DinnerSelector

Web Retrieval and Mining - Spring 2019 - Final Term Project

R07922096 莊昕寰 R07922101 游子緒 R07922137 王韻華 R07922139 蕭皓仁

Introduction

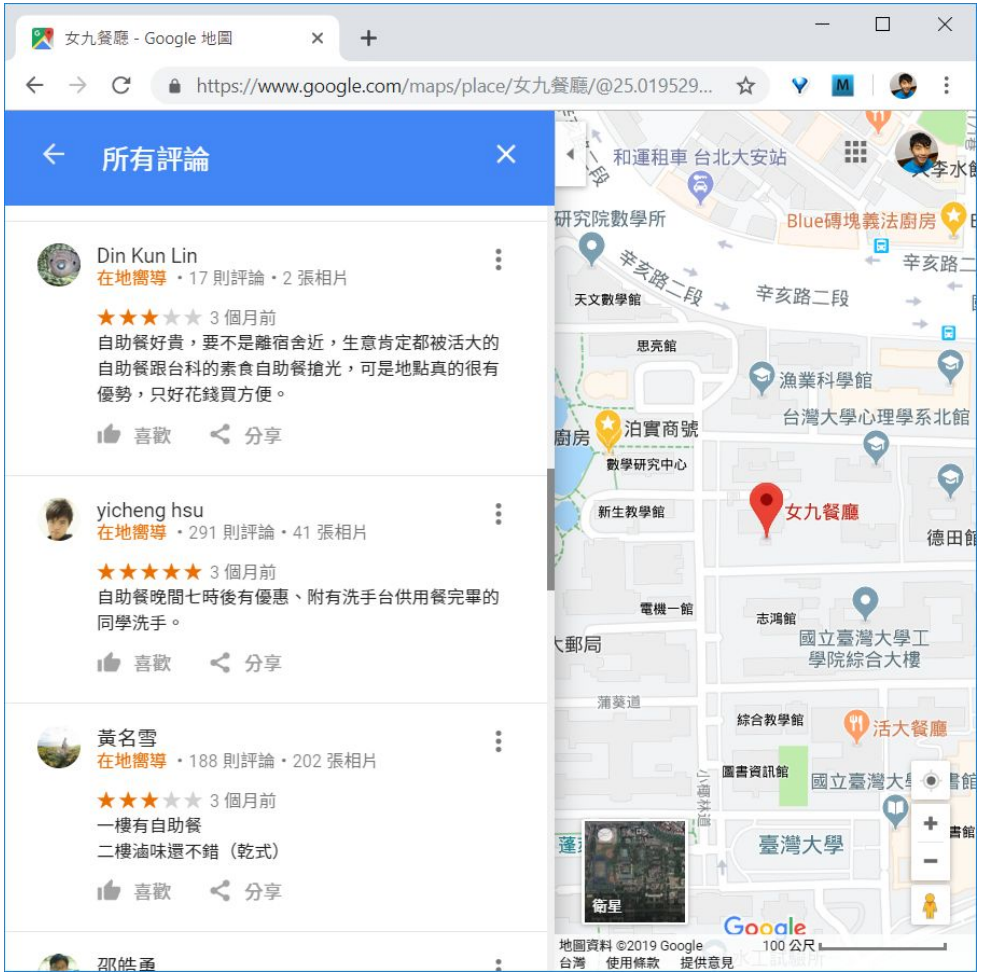
Motivation

The question we ask every day:

What should we have for dinner?

Data: Google Maps

- For the areas around NTU, Google Maps has the most reviews among all platforms
- $200 USD credit available in Google Places API each month
- Problem: Only 5 reviews per a place available
- Solution: Crawl the reviews from web pages

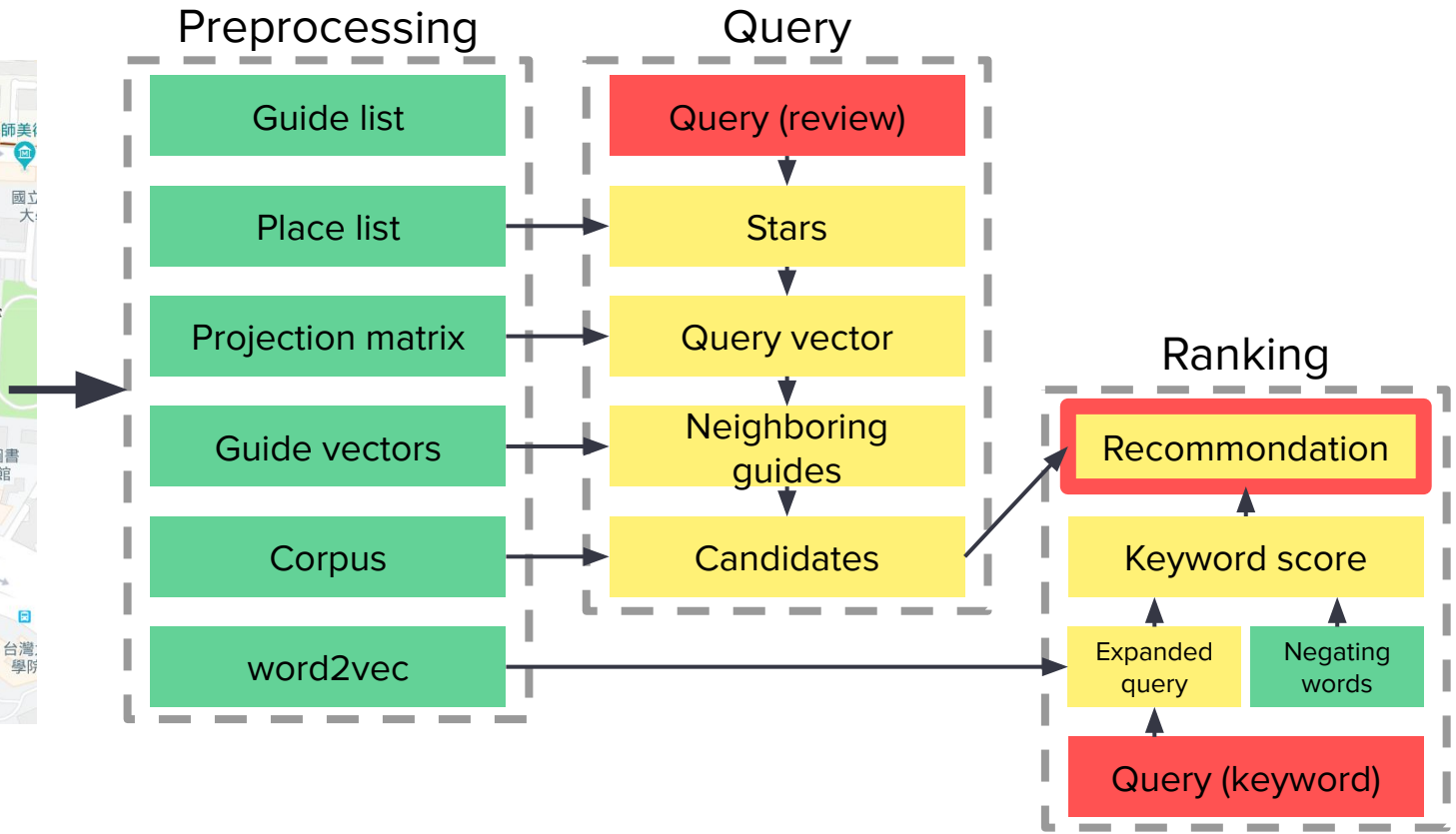


Data: Google Maps (cont.)

- In a review:
  - #star
  - comment
  - photos (not used)
- In web page of a place:
  - name
  - address
  - reviews with
    - reviewer IDs
    - review counts
- In web page of a reviewer:
  - reviewer ID
  - review count
  - reviews with
    - names
    - addresses

System Design

Area of interest



Details

Data collection

1. Pick a set R of 41 restaurants from Lane 118.
2. Collect reviewers from R.
3. Filter the reviewers by: (1) review count > 100, (2) reviewed > 5 restaurants in R. These reviewers will be called guides. **92 guides** were found.
4. Collect reviews from guides.
5. Filter the places by: (1) reviewed by more than 10 Guides.
   180 places were found.

Preprocessing

1. All the reviews of the 92 guides will be taken as the corpus.
2. The #star of each guide is used to construct a guide vector:
   a. Normalize #star so that μ=0 and σ=1.
   b. Assign the places not reviewed 0.
   c. Construct a 180-dimensional vector for each guide.
   d. **LSI**: Project the 180-dimensional vector into a **20-dimensional latent space**.
   e. Keep the projection matrix **P** for future use.

	restaurant #1	restaurant #2	restaurant #3
guide #1	0.9	-0.2	0.4
guide #2	1.1	0.8	-0.2

Query

The query consists of a list of reviews and some optional constraint. e.g.:

```
“reviews”:[
    {"place":"小川拉麵",                    "stars":5,    },
    {"place":"一品日式拉麵專門店 ",          "stars":5,    },
    {"place":"胖老爹美式炸雞大安店 ",        "stars":5,    },
    {"place":"雞二拉麵",                    "stars":5,    },
    {"place":"七里亭",                      "stars":1,    },
    {"place":"巧味快餐",                    “stars":1,    },...]
"constraints":[
    {"keyword":"便宜"},
    {"keyword":"飲料"},
    {"keyword":"衛生"}, ...]
```

Neighboring Guides in Latent Space

1. Project the query into the latent space by LSI projection matrix P
2. Calculate Euclidean distances between the query vector and all guide vectors
3. Select the 20 nearest guides as “Taste Neighbors”
4. Collect all the places reviewed by Taste Neighbors and count them.
5. Roughly Rank the restaurants with
   score = 0.1 * review count
   + average$_{\text{Taste Neighbors}}$(normalized #stars)
6. Select the 30 restaurants with the highest scores as candidates

Keyword Expansion

1. Preprocessing:
   Train a word2vec with a corpus collected from reviews to restaurants
2. Runtime:
   Search for similar words in the corpus with word embedding from word2vec

Keywords: 安靜	Keywords: 便宜	Keywords: 飲料
悠閒 0.9737377762794495	划算 0.8164801001548767	紅茶 0.8619219660758972
放鬆 0.9711542129516602	料好 0.8082225322723389	暢飲 0.8571416735649109
輕鬆 0.9618710279464722	佛心 0.8076516389846802	自助 0.8523671627044678
好友 0.9583859443664551	不貴 0.8062626123428345	飲品 0.8502625226974487
佈置 0.9571378231048584	實惠 0.7958357334136963	取用 0.850176215171814
燈光 0.9568617343902588	物超所 0.7875328660011292	湯品 0.8456183671951294
復古 0.9562684297561646	物美價廉 0.784941554069519	自取 0.8402163982391357
咖啡店 0.9555534720420837	平民 0.7785834074020386	無限 0.8372657299041748
約會 0.9539598226547241	公道 0.7776854038238525	附 0.8345868587493896
場所 0.9444922804832458	物價 0.7765038013458252	果汁 0.8340500593185425

Term Frequency & Negating Words

1. A restaurant is considered a document composed of all of its review contents.
2. Check if each word is a negating word (e.g.「不」,「很差」)
3. Multiply term frequency by -2 if the context is negating

	覺得	很	不	衛生
TF	1	1	1	1
reversed	1	-2	1	-2

"山西刀削麵"由於「價格還算實惠！小菜一盤25元也算便宜」中的"實惠"而加1分
"山西刀削麵"由於「價格實惠，口味稍重。」中的"實惠"而加1分
"山西刀削麵"由於「推薦道地的口感超Q的刀削麵，價實惠。」中的"實惠"而加1分
"山西刀削麵"由於「好吃，價格實惠」中的"實惠"而加1分
"台一牛奶大王"由於「口味一般，不便宜。都是緊鄰高級學府，建中旁的冰店也不差。（雖然是大學比中學😝）」中的"便宜"而加-2分

Ranking

1. Keyword expansion : “便宜” ➔ [“便宜”, “實惠”, “大碗”]
2. Calculate (inverted) term frequency for each keyword
3. Normalize term frequency with the total review count of each restaurant and scale with a constant: TF_norm(kwd, doc) = TF * 100 / review_count(doc)
4. Rank the candidates with:
   score(restaurant) = 0.1 * review count
   + 2 * average$_{\text{Taste Neighbors}}$(normalized #stars)
   + average$_{\text{keyword}}$(TF_norm) [if keywords are specified]
5. Output the ranked candidates

Evaluation

Query 1 (MVNLab)

{"place":"滇味小廚","stars":5},
{"place":"親來食堂","stars":5},
{"place":"雞二拉麵","stars":4},
{"place":"小川拉麵","stars":4},
{"place":"胖老爹美式炸雞大安店","stars":4},
{"place":"一品日式拉麵專門店","stars":1},
{"place":"巧味快餐","stars":1},
{"place":"炒飯仙人","stars":1}

Evaluation of Query 1 (MVNLab)

MAP = 43.2%, Precision = 36.7%

Match | Mismatch

1	春山茶水舖	呂 巷仔口米粉湯	松田日式飯糰	蘇草salvia	好食早餐	鼎泰豐 信義店
7	一極拌福州 乾拌麵	初牛 台北公館店	七里亭茶食館	蔣記家蘑麵	合益佳雞肉飯	湄賽雲泰料理
15	宮原眼科	姊姊的廚房	親來食堂	臺大黑飯糰	池先生咖哩屋	忠誠山東蔥油 餅
21	二八麵堂	樂食堂	樂業麵線	健康滷味	すき家Sukiya 公館店	香料廚房
27	小木屋鬆餅 (台大店)	大李水餃	貝菈小屋	阿玉水餃 (生水餃專賣店)	長興小舖	願有記台大店

Query 2 (huzixiao)

{"place":"淘客美式漢堡公館店","stars":5},
{"place":"好食早餐","stars":5},
{"place":"好好味港式菠蘿包","stars":5},
{"place":"詹記麻辣火鍋-敦南店","stars":4},
{"place":"台南阿輝炒鱔魚","stars":4},
{"place":"二八麵堂","stars":3},
{"place":"美美平價火鍋","stars":2},
{"place":"台越美食","stars":1},
{"place":"奇美博物館","stars":1}

Evaluation of Query 2 (huzixiao)

MAP = 46.1%, Precision = 56.7%

Match | Mismatch

1	春山茶水舖	松田日式飯糰	呂 巷仔口米粉湯	蘇草salvia	SUKIYA すき家 古亭店	好食早餐
7	健康滷味	七里亭茶食館	蔣記家蒯麵	湄賽雲泰料理	宮原眼科	合益佳雞肉飯
15	長興小舖	姊姊的廚房	忠誠山東蔥油 餅 - 此燈亮有餅	二八麵堂	池先生咖哩屋	臺大黑飯糰
21	鼎泰豐 信義店	樂業麵線	親來食堂	阿玉水餃 (生水餃專賣店)	すき家Sukiya 公館店	香料廚房
27	樂食堂	小木屋鬆餅 (台大店)	初牛 台北公館店	鳳城燒臘粵菜	蠶居	漢來海港餐廳- 敦化店

Evaluation of Negating words

“冷氣” Accuracy = 87.5%	Positive	Negative
Predicted positive	55 (85.9%)	7 (10.9%)
Predicted negative	1 (1.6%)	1 (1.6%)

“乾淨” Accuracy = 92.1%	Positive	Negative
Predicted positive	197 (86.4%)	14 (6.1%)
Predicted negative	4 (1.8%)	13 (5.7%)

“發票” Acc = 92.8%	Positive	Negative
Predicted positive	10 (71.4%)	1 (7.1%)
Predicted negative	0 (0.0%)	3 (21.4%)

Demo

```
66  keywords = ['便宜', '衛生', '飲料']
```

PROBLEMS 22 OUTPUT DEBUG CONSOLE TERMINAL

stars	count	kw	score	place
1.92	0.02	2.38	4.32	宮原眼科
2.1	0.02	1.71	3.83	健康滷味
2.06	0.02	1.64	3.72	七里亭茶食館
1.34	0.09	2.10	3.53	呂 巷仔口米粉湯
1.38	0.05	1.98	3.41	湄赛雲泰料理
2.28	0.02	0.87	3.17	SUKIYA すき家 古亭店
1.92	0.06	1.05	3.04	松田日式飯糰
1.16	0.03	1.77	2.96	蠶居
1.8	0.04	1.09	2.93	好食早餐
2.12	0.04	0.53	2.69	蘇草salvia
1.76	0.01	0.89	2.66	すき家Sukiya 公館店
1.78	0.02	0.62	2.42	臺大黑飯糰
1.72	0.01	0.64	2.37	初牛 台北公館店
1.6	0.03	0.71	2.34	二八麵堂
1.08	0.06	0.82	1.96	合益佳雞肉飯
0.94	0.05	0.96	1.95	小木屋鬆餅(台大店)
1.24	0.05	0.64	1.93	姊姊的廚房
1.32	0.11	0.48	1.91	春山茶水舖
1.46	0.04	0.34	1.84	長興小舖
1.54	0.03	0.27	1.84	樂業麵線
1.76	0.02	0.05	1.83	鼎泰豐 信義店
1.14	0.03	0.49	1.66	漢來海港餐廳-敦化店
0.74	0.06	0.67	1.47	樂食堂
0.62	0.08	0.73	1.43	忠誠山東蔥油餅 - 此燈亮有餅
0.6	0.08	0.51	1.19	池先生咖哩屋
0.66	0.06	0.39	1.11	鳳城燒臘粵菜
0.74	0.06	0.28	1.08	香料廚房
1.02	0.06	-0.1	0.96	蔣記家薾麵
0.4	0.08	0.29	0.77	阿玉水餃 (生水餃專賣店)
-0.0	0.11	0.43	0.54	親來食堂

```
66  keywords = ['高級', '約會']
```

PROBLEMS 22 OUTPUT DEBUG CONSOLE TERMINAL

stars	count	kw	score	place
2.28	0.02	0.02	2.32	SUKIYA すき家 古亭店
2.12	0.04	0.14	2.30	蘇草salvia
2.1	0.02	0.0	2.12	健康滷味
2.06	0.02	0.0	2.08	七里亭茶食館
1.92	0.02	0.11	2.05	宮原眼科
1.92	0.06	0.0	1.99	松田日式飯糰
1.8	0.04	0.04	1.88	好食早餐
1.76	0.02	0.07	1.85	鼎泰豐 信義店
1.78	0.02	0.0	1.8	臺大黑飯糰
1.76	0.01	0.01	1.78	すき家Sukiya 公館店
1.72	0.01	0.0	1.73	初牛 台北公館店
1.6	0.03	0.0	1.63	二八麵堂
1.54	0.03	0.01	1.58	樂業麵線
1.46	0.04	0.0	1.5	長興小舖
1.34	0.09	0.03	1.46	呂 巷仔口米粉湯
1.32	0.11	0.0	1.43	春山茶水舖
1.38	0.05	0.0	1.43	湄赛雲泰料理
1.24	0.05	0.0	1.29	姊姊的廚房
1.14	0.03	0.07	1.24	漢來海港餐廳-敦化店
1.16	0.03	0.0	1.19	蠶居
1.08	0.06	0.0	1.14	合益佳雞肉飯
1.02	0.06	0.03	1.12	蔣記家薾麵
0.74	0.06	0.21	1.01	香料廚房
0.94	0.05	0.0	0.99	小木屋鬆餅(台大店)
0.74	0.06	0.0	0.8	樂食堂
0.66	0.06	0.0	0.72	鳳城燒臘粵菜
0.62	0.08	0.0	0.7	忠誠山東蔥油餅 - 此燈亮有餅
0.6	0.08	0.0	0.67	池先生咖哩屋
0.4	0.08	0.0	0.48	阿玉水餃 (生水餃專賣店)
-0.0	0.11	0.0	0.11	親來食堂

Q&A