自然語言處理課程學期專題
Aspect-based Sentiment Analysis

April 24, 2014

Opinion Mining & Sentiment Analysis

我們的房間非常棒。飯店員工很不錯，而且總是會有會說英文的人可以服務我們。他們提供非常好的用餐建議，並確認是否有優良計程車司機可以為我們服務。地點很適合商務旅行，只要走一點路就可到達餐廳、銀行和服務業。飯店自助餐很不錯，咖啡館也是。總之，這是一個不錯的住宿經驗。

Aspects: 房間、員工、地點……

Opinion words: 棒、不錯……

專題說明

- Teamwork：三到四人一組
  - 登記組員：http://goo.gl/V3xEst
- 第一階段：給定一組含有意見標記的評論文章，建立兩個辭典
  - Aspects
  - Opinion words
- 第二階段：給定一組不含意見標記的評論文章，標記每篇文章
  - 各個Aspect的極性（正、負、中性）
  - 整篇文章全體的極性（正、負）

資料：旅館評論

- 位置離我們單位很近,從價格來說,性價比很高.我要的大床房,168元,前台服務員態度很好,房間硬件一般,但是想>想價格也就這樣了.還算乾淨,就是床墊子太硬.BTW:沒有騷擾電話,這個很好補充點評 2008年3月4日 ：對了,這裡>能夠免費上網,剛好上網加班一晚,很不錯

第一階段

- 提供三千筆旅館評論。
  - 1500筆正面評論（標記為1）
  - 1500筆負面評論（標記為2）
  - 中文繁體，UTF-8編碼
- 輸出資料
  - 最多100個不重覆的Aspects，按重要性排序
  - 最多500個不重覆的Opinion Words，按重要性排序
  - 中文繁體，UTF-8編碼

輸入資料格式

- 檔名：*hotel_training.txt*
- 一篇評論有兩行
  - 整體評價：正面（1）、負面（2）
  - 文章ID | 評論主文

輸入資料範例

1

18|很不錯的酒店，儘管在火車站，但很安靜。房間乾淨，早餐一般。

1

32|地點很方便，房間很舒服，服務也很好，就是價格不便宜啊！

2

3477|房間超爛，看上去很破很舊，而且房間內還有一股長時間不開窗子的那種霉味，被子也很不舒服，而且也有味道，燒水的壺超級髒，根本沒法用，房間也不提供收費的礦泉水，只能叫服務台的人給送，可人家竟然說要先付款>，然後給我出去買，這是什麼酒店啊，總之一個字——爛！勸大家最好別去！

輸出資料範例

aspect_{#team_id}.txt
地點 # top 1
房間 # top 2
床墊

opinion_{#team_id}.txt
不錯 # top 1
很好 # top 2
棒
髒

寄到 hhhuang@nlg.csie.ntu.edu.tw

第一階段評估

- 根據全班所有組別上傳的aspects與opinion words，建立兩個共識最高的辭典。
  - Aspect dictionary
  - Opinion dictionary
- 人工審查每一組上傳的aspects與opinion words，得到ground-truth。
- 計算各組結果與ground-truth的距離，選出表現最佳的六組，於5/29日上課時間報告作法。

第二階段

- 輸入一篇旅館評論，預測：
  - 提到了哪些aspect，對這些aspect的意見
  - 評論的整體評價：正（1）、負（2）
- 提供第一階段全班所得之 aspect dictionary與opinion dictionary。
- 1000筆測試資料

測試資料輸入格式

- 檔名：*hotel_test.txt*
- 一行一篇評論
- 文章ID|評論主文
- 中文繁體UTF-8，無BOM。

測試資料範例

18|很不錯的酒店，儘管在火車站，但很安靜。房間乾淨，早餐一般。

32|地點很方便，房間很舒服，服務也很好，就是價格不便宜啊！

3477|房間超爛，看上去很破很舊，而且房間內還有一股長時間不開窗子的那種霉味，被子也很不舒服，而且也有味道，燒水的壺超級髒，根本沒法用，房間也不提供收費的礦泉水，只能叫服務台的人給送，可人家竟然說要先付款>，然後給我出去買，這是什麼酒店啊，總之一個字——爛！勸大家最好別去！

輸出範例格式

- *hotel_test_{#team_id}.out*
- 一篇文章有四行輸出
  - 文章的ID
  - 正面aspects，tab分隔（無則留一行空白）
  - 負面aspects，tab分隔（無則留一行空白）
  - 總體評價 (1 正面、2 負面）
- 中文繁體，UTF-8編碼，無BOM。

輸出範例

```
18                # 文章的 ID
隔音 清潔          # 內文中不一定會有完全相同的字樣，tab 分隔
                  # 無則留一行空白
1                 # 總體評價，1為正面、2為負面。
32
地點 房間 服務
價格
1
3477

房間 棉被
2
```

第二階段成果繳交

- 輸出檔 *hotel_test_{#team_id}.out*
- 專題報告
  - 構想
  - 方法
  - 實驗設計
  - 效能評估
  - 心得
  - 需涵蓋兩階的專題內容
- 系統原始碼
- 寄到 hhhuang@nlg.csie.ntu.edu.tw

時程

時間	事項
2014-04-24	公布題目與第一階段資料
2014-04-30 23:59	登記組別資料 http://goo.gl/V3xEst
2014-05-15 23:59	繳交第一階段成果
2014-05-17	公布Aspect字典與Opinion字典
2014-05-29 上課時間	專題報告
2014-06-02 12:00	公布第二階段測試資料
2014-06-08 23:59	繳交第二階段成果與專題報告
2014-06-19	期末考

相關工具與資源

- 簡繁轉換
  - iconv
- 中文斷詞
  - Stanford Chinese Word Segmenter
    http://nlp.stanford.edu/software/segmenter.shtml
- 詞性標記
  - Stanford POS Tagger
    http://nlp.stanford.edu/software/tagger.shtml
- 句子剖析
  - Stanford Parser
    http://nlp.stanford.edu/software/lex-parser.shtml
- NTUSD（台大情緒辭典）
  - https://docs.google.com/spreadsheet/pub?key=0Ar4-24sx1v8GdDFpUl9WTVFONTBpU0YxYTVvcldjX1E&output=csv