA	5.0	43.0	6.0	10.0	19.0	16.6
5	wNUea3IXZWD63bbOQaOH-g	13.0	36.0	2.0	8.0	13.0	14.4
6	nMHhuYan8e3cONo3PornJA	9.0	44.0	6.0	6.0	19.0	16.8
7	e9nN4XxjdHj4qtKCOPq_vg	-2.0	24.0	-1.0	5.0	22.0	9.6
8	h53YuCilDfEFSJCQpk8v1g	2.0	6.0	0.0	0.0	7.0	3.0
9	yc5AH9H71xJidA_J2mChLA	8.0	8.0	10.0	4.0	13.0	8.6

非關聯式
資料庫

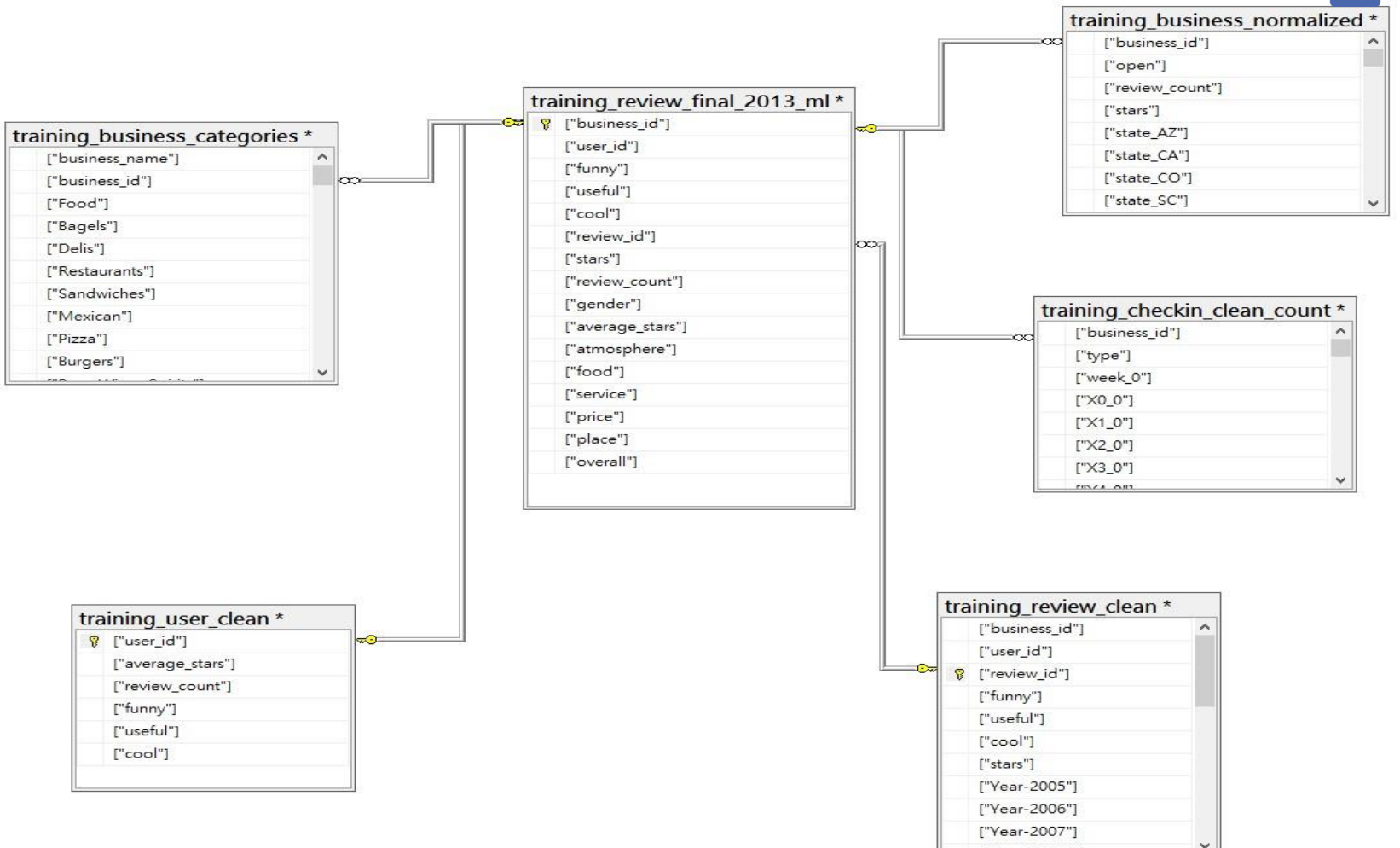
資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析

Microsoft® SQL Server®


training_business_normalized *
["business_id"]
["open"]
["review_count"]
["stars"]
["state_AZ"]
["state_CA"]
["state_CO"]
["state_SC"]
training_business_categories *
["business_name"]
["business_id"]
["Food"]
["Bagels"]
["Delis"]
["Restaurants"]
["Sandwiches"]
["Mexican"]
["Pizza"]
["Burgers"]
training_review_final_2013_ml *
["business_id"]
["user_id"]
["funny"]
["useful"]
["cool"]
["review_id"]
["stars"]
["review_count"]
["gender"]
["average_stars"]
["atmosphere"]
["food"]
["service"]
["price"]
["place"]
["overall"]
training_checkin_clean_count *
["business_id"]
["type"]
["week_0"]
["X0_0"]
["X1_0"]
["X2_0"]
["X3_0"]
training_user_clean *
["user_id"]
["average_stars"]
["review_count"]
["funny"]
["useful"]
["cool"]
training_review_clean *
["business_id"]
["user_id"]
["review_id"]
["funny"]
["useful"]
["cool"]
["stars"]
["Year-2005"]
["Year-2006"]
["Year-2007"]

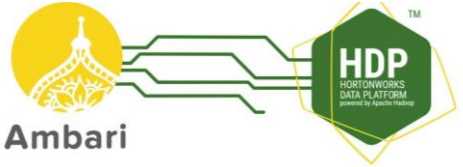
非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析




hadoop

ANSIBLE

多集群一致化管理

利用SSH與遠端Server溝通並控制機群
一致化配置及安裝,並同步更新及環境維護

藉由撰寫”Playbook”，
實現從server端管理多機群架構
可依需求改變Client的群組及
接受命令的對象
維持多機群環境一致化管理

```
---
- name: install maven
  hosts: hadoop_other
  remote_user: student
  become: yes
  become_method: sudo
  become_user: root


  tasks:
   - name: set variables
     set_fact: dir="/opt/modules"
   - name: rm
     file: path=/opt/modules/maven state=absent
   - name: untar
     unarchive:
      src=/opt/softwares/apache-maven-3.6.1-bin.tar.gz
      dest={{dir}}
   - name: mv dirctory
     shell: mv /opt/modules/apache-maven-3.6.1 /opt/modules/maven
   - name: set env
     lineinfile:
         dest: /etc/profile
         line: "{{ item.line }}"
     with_items:
         - { line: '# MAVEN_HOME' }
         - { line: 'export MAVEN_HOME=/opt/modules/maven' }
         - { line: 'export PATH=$MAVEN_HOME/bin:$PATH' }
```

藉由撰寫”Playbook”，
實現從server端管理多機群架構
可依需求改變Client的群組及
接受命令的對象
維持多機群環境一致化管理

```
---
- name: install maven
  hosts: hadoop_other
  remote_user: student
  become: yes
  become_method: sudo
  become_user: root


  tasks:
   - name: set variables
     set_fact: dir="/opt/modules"
   - name: rm
     file: path=/opt/modules/maven state=absent
   - name: untar
     unarchive:
      src=/opt/softwares/apache-maven-3.6.1-bin.tar.gz
      dest={{dir}}
   - name: mv dirctory
     shell: mv /opt/modules/apache-maven-3.6.1 /opt/modules/maven
   - name: set env
     lineinfile:
         dest: /etc/profile
         line: "{{ item.line }}"
     with_items:
         - { line: '# MAVEN_HOME' }
         - { line: 'export MAVEN_HOME=/opt/modules/maven' }
         - { line: 'export PATH=$MAVEN_HOME/bin:$PATH' }
```

藉由撰寫”Playbook”，
實現從server端管理多機群架構
可依需求改變Client的群組及
接受命令的對象
維持多機群環境一致化管理

```
---
- name: install maven
  hosts: hadoop_other
  remote_user: student
  become: yes
  become_method: sudo
  become_user: root


  tasks:
   - name: set variables
     set_fact: dir="/opt/modules"
   - name: rm
     file: path=/opt/modules/maven state=absent
   - name: untar
     unarchive:
      src=/opt/softwares/apache-maven-3.6.1-bin.tar.gz
      dest={{dir}}
   - name: mv dirctory
     shell: mv /opt/modules/apache-maven-3.6.1 /opt/modules/maven
   - name: set env
     lineinfile:
         dest: /etc/profile
         line: "{{ item.line }}"
     with_items:
         - { line: '# MAVEN_HOME' }
         - { line: 'export MAVEN_HOME=/opt/modules/maven' }
         - { line: 'export PATH=$MAVEN_HOME/bin:$PATH' }
```

藉由撰寫”Playbook”，
實現從server端管理多機群架構
可依需求改變Client的群組及
接受命令的對象
維持多機群環境一致化管理

```
---
- name: install maven
  hosts: hadoop_other
  remote_user: student
  become: yes
  become_method: sudo
  become_user: root


  tasks:
   - name: set variables
     set_fact: dir="/opt/modules"
   - name: rm
     file: path=/opt/modules/maven state=absent
   - name: untar
     unarchive:
      src=/opt/softwares/apache-maven-3.6.1-bin.tar.gz
      dest={{dir}}
   - name: mv dirctory
     shell: mv /opt/modules/apache-maven-3.6.1 /opt/modules/maven
   - name: set env
     lineinfile:
         dest: /etc/profile
         line: "{{ item.line }}"
     with_items:
         - { line: '# MAVEN_HOME' }
         - { line: 'export MAVEN_HOME=/opt/modules/maven' }
         - { line: 'export PATH=$MAVEN_HOME/bin:$PATH' }
```

實作統一安裝
Apache Maven與HDP-util
Install_Maven_HDP.yml

```
changed: [cluster10]
changed: [cluster12]
changed: [cluster11]

TASK [untar HDP-UTILS] *********************************************************
changed: [cluster12]
changed: [cluster10]
changed: [cluster11]

TASK [change HDP-UTILS owner and group] ****************************************
changed: [cluster10]
changed: [cluster11]
changed: [cluster12]

TASK [install apache2] *********************************************************
 [WARNING]: Consider using the apt module rather than running 'apt-get'.  If you need
to use command because apt is insufficient you can add 'warn: false' to this command
task or set 'command_warnings=False' in ansible.cfg to get rid of this message.

changed: [cluster12]
changed: [cluster11]
changed: [cluster10]

TASK [copy modules into /var/www/html] *****************************************
changed: [cluster10]
changed: [cluster12]
changed: [cluster11]

PLAY RECAP *********************************************************************
cluster10                  : ok=8    changed=6    unreachable=0    failed=0    skipped=0    rescued=0    i
gnored=0
cluster11                  : ok=8    changed=6    unreachable=0    failed=0    skipped=0    rescued=0    i
gnored=0
cluster12                  : ok=8    changed=6    unreachable=0    failed=0    skipped=0    rescued=0    i
gnored=0
```

實作叢集ssh無密碼登入

sshpasswordless.yml

```
---
- name: sshpasswordless
  hosts: ssh1

  tasks:
   - name: close ssh check
     shell: sudo sed -i "s/^.*StrictHostKeyChecking.*$/ StrictHostKeyChecking no/g" /etc/ssh/ssh_config
   - name: delete /home/stu/.ssh/
     file: path=/home/stu/.ssh/ state=absent
   - name: generating public/private rsa key pair
     shell: ssh-keygen -t rsa -b 2048 -N '' -f /home/stu/.ssh/id_rsa
   - name: view id_rsa.pub
     shell: cat /home/stu/.ssh/id_rsa.pub
     register: sshinfo
   - set_fact: sshpub={{sshinfo.stdout}}
   - name: add ssh record
     local_action: shell echo {{sshpub}} >> /etc/ansible/roles/templates/authorized_keys.j2
   - name: copy authorized_keys.j2 to all
     template: src=/etc/ansible/roles/templates/authorized_keys.j2 dest=/home/stu/.ssh/authorized_keys mode=0600
```

實作叢集ssh無密碼登入

sshpasswordless.yml

```
---
- name: sshpasswordless
  hosts: ssh1

  tasks:
   - name: close ssh check
     shell: sudo sed -i "s/^.*StrictHostKeyChecking.*$/ StrictHostKeyChecking no/g" /etc/ssh/ssh_config
   - name: delete /home/stu/.ssh/
     file: path=/home/stu/.ssh/ state=absent
   - name: generating public/private rsa key pair
     shell: ssh-keygen -t rsa -b 2048 -N '' -f /home/stu/.ssh/id_rsa
   - name: view id_rsa.pub
     shell: cat /home/stu/.ssh/id_rsa.pub
     register: sshinfo
   - set_fact: sshpub={{sshinfo.stdout}}
   - name: add ssh record
     local_action: shell echo {{sshpub}} >> /etc/ansible/roles/templates/authorized_keys.j2
   - name: copy authorized_keys.j2 to all
     template: src=/etc/ansible/roles/templates/authorized_keys.j2 dest=/home/stu/.ssh/authorized_keys mode=0600
```

實作叢集ssh無密碼登入

sshpasswordless.yml

```
---
- name: sshpasswordless
  hosts: ssh1

  tasks:
   - name: close ssh check
     shell: sudo sed -i "s/^.*StrictHostKeyChecking.*$/ StrictHostKeyChecking no/g" /etc/ssh/ssh_config
   - name: delete /home/stu/.ssh/
     file: path=/home/stu/.ssh/ state=absent
   - name: generating public/private rsa key pair
     shell: ssh-keygen -t rsa -b 2048 -N '' -f /home/stu/.ssh/id_rsa
   - name: view id_rsa.pub
     shell: cat /home/stu/.ssh/id_rsa.pub
     register: sshinfo
   - set_fact: sshpub={{sshinfo.stdout}}
   - name: add ssh record
     local_action: shell echo {{sshpub}} >> /etc/ansible/roles/templates/authorized_keys.j2
   - name: copy authorized_keys.j2 to all
     template: src=/etc/ansible/roles/templates/authorized_keys.j2 dest=/home/stu/.ssh/authorized_keys mode=0600
```

實作叢集ssh無密碼登入

sshpasswordless.yml

```
---
- name: sshpasswordless
  hosts: ssh1

  tasks:
   - name: close ssh check
     shell: sudo sed -i "s/^.*StrictHostKeyChecking.*$/ StrictHostKeyChecking no/g" /etc/ssh/ssh_config
   - name: delete /home/stu/.ssh/
     file: path=/home/stu/.ssh/ state=absent
   - name: generating public/private rsa key pair
     shell: ssh-keygen -t rsa -b 2048 -N '' -f /home/stu/.ssh/id_rsa
   - name: view id_rsa.pub
     shell: cat /home/stu/.ssh/id_rsa.pub
     register: sshinfo
   - set_fact: sshpub={{sshinfo.stdout}}
   - name: add ssh record
     local_action: shell echo {{sshpub}} >> /etc/ansible/roles/templates/authorized_keys.j2
   - name: copy authorized_keys.j2 to all
     template: src=/etc/ansible/roles/templates/authorized_keys.j2 dest=/home/stu/.ssh/authorized_keys mode=0600
```

實作叢集ssh無密碼登入

sshpasswordless.yml

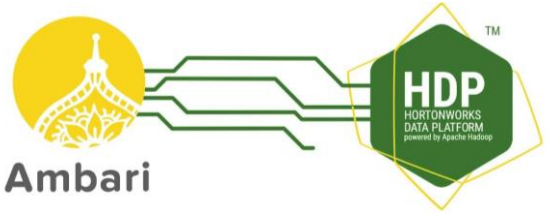
```
TASK [delete /home/stu/.ssh/] ***************************************************************************************************************************
ok: [192.168.35.201]
changed: [192.168.35.200]

TASK [generating public/private rsa key pair] ***********************************************************************************************************
changed: [192.168.35.200]
changed: [192.168.35.201]

TASK [view id_rsa.pub] **********************************************************************************************************************************
changed: [192.168.35.200]
changed: [192.168.35.201]

TASK [set_fact] *****************************************************************************************************************************************
ok: [192.168.35.200]
ok: [192.168.35.201]

TASK [add ssh record] ***********************************************************************************************************************************
changed: [192.168.35.200 -> localhost]
changed: [192.168.35.201 -> localhost]

TASK [copy authorized_keys.j2 to all] *******************************************************************************************************************
changed: [192.168.35.200]
changed: [192.168.35.201]

PLAY RECAP **********************************************************************************************************************************************
192.168.35.200             : ok=8    changed=6    unreachable=0    failed=0    skipped=0    rescued=0    ignored=0
192.168.35.201             : ok=8    changed=5    unreachable=0    failed=0    skipped=0    rescued=0    ignored=0
```

Ambari

HDP HORTONWORKS DATA PLATFORM powered by Apache Hadoop

叢集自動化搭建&監控

圖形化介面安裝、配置叢集
動態更新、升級、修改叢集功能及各節點配置
監控個節點狀態及資源使用

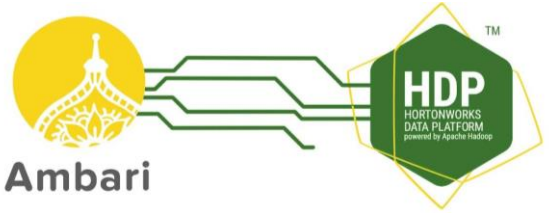


- 圖形化介面選擇欲安裝的專案及功能
- 依需求快速修改叢集配置及安裝套件
- 有升級需求時可以快速使用圖形化介面進行升級

Choose Services

Choose which services you want to install on your cluster.

	Service	Version	Description
☑	YARN + MapReduce2	3.1.1	Apache Hadoop NextGen MapReduce (YARN)
☐	Tez	0.9.1	Tez is the next generation Hadoop Query Processing framework written on top of YARN.
☑	Hive	3.1.0	Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
☑	HBase	2.0.2	Non-relational distributed database and centralized service for configuration management & synchronization
☐	Pig	0.16.0	Scripting platform for analyzing large datasets
☑	Sqoop	1.4.7	Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases
☐	Oozie	4.3.1	System for workflow coordination and execution of Apache Hadoop jobs. This also includes the installation of the optional Oozie Web Console which relies on and will install the ExtJS Library.
☑	ZooKeeper	3.4.6	Centralized service which provides highly reliable distributed coordination
☐	Storm	1.2.1	Apache Hadoop Stream processing framework
☐	Accumulo	1.7.0	Robust, scalable, high performance distributed key/value store.
☐	Infra Solr	0.1.0	Core shared service used by Ambari managed components.
☑	Ambari Metrics	0.1.0	A system for metrics collection that provides storage and retrieval capability for metrics collected from the cluster
☐	Atlas	1.1.0	Atlas Metadata and Governance platform
☑	Kafka	2.0.0	A high-throughput distributed messaging system
☐	Knox	1.0.0	Provides a single point of authentication and access for Apache Hadoop services in a cluster
☐	Log Search	0.5.0	Log aggregation, analysis, and visualization for Ambari managed services. This service is **Technical Preview**.
☐	Ranger	1.2.0	Comprehensive security for Hadoop
☐	Ranger KMS	1.2.0	Key Management Server
☑	SmartSense	1.5.1.2.7.3.0-139	SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
☑	Spark2	2.3.2	Apache Spark 2.3 is a fast and general engine for large-scale data processing.



- 圖形化介面選擇欲安裝的套件及功能
- 依需求快速修改叢集配置及安裝套件
- 有升級需求時可以快速使用圖形化介面進行升級

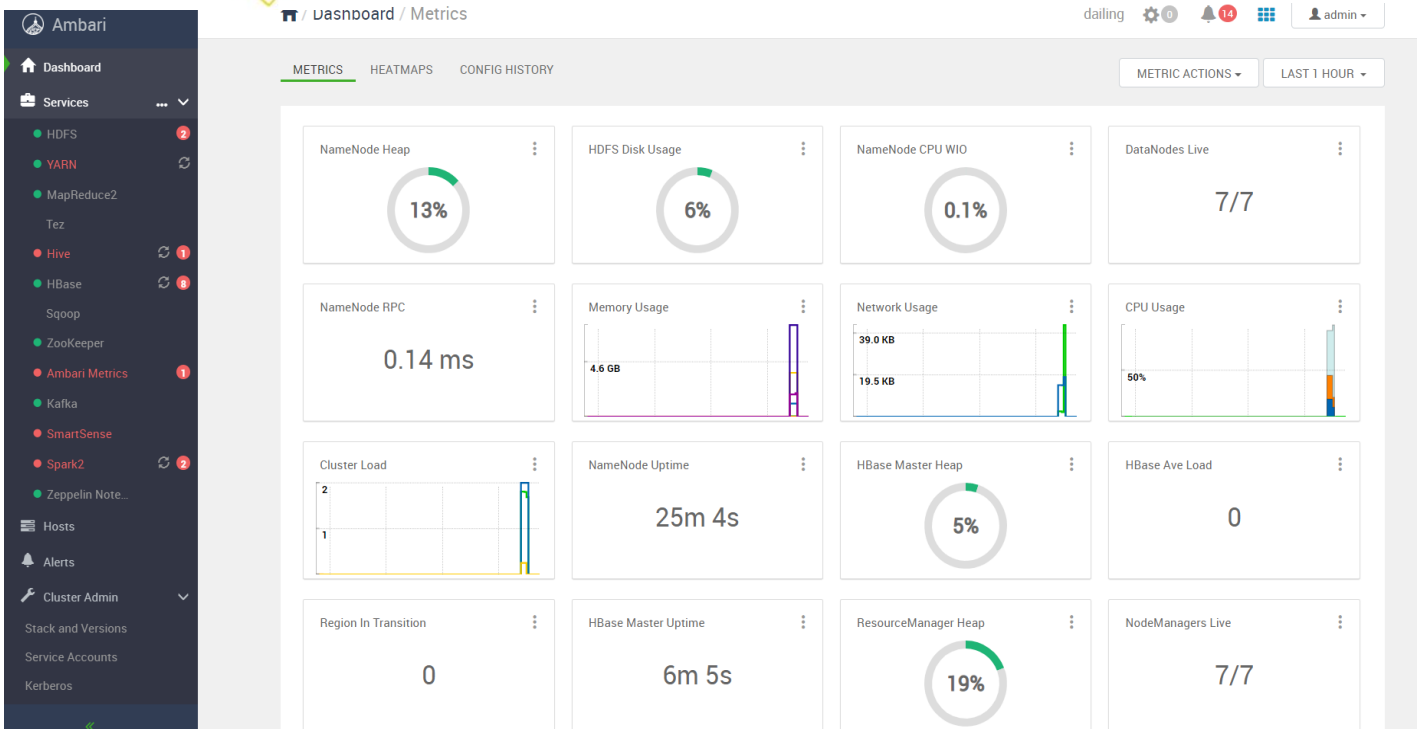
Assign Slaves and Clients

Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with *.
"Client" will install HDFS Client, YARN Client, MapReduce2 Client, Tez Client, Hive Client, HBase Client, Sqoop Client, ZooKeeper Client and Spark2 Client.

Host	all | none	all | none	all | none	all | none	all | none	all | none	all | none	all | none
cluster01.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster02.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster03.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster04.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster05.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster06.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster07.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster08.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster09.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster11.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client
cluster12.example.org*	DataNode	NFSGateway	NodeManager	RegionServer	Phoenix Query Server	Livy for Spark2 Server	Spark2 Thrift Server	Client

Ambari

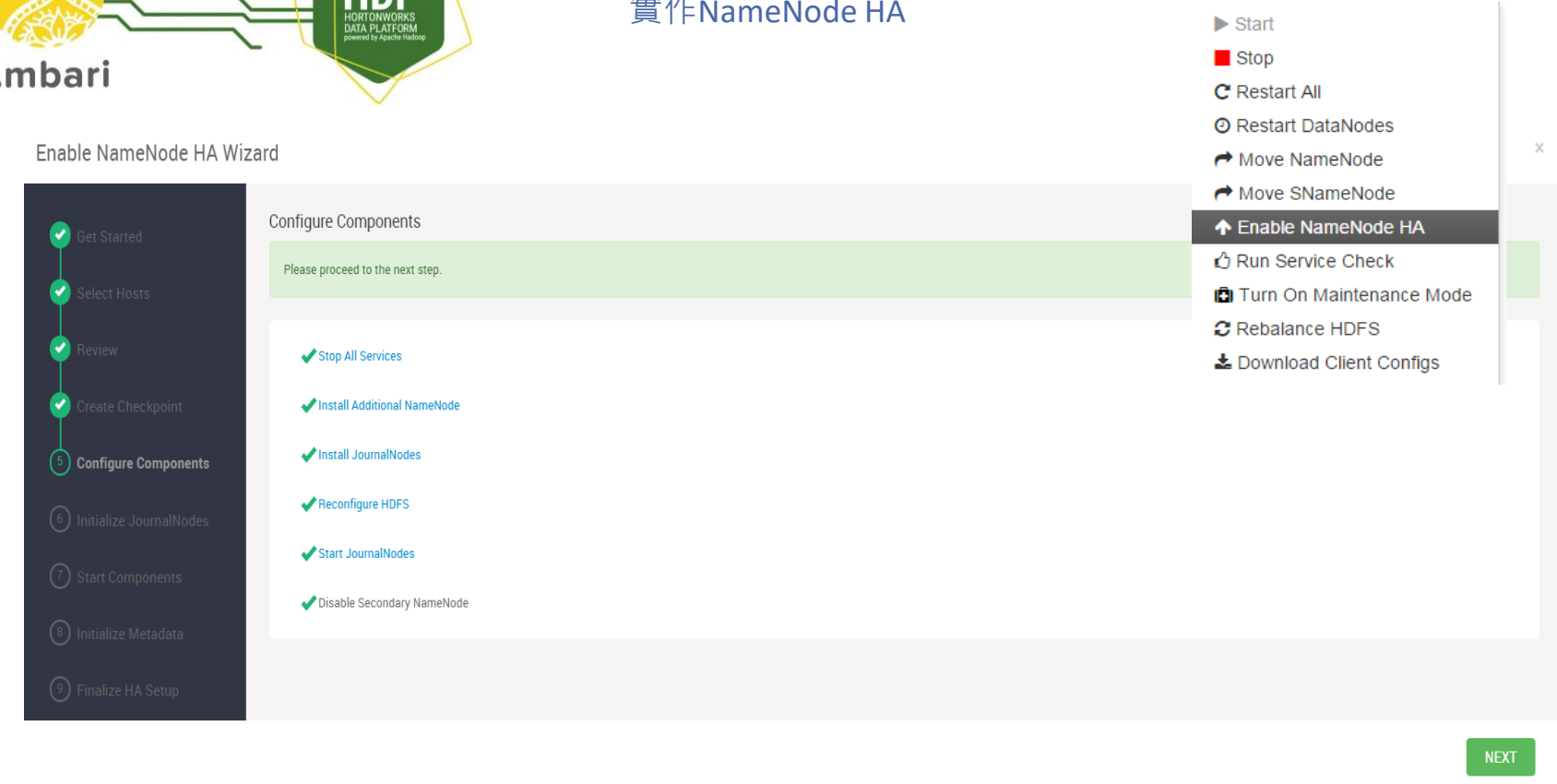
- 圖形化介面快速監控各節點及功能運行狀態、資源使用效率等
- 圖形介面操作快速啟動關閉節點功能



Ambari

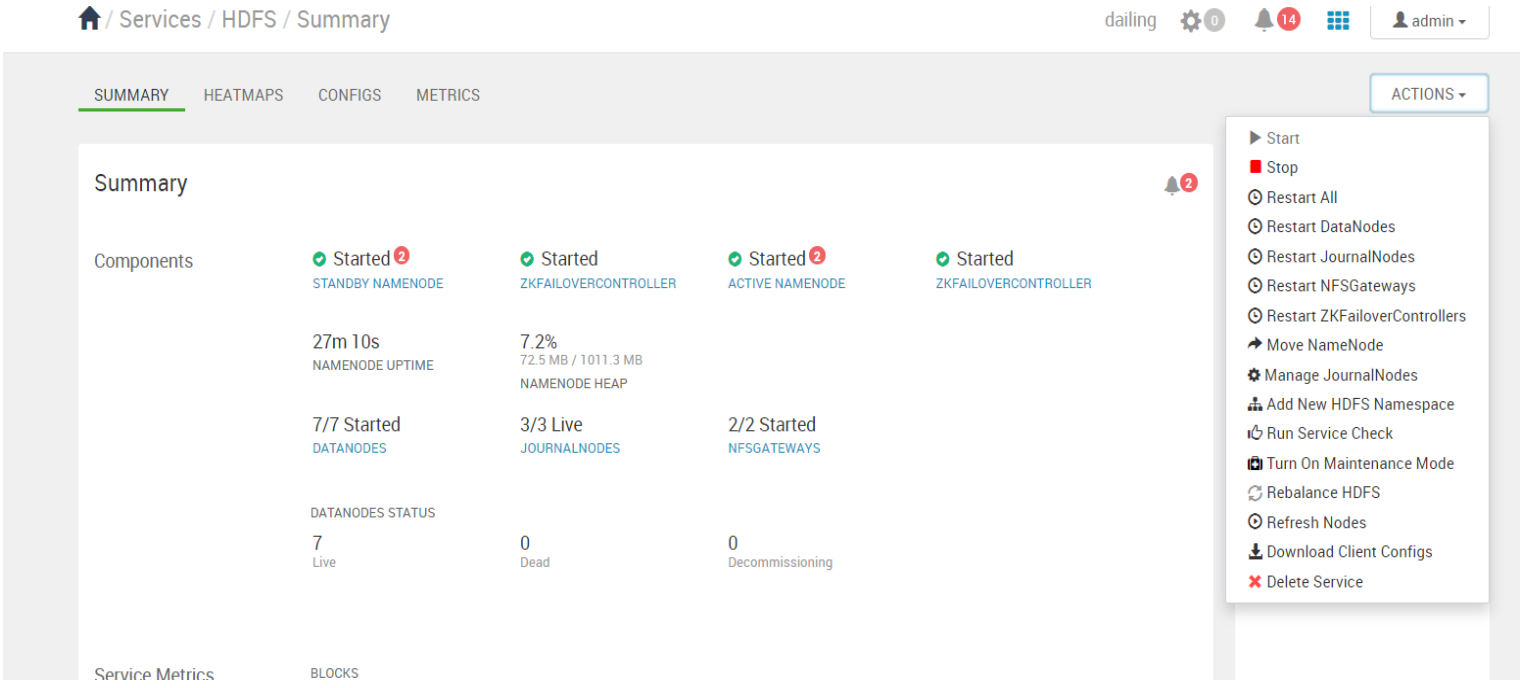
HDP HORTONWORKS DATA PLATFORM powered by Apache Hadoop

圖形化介面管理範例
實作NameNode HA



Ambari

- 圖形化介面快速監控各節點及功能運行狀態、資源使用效率等
- 圖形介面操作快速啟動關閉節點功能

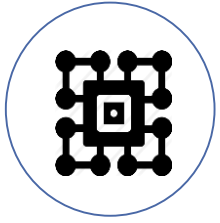


主機
4 台

虛擬機
12 台

趙上涵	192.168.35.85 192.168.35.89 192.168.35.90
林庭宇	192.168.35.141 192.168.35.142 192.168.35.146
薛正暐	192.168.35.197 192.168.35.198 192.168.35.199
吳岱凌	192.168.35.92 192.168.35.93 192.168.35.94

記憶體
96 G



CPU
24

HDD
6T

Client

bdse93
192.168.35.93

bdse142
192.168.35.142

bdse94
192.168.35.94

bdse146
192.168.35.146

NameNode

bdse85
192.168.35.85

bdse92
192.168.35.92

Hadoop Cluster

ANSIBLE

Ambari

HDP

bdse198
192.168.35.198

ResourceManager

bdse141
192.168.35.141

bdse197
192.168.35.197

Worker

Apache Zookeeper

JOURNAL NODE

bdse90
192.168.35.90

Apache Zookeeper

JOURNAL NODE

bdse93
192.168.35.93

Apache Zookeeper

JOURNAL NODE

bdse142
192.168.35.142

HISTORY SERVER

bdse89
192.168.35.89

bdse94
192.168.35.94

bdse146
192.168.35.146

bdse199
192.168.35.199

Client

bdse93
192.168.35.93

bdse142
192.168.35.142

bdse94
192.168.35.94

bdse146
192.168.35.146

NameNode

bdse85
192.168.35.85

bdse92
192.168.35.92

Hadoop Cluster

ANSIBLE

Ambari

HDP

bdse198
192.168.35.198

ResourceManager

bdse141
192.168.35.141

bdse197
192.168.35.197

Worker

Apache Zookeeper

JOURNAL NODE

bdse90
192.168.35.90

Apache Zookeeper

JOURNAL NODE

bdse93
192.168.35.93

Apache Zookeeper

JOURNAL NODE

bdse142
192.168.35.142

HISTORY SERVER

bdse89
192.168.35.89

bdse94
192.168.35.94

bdse146
192.168.35.146

bdse199
192.168.35.199

Client

bdse93
192.168.35.93

bdse142
192.168.35.142

bdse94
192.168.35.94

bdse146
192.168.35.146

NameNode

bdse85
192.168.35.85

bdse92
192.168.35.92

ANSIBLE

Ambari

HDP

bdse198
192.168.35.198

Hadoop Cluster

ResourceManager

bdse141
192.168.35.141

bdse197
192.168.35.197

Worker

Apache Zookeeper

JOURNAL NODE

bdse90
192.168.35.90

Apache Zookeeper

JOURNAL NODE

bdse93
192.168.35.93

Apache Zookeeper

JOURNAL NODE

bdse142
192.168.35.142

HISTORY SERVER

bdse89
192.168.35.89

bdse94
192.168.35.94

bdse146
192.168.35.146

bdse199
192.168.35.199

Client


bdse93
192.168.35.93


bdse142
192.168.35.142


bdse94
192.168.35.94


bdse146
192.168.35.146

NameNode


bdse85
192.168.35.85

bdse92
192.168.35.92

ANSIBLE

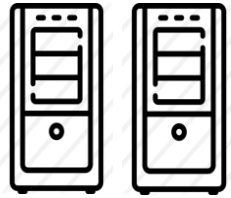
Ambari

HDP


bdse198
192.168.35.198

Hadoop Cluster

ResourceManager


bdse141
192.168.35.141

bdse197
192.168.35.197

Worker



Apache Zookeeper

JOURNAL NODE
bdse90
192.168.35.90



Apache Zookeeper

JOURNAL NODE
bdse93
192.168.35.93



Apache Zookeeper

JOURNAL NODE
bdse142
192.168.35.142



HISTORY SERVER
bdse89
192.168.35.89


bdse94
192.168.35.94


bdse146
192.168.35.146


bdse199
192.168.35.199

Client

bdse93
192.168.35.93

bdse142
192.168.35.142

bdse94
192.168.35.94

bdse146
192.168.35.146

NameNode

bdse85
192.168.35.85

bdse92
192.168.35.92

Hadoop Cluster

ANSIBLE

Ambari

HDP

bdse198
192.168.35.198

ResourceManager

bdse141
192.168.35.141

bdse197
192.168.35.197

Worker

bdse90
192.168.35.90
Apache Zookeeper
JOURNAL NODE

bdse93
192.168.35.93
Apache Zookeeper
JOURNAL NODE

bdse142
192.168.35.142
Apache Zookeeper
JOURNAL NODE

bdse89
192.168.35.89
HISTORY SERVER

bdse94
192.168.35.94

bdse146
192.168.35.146

bdse199
192.168.35.199

04 專案流程

非關聯式
資料庫

資料預處理

匯入
關聯式資料庫

大數據
平台架設

資料分析

吳岱凌

專案目標

進行用戶評分星等預測

機器學習

深度學習

Collaborative Filtering

吳岱凌

專案目標

進行用戶評分星等預測

機器學習

深度學習

Collaborative Filtering

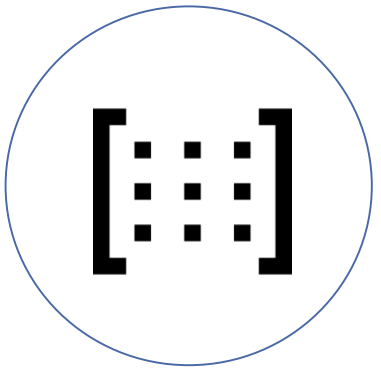
小丙

韓食館

XGBoost

LUDWIG

Collaborative Filtering



XGBoost

LUDWIG

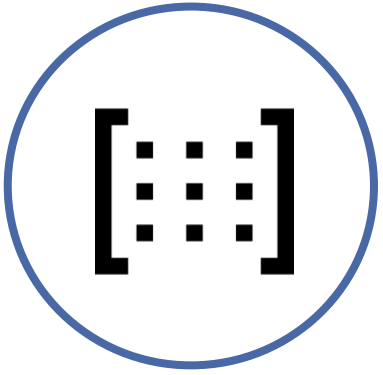
Collaborative Filtering

由Uber開發，
實作於Tensorflow上的深度學習API

XGBoost

LUDWIG

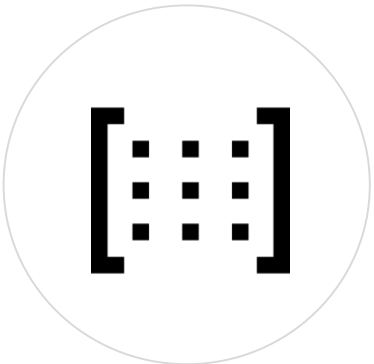
Collaborative Filtering



XGBoost

LUDWIG

Collaborative Filtering



建立特徵欄位
做評分星等預測

XGBoost

LUDWIG

Collaborative Filtering

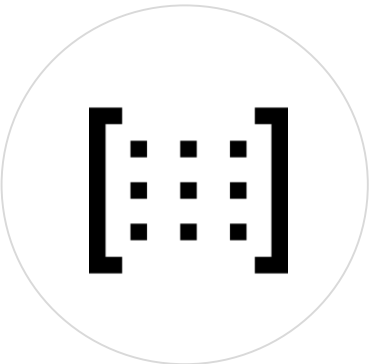
	九州鬆餅	韓食館	樂麵屋	操場酒吧
小甲	3	5	4	4
小乙	2	2	2	1
小丙	4	?	5	4

用戶評分星等

XGBoost

LUDWIG

Collaborative Filtering



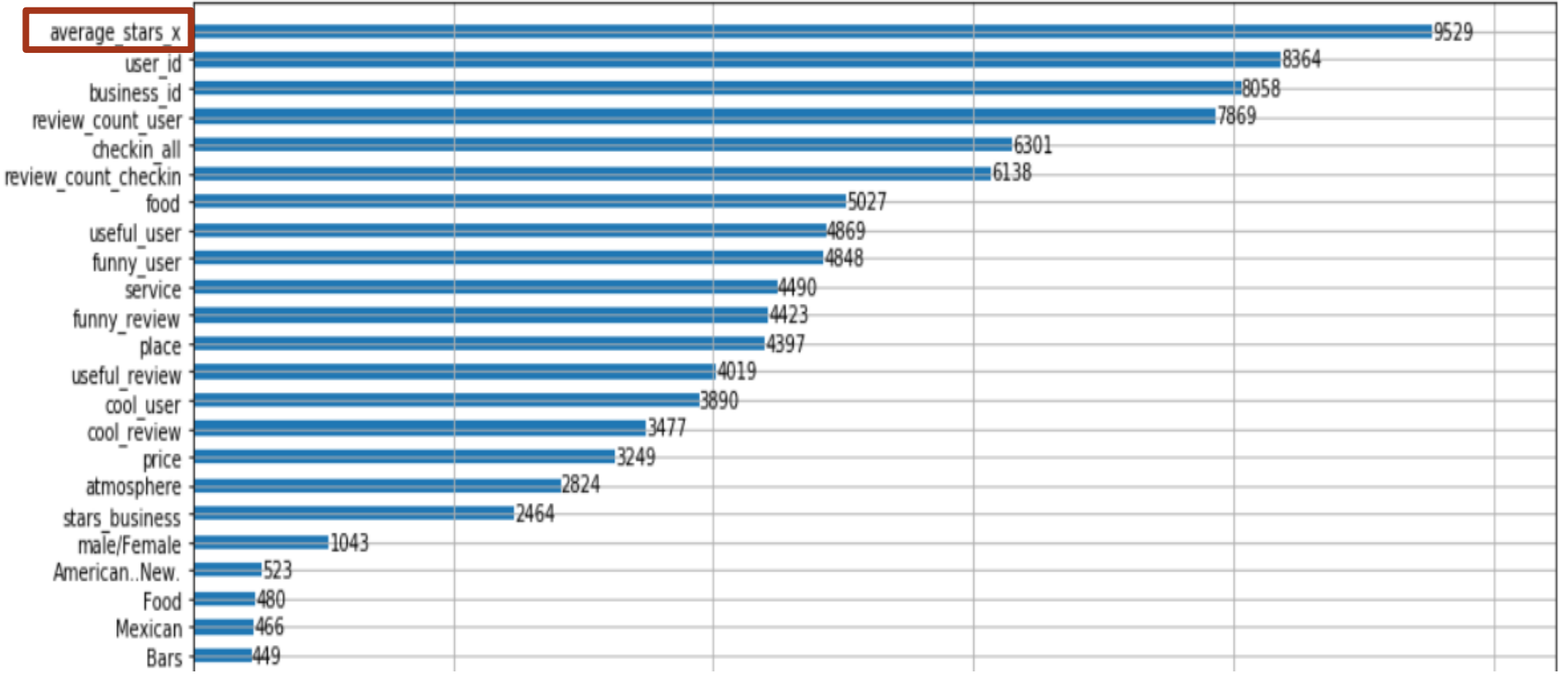
XGBOOST

Ludwig

協同過濾

特徵欄位重要程度

用戶過去給出的平均星等

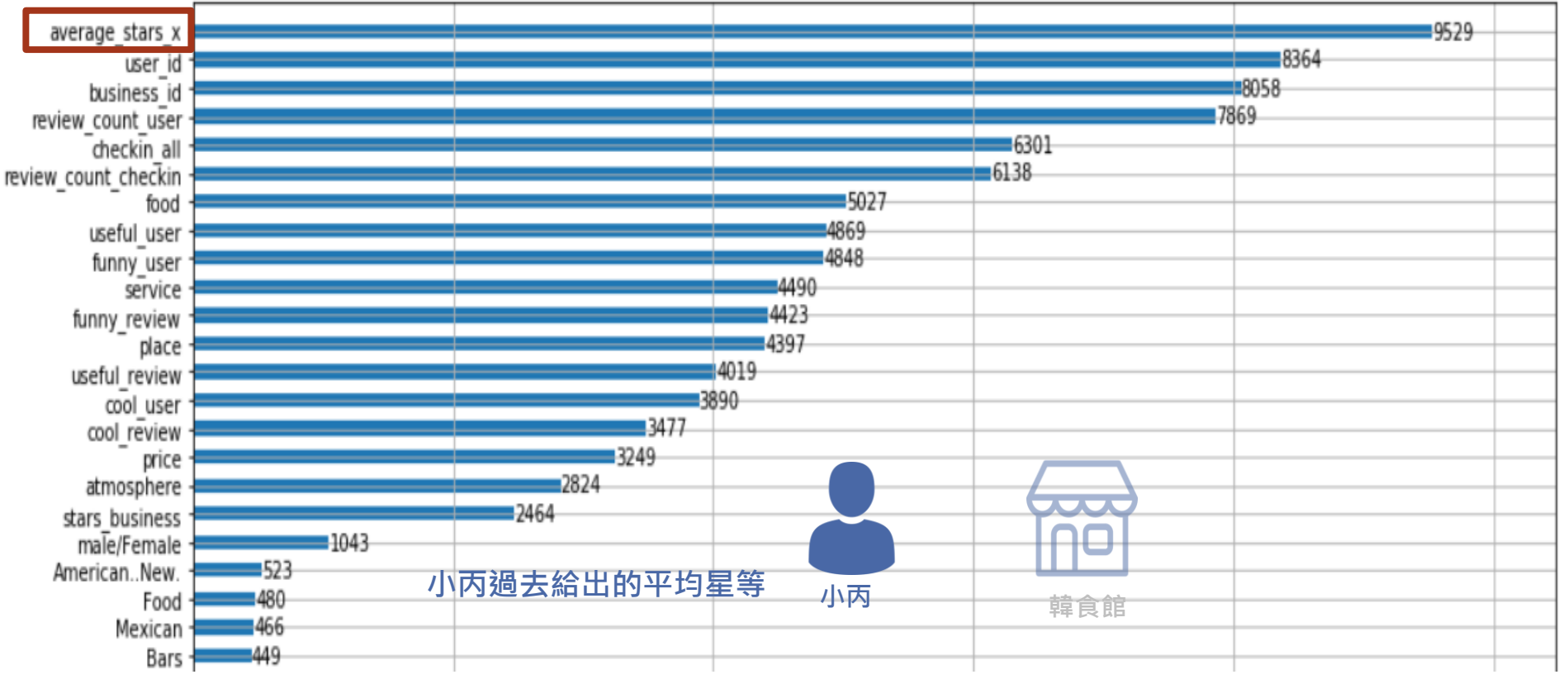


XGBOOST

Ludwig

協同過濾

特徵欄位
重要程度

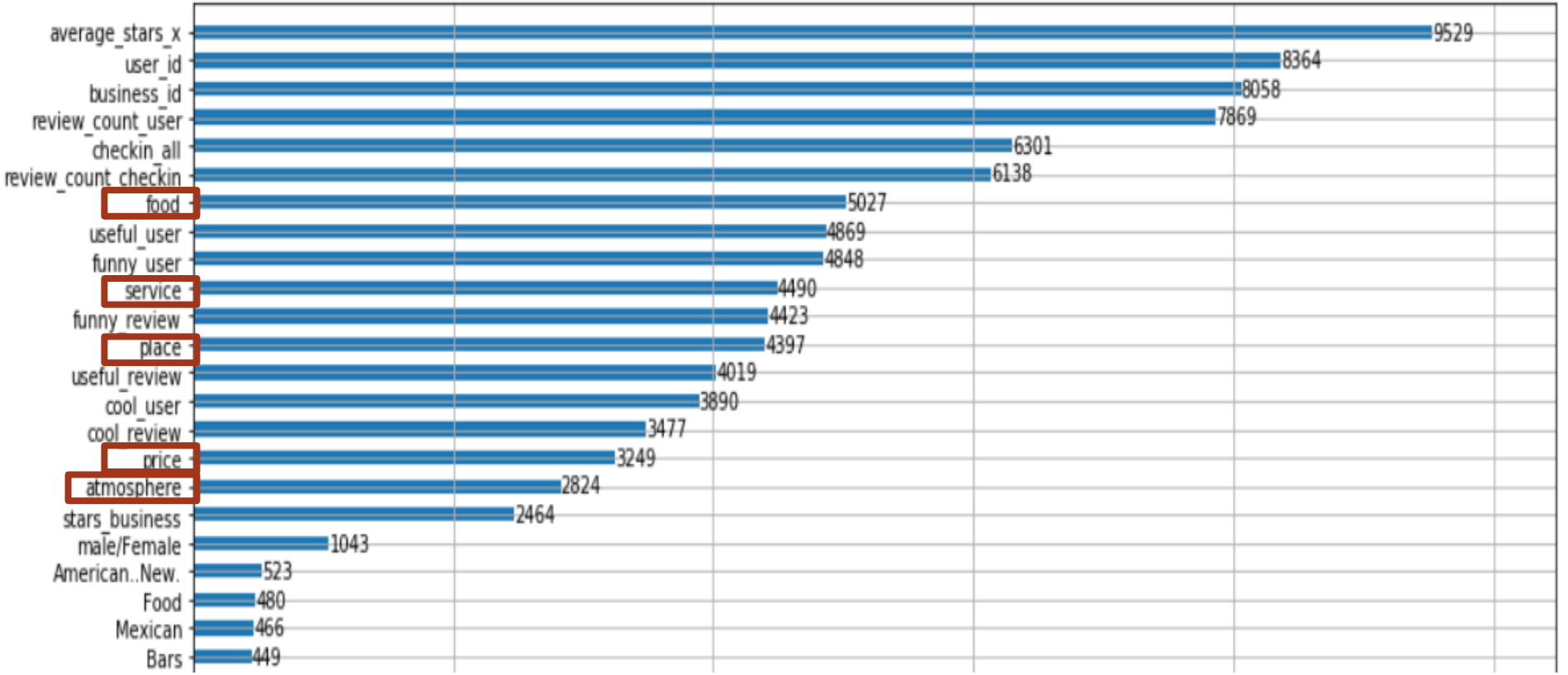


XGBOOST

Ludwig

協同過濾

特徵欄位重要程度



對評論做文本分析
建立五大特質

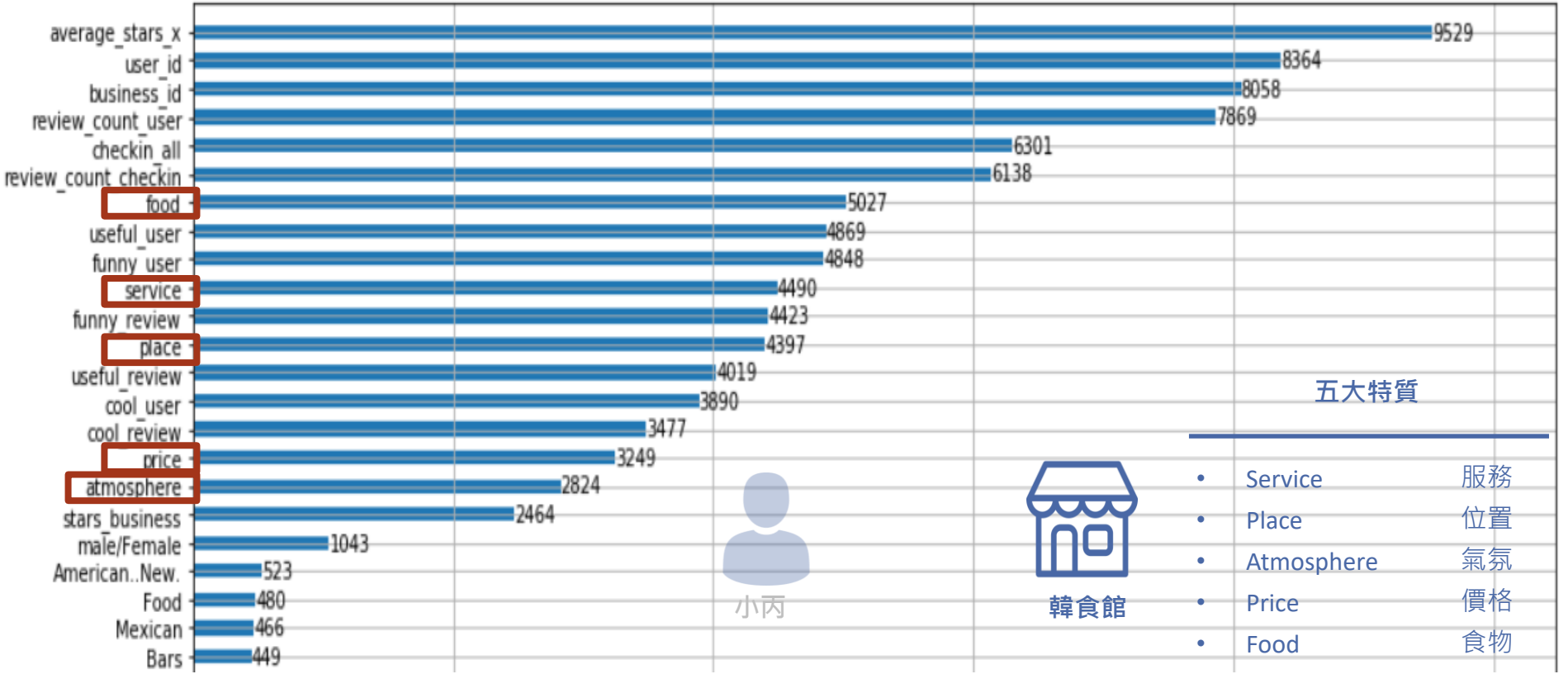
	business_id	atmosphere	food	service	price	place	overall
0	9yKzy9PApeiPPOUJEtnvkg	6.0	15.0	2.0	0.0	33.0	11.2
1	ZRJwVLyzEJq1VAihDhYiow	5.0	30.0	17.0	3.0	-1.0	10.8
2	6oRAC4uyJCsJl1X0WZpVSA	2.0	3.0	11.0	16.0	13.0	9.0
3	-yxfBYGB6SEqszmxJxd97A	-1.0	0.0	-1.0	0.0	0.0	-0.4
4	zp713qNhx8d9KCJJnrw1xA	5.0	43.0	6.0	10.0	19.0	16.6
5	wNUea3IXZWD63bbOQaOH-g	13.0	36.0	2.0	8.0	13.0	14.4
6	nMHhuYan8e3cONo3PornJA	9.0	44.0	6.0	6.0	19.0	16.8
7	e9nN4XxjdHj4qtKCOPq_vg	-2.0	24.0	-1.0	5.0	22.0	9.6
8	h53YuCilDfEFSJCQpk8v1g	2.0	6.0	0.0	0.0	7.0	3.0
9	yc5AH9H71xJidA_J2mChLA	8.0	8.0	10.0	4.0	13.0	8.6

XGBOOST

Ludwig

協同過濾

特徵欄位重要程度



小丙

韓食館

五大特質

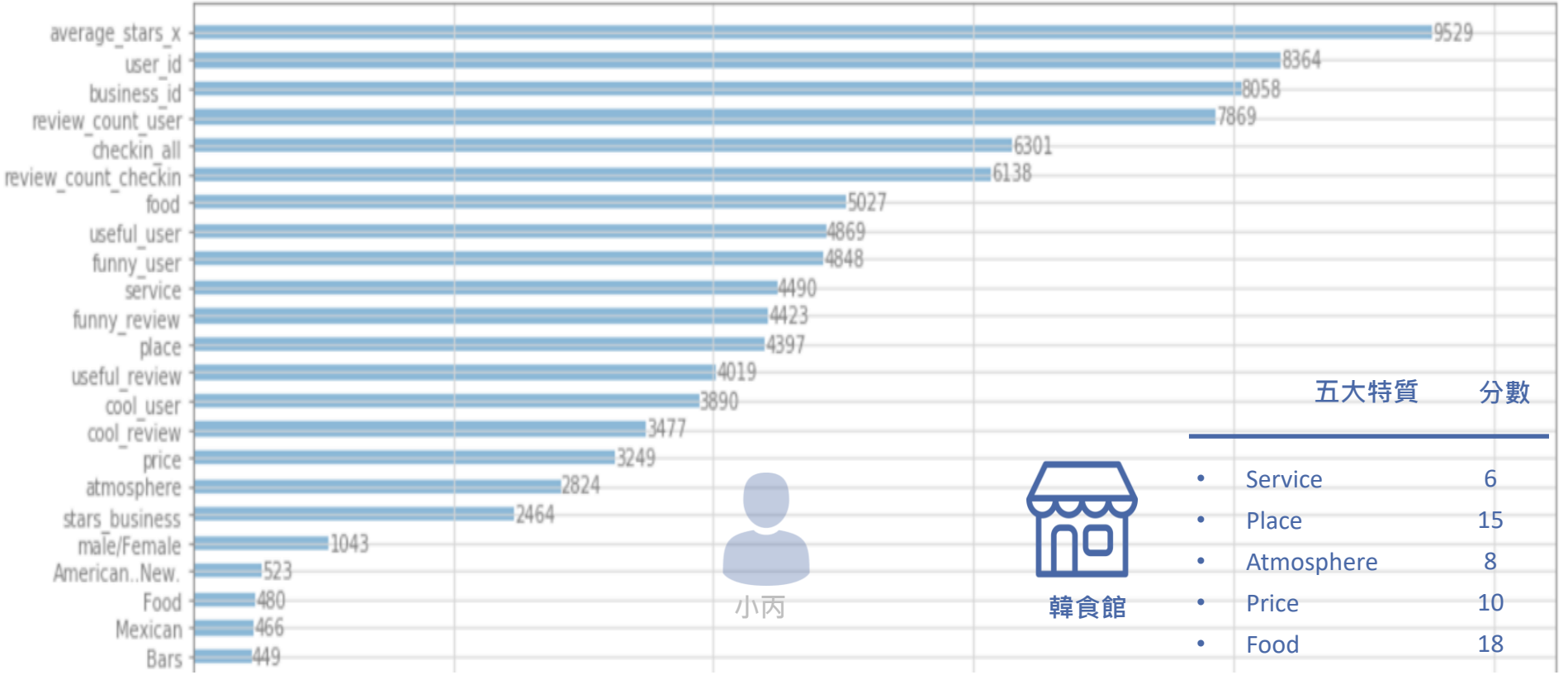
- Service 服務
- Place 位置
- Atmosphere 氣氛
- Price 價格
- Food 食物

XGBOOST

Ludwig

協同過濾

特徵欄位重要程度



小丙

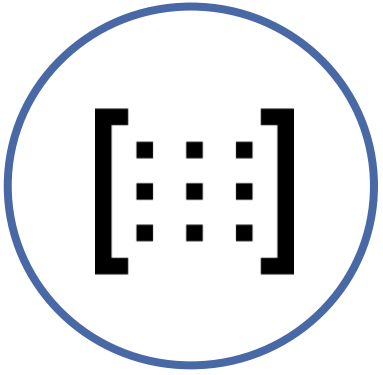
韓食館

五大特質	分數
• Service	6
• Place	15
• Atmosphere	8
• Price	10
• Food	18

XGBoost

LUDWIG

Collaborative Filtering



XGBOOST
Ludwig
協同過濾

相似度計算
評分星等預測

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	5	4	4	4
小乙	2	2	2	1	2
小丙	4	?	5	4	3

XGBOOST
Ludwig
協同過濾

相似度計算

評分星等預測

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	5	4	4	4
小乙	2	2	2	1	2
小丙	4	?	5	4	3

$$\text{甲丙相似度} = \frac{(3-4)\times(4-3)+(4-4)\times(5-3)+(4-4)\times(4-3)}{\sqrt{(3-4)^2+(4-4)^2+(4-4)^2}\times\sqrt{(4-3)^2+(5-3)^2+(4-3)^2}} = 0.42$$

乙丙相似度 = 0.31

XGBOOST
Ludwig
協同過濾

相似度計算

評分星等預測

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	5	4	4	4
小乙	2	2	2	1	2
小丙	4	?	5	4	**3**

甲丙相似度 0.42

乙丙相似度 0.31

甲丙相似度　甲對韓食館的評分 – 甲的評分平均　乙丙相似度　乙對韓食館的評分 – 乙的評分平均

$$\text{預測小丙對韓食館的評分} = \underbrace{3}_{\text{丙評分平均}} + \frac{\overbrace{0.42}^{\text{甲丙相似度}} \times (5-4) + \overbrace{0.31}^{\text{乙丙相似度}} \times (2-2)}{0.42 + 0.31}$$

XGBOOST
Ludwig
協同過濾

相似度計算

評分星等預測 =

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	5	4	4	4
小乙	2	2	2	1	2
小丙	4	?	5	4	3

甲丙相似度 0.42

乙丙相似度 0.31

甲丙相似度 甲對韓食館的評分 – 甲的評分平均 **乙丙相似度** 乙對韓食館的評分 – 乙的評分平均

$$\text{預測小丙對韓食館的評分} = \underbrace{3}_{\text{丙評分平均}} + \frac{\overbrace{0.42}^{\text{甲丙相似度}} \times (5-4) + \overbrace{0.31}^{\text{乙丙相似度}} \times (2-2)}{0.42 + 0.31}$$

XGBOOST
Ludwig
協同過濾

相似度計算

評分星等預測

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	**5**	4	4	**4**
小乙	2	**2**	2	1	**2**
小丙	4	?	5	4	3

甲丙相似度 0.42

乙丙相似度 0.31

甲丙相似度 **甲對韓食館的評分 – 甲的評分平均** 乙丙相似度 **乙對韓食館的評分 – 乙的評分平均**

$$\text{預測小丙對韓食館的評分} = \underbrace{3}_{\text{丙評分平均}} + \frac{\overbrace{0.42}^{\text{甲丙相似度}} \times (5-4) + \overbrace{0.31}^{\text{乙丙相似度}} \times (2-2)}{0.42 + 0.31}$$

XGBOOST
Ludwig
協同過濾

相似度計算

評分星等預測

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	5	4	4	4
小乙	2	2	2	1	2
小丙	4	**?**	5	4	3

甲丙相似度 0.42

乙丙相似度 0.31

甲丙相似度　甲對韓食館的評分 – 甲的評分平均　乙丙相似度　乙對韓食館的評分 – 乙的評分平均

$$\text{預測小丙對韓食館的評分} = \underbrace{3}_{\text{丙評分平均}} + \frac{0.42 \times (5-4) + 0.31 \times (2-2)}{\underbrace{0.42 + 0.31}_{\text{甲丙相似度}\quad\text{乙丙相似度}}} = \mathbf{3.65}$$

相似度計算

評分星等預測

	九州鬆餅	韓食館	樂麵屋	操場酒吧	平均
小甲	3	5	4	4	4
小乙	2	2	2	1	2
小丙	4	?	5	4	3

甲丙相似度 0.42

乙丙相似度 0.31

3.65

甲丙相似度　甲對韓食館的評分 – 甲的評分平均　乙丙相似度　乙對韓食館的評分 – 乙的評分平均

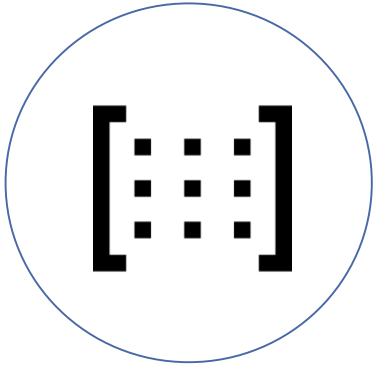
$$\text{預測小丙對韓食館的評分} = \underbrace{3}_{\text{丙評分平均}} + \frac{\overbrace{0.42}^{\text{甲丙相似度}} \times (5-4) + \overbrace{0.31}^{\text{乙丙相似度}} \times (2-2)}{0.42 + 0.31} = 3.65$$

$$RMSE = \sqrt{\frac{\Sigma(\text{預測星等} - \text{真實星等})^2}{\text{總數}}}$$

XGBoost

LUDWIG

Collaborative Filtering



$$RMSE = \sqrt{\frac{\Sigma(\text{預測星等} - \text{真實星等})^2}{\text{總數}}}$$

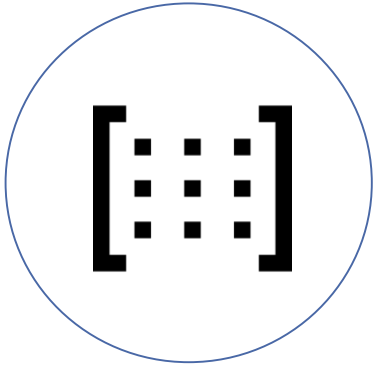
XGBoost

0.93

LUDWIG

0.95

Collaborative Filtering

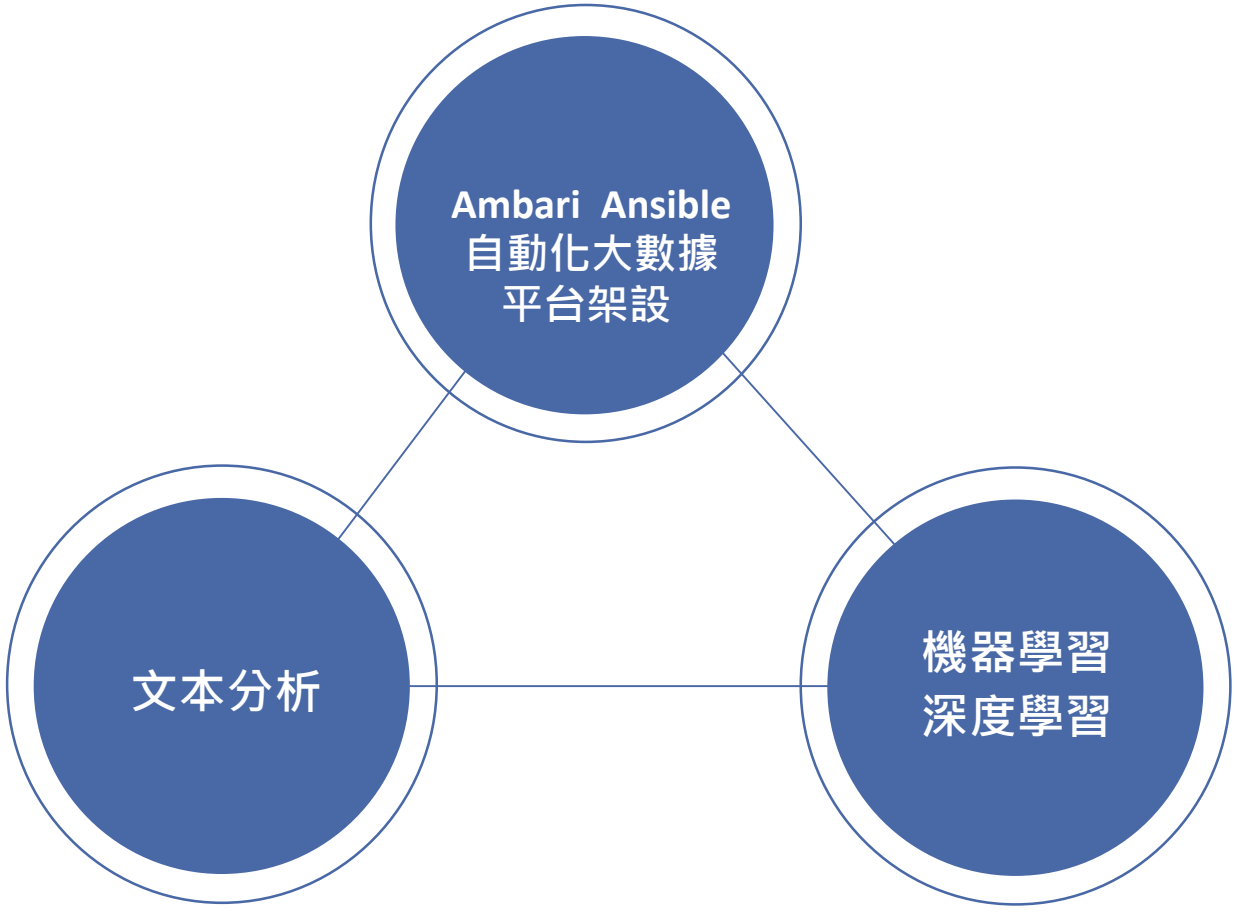


0.96

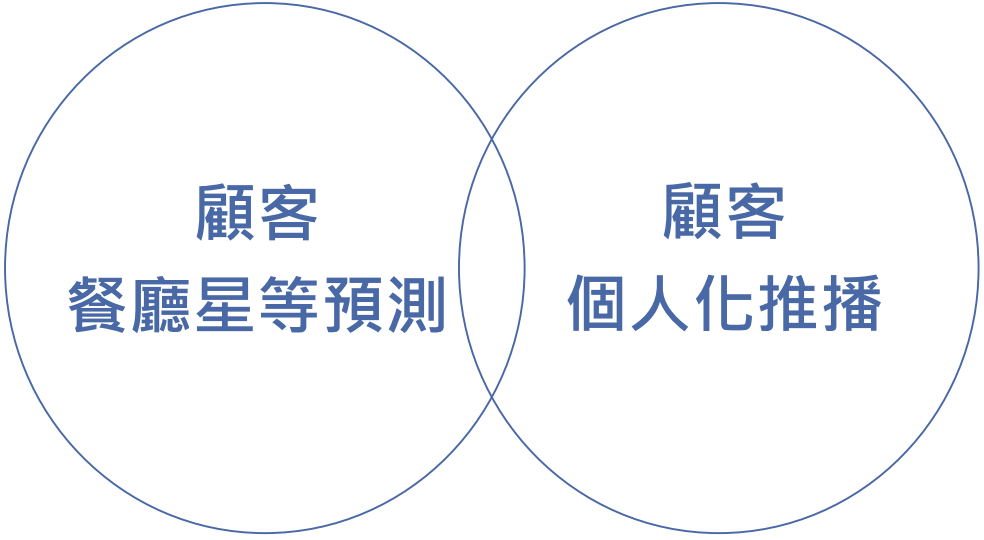
05

結論

吳岱凌



吳岱凌



Thank you for your listening !

如對簡報有任何疑問

歡迎寄信至 richiechao95@gmail.com