# Project 1: Hotel Review Opinion Mining

## NATURAL LANGUAGE PROCESSING, NTU CSIE, SPRING 2017

授課教師: 陳信希 教授

助教: 顏安孜、薛祐婷

{d04922005, r04922029}@ntu.edu.tw

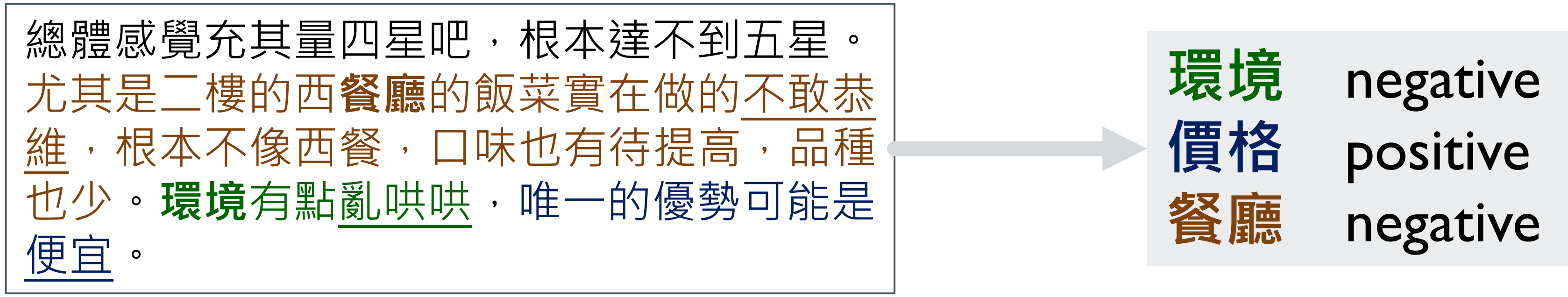
| Date | Schedule |
|---|---|
| 2017/04/06 | Project 1 release |
| 2017/04/20 | Midterm exam |
| 2017/05/05 | Kaggle submission due |
| 2017/05/07 | Report submission due |
| 2017/05/11 | In-class presentation (tentative) |

# OUTLINE

- Task Description
- Data
- Submission & Evaluation
- Grading & Rules
- Report & Presentation
- Possible Direction
- Tools & Resources

# TASK DESCRIPTION

- Can machine understand a review and summarize the reviewer's opinion on different aspects?



總體感覺充其量四星吧，根本達不到五星。尤其是二樓的西**餐廳**的飯菜實在做的不敢恭維，根本不像西餐，口味也有待提高，品種也少。**環境**有點亂哄哄，唯一的優勢可能是便宜。

| | |
|---|---|
| **環境** | negative |
| **價格** | positive |
| **餐廳** | negative |

# DATA

- List of provided data

1. Small review dataset with aspect-polarity labels
2. Large review dataset with polarity labels
3. NTUSD: sentiment dictionary
4. Aspect term examples

5. Test review dataset
6. Test questions
7. Sample submission

# DATA

- **`aspect_review.txt`** : small review dataset with aspect-polarity labels
  - 200 Chinese reviews
  - Each review is composed of 4 lines
    1. Review_id
    2. Content
    3. Positive aspects [separated by tab]
    4. Negative aspects [separated by tab]

```
...
32
這個賓館就在王府井步行街旁.地點相當好.價格也算優惠 ...
交通     價格
環境
33
比較老的飯店，房間等設施也比較老，服務還是可以的
服務
環境
...
```

- An aspect must be one of {服務, 環境, 價格, 交通, 餐廳}
  - But might not appear with exactly the same form in the review text
    E.g.“唯一的優勢可能是便宜” → aspect = **價格**

# DATA

- `polarity_review.txt`: large review dataset with only polarity labels
  - 258,003 reviews
  - Each line is a review, with two columns separated by a tab
    1. Label
       - `1`: positive
       - `-1`: negative
    2. Content

```
...
1		房間很大，早餐不錯。
-1		早餐太晚，不利於出行。
1		便宜且舒適
-1		入住時間不好
1		環境好！
...
```

- You can assume that **all statements** in a review are of the given **polarity** (label)
  - The source website requested users to separate positive and negative parts when they post a review

# DATA

- **`NTUSD_pos.txt`** / **`NTUSD_neg.txt`**: sentiment dictionary
  - List of opinion words that stand for positive/negative polarity
  - One word per line

  - The polarity of a word **might not apply to all possible context**
    - E.g. "更久" is in the list of *positive* opinion words
    - Be aware that there are **noises** in this dictionary, and use it carefully!

```
NTUSD_pos.txt

...
好景
好評
好開心
...
```

```
NTUSD_neg.txt

...
小衝突
小器
小題大作
...
```

# DATA

- **aspect_term.txt** : example terms that refer to certain aspect
  - Format:
    ASPECT<*tab*>TERM1<*space*>TERM2<*space*> ...
  - Use these terms as **seed** to find more useful terms

```
服務	態度 人員
環境	客房 設備 空調
價格	房價
交通	地理
餐廳	早餐
```

# DATA

- **test_review.txt** : test reviews to extract aspect-specific opinions
  - Same as aspect_review.txt but without the two list of aspects (same source website)
    1. **Review_id**
    2. Content
  - In the test questions, we use **Review_id** to specify a certain review
- **test.csv** : test questions
  1. **Id**: question id (1,737 questions in total)
  2. **Review_id**: id of the review to extract opinions
  3. Aspect = {服務, 環境, 價格, 交通, 餐廳}
     - For each review, we will ask the polarity of 3 aspects

```
...
277
裝修不錯，都是新的。服務非常好，很細心周到。 ...
278
位置尚可，但距離海邊的位置比預期的要差的多， ...
...
```

```
Id,Review_id,Aspect
...
9,276,價格
10,277,餐廳
11,277,交通
12,277,環境
13,278,交通
...
```

# DATA

```
test_review.txt

...
294
在南山區辦事的話住這裡還算方便，門口往哪個方向的
車都有，就是稍微有點吵啊
...
```

- **`sample_submission.csv`** : the format of submission
  - CSV Columns:
    1. **Id**: question id
    2. **Label**: the polarity of the specified aspect in the review
       - `1`: positive
       - `0`: not mentioned (or neutral)
       - `-1`: negative
- You must replace the values in the **Label** column with your answers!

```
test.csv

Id,Review_id,Aspect
...
58,294,價格
59,294,環境
60,294,交通
...
```

```
Id,Label
...
58,-2          → 0
59,-2          → -1
60,-2          → 1
...
```

# SUBMISSION & EVALUATION

- Submit your answers to the Kaggle platform
  - Join link: https://kaggle.com/join/ntunlp2017project1
    - Email to TA if you have problem joining the competition
  - A group may submit a maximum of 5 entries per day / select up to 2 final submissions for grading
  - Deadline: **2017/05/05 (Fri.) 23:59**
- Evaluation: $\text{accuracy} = \frac{\#\ \text{correct}}{\#\ \text{test questions}}$
  - Public score (50% test data): will update in real time during the project
  - Private score (50% test data): will be announced after the deadline

# GRADING & RULES

- Grading
  - Performance: 40%
    - Grade relatively
    - Mainly based on **private** score
  - Report: 30%
  - Presentation: 30%
  - Each group will be treated equally **regardless of # members**

- Rules
  - You can
    - Use any toolkit/library
    - Use external text corpus
  - You **CANNOT**
    - Find the answer somewhere on the web
    - Manually build aspect-term / opinion word dictionary<br>→ You need to propose an **automatic** approach
    - Answer the questions by yourself
  - **Contact the TAs if your are not sure!**

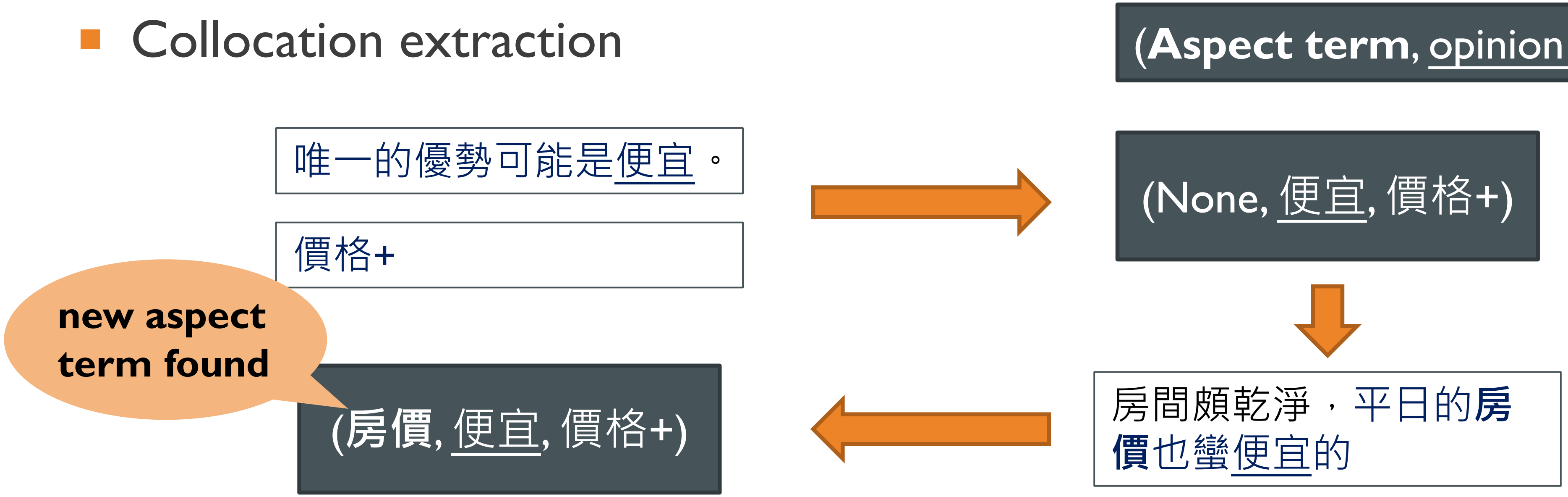
# REPORT & PRESENTATION

- Report
  - Language: Chinese or English (Be clear in meaning!)
  - Pages: **no more than 6** (with readable font size)
  - Must include:
    - Name and student ID of each group member
    - Kaggle team name & email address for registering Kaggle
    - Methodology
    - Experiments
    - Discussions
  - Deadline: **2017/05/07 (Sun.) 23:59**
    - Upload to CEIBA, one copy per group

- Presentation
  - Each group will have **5 minutes**
    - Be concise!
  - Date: 2017/05/11 (Thu.) in class
    - The next week after the report deadline
    - ***Subject to change***
  - Details will be announced later

# POSSIBLE DIRECTION

- Collocation extraction



(**Aspect term**, opinion word, aspect +/-)

唯一的優勢可能是便宜。

價格+

(None, 便宜, 價格+)

房間頗乾淨，平日的**房價**也蠻便宜的

(**房價**, 便宜, 價格+)

**new aspect term found**

  - Definition of **co-occurrence**: same review / same clause / within a distance of several words
  - Filter the set of possible aspect terms by POS tags
- You are encouraged to try other methods!

# TOOLS & RESOURCES

- Stanford CoreNLP http://stanfordnlp.github.io/CoreNLP/
  - Works better with simplified Chinese
  - Word segmentation / POS tagging / Parsing
- jieba中文分詞 https://github.com/fxsjy/jieba
- CKIP中文斷詞系統 http://ckipsvr.iis.sinica.edu.tw/

- NLTK collocation extraction
  - Document: http://www.nltk.org/howto/collocations.html

# QUESTIONS?