Amazon Reviewers' <mark>Aspect</mark> Detection

鄒雅雯 林明廣

Agenda

研究動機

文獻回顧

研究方法

研究結果

研究動機

研究動機

VS

從評論中找解答

心好累

除了使用情緒分析知道商品好壞之外，

想要更細節了解產品的==使用心得==*以及*==介紹==

"

2 **文獻回顧**

Aspect Detection

Aspect Detection

- Aspect-based opinion mining：the extraction of aspects (or product features) from opinionated text and on the inference of polarity values associated with these

- Two types of aspects：Explicit and Implicit

    - Explicit：concepts that explicitly denote targets in the opinionated sentence.

      ex. I love the touchscreen of my phone but the battery life is so short

    - Implicit: an aspect be expressed indirectly through an implicit aspect clue (IAC)

      ex. This camera is sleek and very affordable → *appearance* and *price*

Aspect Detection

- **Unsupervised Learning**

  1. A Rule-Based Approach:

     - 透過建立 Dependency parse tree，抓取潛在 IAC

     - 使用 Camera Corpus，當中包含 IAC 以及其所對應的類別，並使用 WordNet 取得 IAC 的相似/反義詞，擴展每個類別中的 IAC

  → 只限於 Camera 的評論，且 WordNet 抓出很多的同義詞，容易有一詞多義的問題 ex. Light

- **Supervised Learning**

  1. 使用 SemEval 上的 datasets 進行訓練
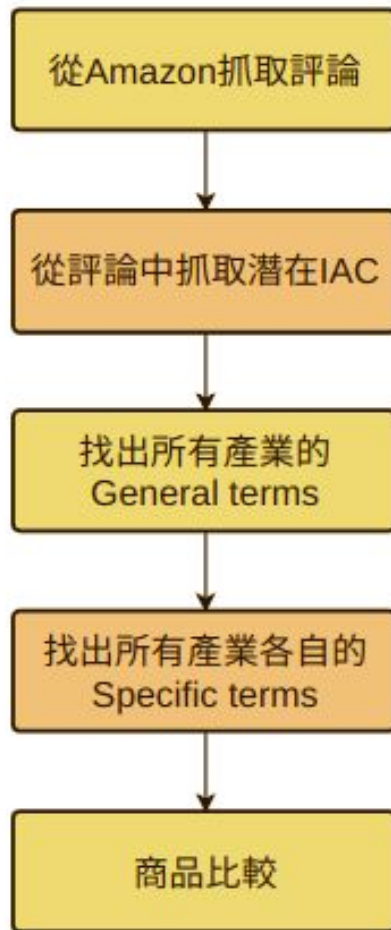
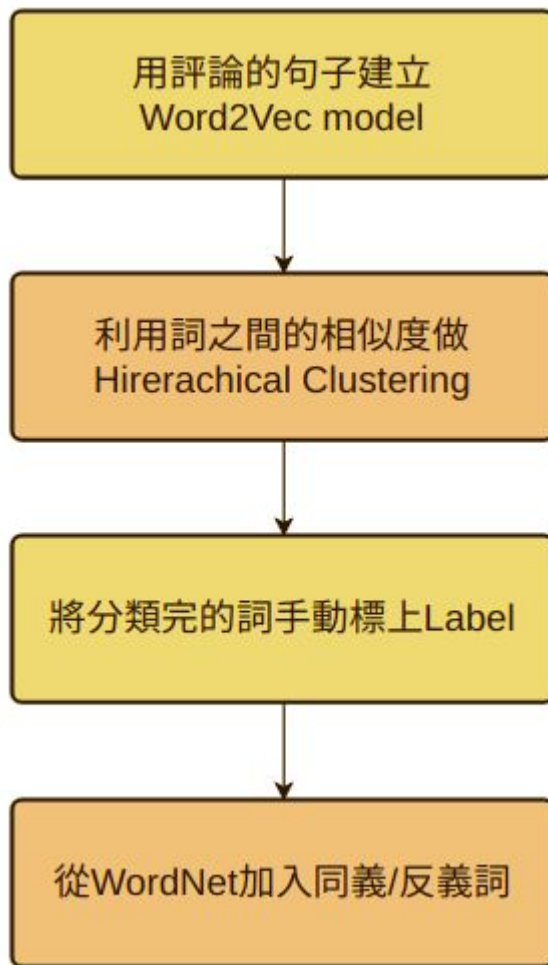 → 受限於 Training Data corpus 的產業，無法適用於不同 產業，實際應用上有限制

3 研究方法

系統流程圖

系統流程圖

📌 General Terms

用評論的句子建立
Word2Vec model

↓

利用詞之間的相似度做
Hirerachical Clustering

↓

將分類完的詞手動標上Label

↓

從WordNet加入同義/反義詞

📌 Specific Terms

用MI, LIFT 選出各產業的
Specific Terms

↓

用先前建好的Word2Vec model
做Hirerachical Clustering

↓

手動Label其中6個產業
並自動Label剩下的IAC

↓

自動擴展到剩下的產業
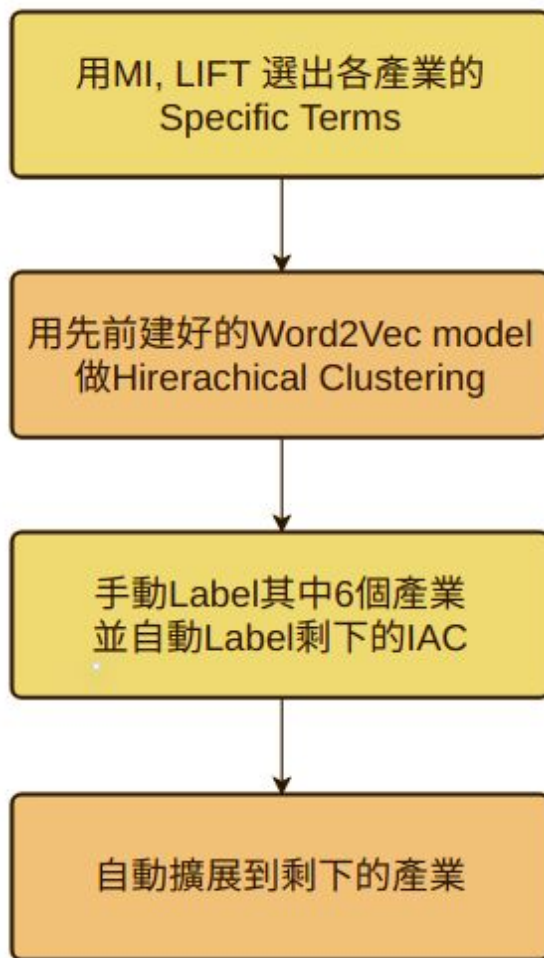
3-1 研究方法

資料來源

Amazon 評論資料

- 針對3C家電

- TreeMall中有交易的3C家電產品，共35項

- 使用Octoparse爬蟲，每項產品中的各商品至多爬200則評論共
  有232,849則評論

萬用鍋	豆漿機	電風扇	烤箱	冰箱	熱水瓶	氣炸鍋	咖啡機	刮鬍刀	冷氣	除濕機	清淨機
烘被機	電子鍋	果汁機	熨斗	吹風機	燈具	吸塵器	麵包機	微波爐	電視	電磁爐	洗衣機
檯燈	電暖器	調理棒	清靜機配件	電動牙刷	吸塵器配件	美容家電	電視配件	壓力鍋	料理機	捕蚊燈	

Data Format

Name	Brand	Price	Title	Score	Time	Text	Product
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Perfect for my 20	5.0 out of 5 stars	22-Jun-17	?Great little unit.	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Great AC	5.0 out of 5 stars	6-Apr-17	Took only about	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Gets the job don	4.0 out of 5 stars	22-May-17	This is about as l	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Nice air conditio	4.0 out of 5 stars	23-Jun-17	Nice air conditio	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Great little a/c - y	5.0 out of 5 stars	14-Aug-16	We bought this	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Frigidaire 5000 B	5.0 out of 5 stars	5-Jun-17	This 5000 b.t.u.	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	I must say these	5.0 out of 5 stars	13-Jul-16	This is my review	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	You'll thank your	5.0 out of 5 stars	30-Jun-17	Awesome little n	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Easy install, cools	5.0 out of 5 stars	19-May-18	Easy installation.	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	May seem small,	5.0 out of 5 stars	13-Jul-18	When I opened	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Light enough for	1.0 out of 5 stars	5-Jul-17	UPDATE: The co	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	I'd pass on this u	2.0 out of 5 stars	21-Jun-18	Actually, this is a	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	It's a decent air c	3.0 out of 5 stars	18-Jun-18	I have mixed rev	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	This taught me n	1.0 out of 5 stars	7-Aug-18	From what I can	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Tough little A/C!	5.0 out of 5 stars	27-May-18	To start, before I	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Great for the pric	5.0 out of 5 stars	25-May-18	It's beautiful and	AirConditioner

3-2 研究方法

抓出潜在IAC

抓出潛在IAC

◉ 資料前處理：使用StanfordCoreNLP

  ○ 將所有評論切成句子

  ○ 句子建立Dependency parse tree(https://corenlp.run/)

  ○ Lemmatization：將所有詞彙還原成字根

◉ 文法規則：共有 13 條

  ○ Subject Noun Rule

  ○ No Subject Noun Relation

  ○ Additional Rules

抓出潛在IAC

- **Subject Noun Rule**:

Trigger：若 active token h 為句子的主詞

  - If the sentence does not have auxiliary verb (be動詞/助動詞) , i.e., is, was, would, should, could:
    1. if the verb t is modified by an adjective or an adverb or it is in adverbial clause modifier relation(修飾詞) with another token, then both h and t are extracted as aspects.

       ex. The <u>battery</u> <u>lasts</u> little.

    2. if t has any direct object relation with a token n and the POS of the token n is Noun, then the token n extracted as aspect term.

       ex. I like the <u>beauty</u> of the screen.

  - A copula is the relation between the complement of a copular verb and the copular verb.(連繫動詞)

    ex. The car is <u>expensive</u>.

抓出潜在IAC

- **No Subject Noun Relation**:
  - If an adjective or adverb h is in infinitival or **open clausal complement** relation with a token t, then h is extracted as an aspect.

    ex. Very <u>big</u> to hold.
  - If a token h is connected to a noun t using a **prepositional relation**, then both h and t are extracted as aspects.

    ex. Love the <u>sleekness</u> of the <u>player</u>.
  - If a token h is in a **direct object relation** with a token t, t is extracted as aspect.

    ex. Not to mention the <u>price</u> of the phone.

抓出潜在IAC

- Additional Rules：
  - For each aspect term extracted above, if an aspect term h is in co-ordination or conjunct relation with another token t, then t is also extracted as an aspect.

    ex. The camera is amazing and easy to use.
  - If t is extracted as an aspect and t has noun compound modifier h, then the aspect h-t is extracted and t is removed from the aspect list.

    ex. We ordered the chicken casserole, but what we got were a few small pieces of chicken, all dark meat and on the bone.

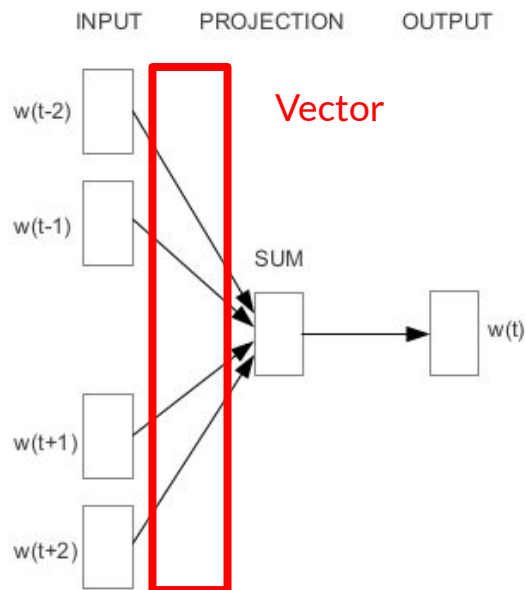## 3-3 研究方法

將相似概念的詞聚類在一起
— General term

Word Representation

- Bag of Words：sparse vector in the size of the vocabulary

  - One-Hot Vector：number of counts or TF-IDF
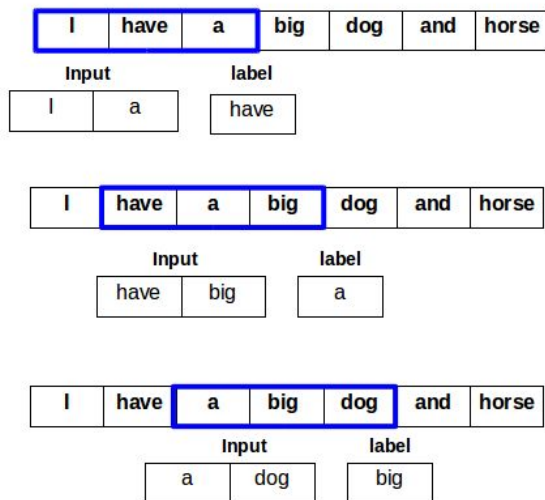
  car $[0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ \ldots\ 0]$

  - Terms are Orthogonal (正交/獨立) → 找不出 implicit word 共同的概念

- Continuous Bag of Words(CBOW)：Word2Vec

  - 考慮上下文關係

  - Dense vectors：Word Embedding

Word2Vec (CBOW)
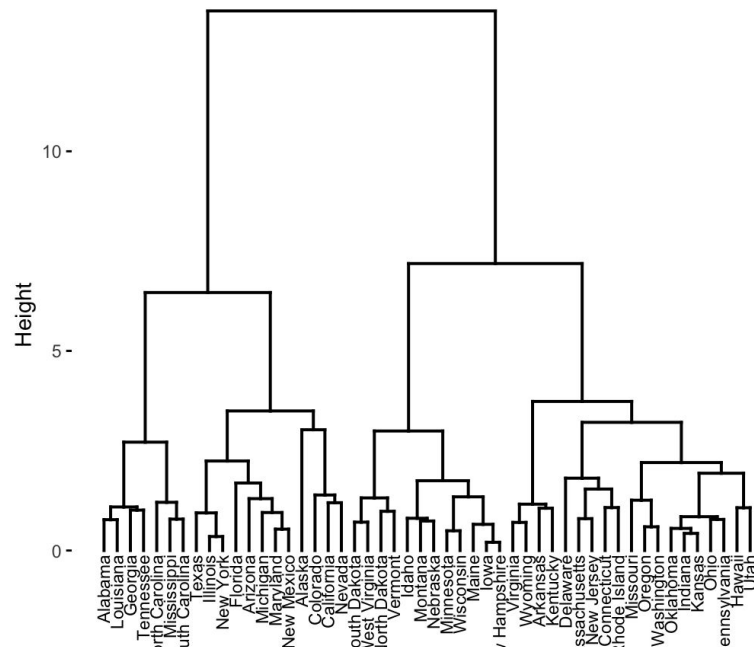

抓出共同概念的詞

Hierarchical Clustering

- ⦿ 目的：將相似概念的詞聚類成一群，並給定類別標籤

- ⦿ 概念：越相似的詞會先聚類起來

- ⦿ 連結方式：**Ward's Minimum Variance**

  - ○ 最小化各群聚內的變異加總

- ⦿ Threshold = 0.6

Hierarchical Clustering

Labeling

Performance
使用後效果

Appearance
外觀/材質

Quality
品質

Functionality
功能

Size
大小/重量

Feeling
使用者體驗

Service
服務

Price
價格

📌 Labeling

```
size:           ['numerous', 'shallow', 'cumbersome', 'max', 'bulky']

performance:    ['disadvantage', 'con', 'greatly', 'sore', 'drastically']

appearance:     ['linen', 'wiry', 'neutral', 'navy', 'colored']

quality:        ['optimum', 'wobbly', 'inaudible', 'disadvantage', 'con']

functionality:  ['hot', 'functionality', 'portability', 'nimble', 'maneuverable']

price:          ['economically', 'worthless', 'cheaply', 'expensively', 'modestly']

feeling:        ['displeased', 'curious', 'marvelously', 'amazingly', 'disappointed']

service:        ['embarrassment', 'prompt', 'shipment', 'ticket', 'apologetic']
```

擴展General Term

- **WordNet**：由普林斯頓大學認識科學實驗室在心理學教授喬治·A·米勒建立英語字典

  - 每一個具有相同意義的字條組稱為一個 synset（同義詞集合）

  - 只包含第一組的同義字與反義字
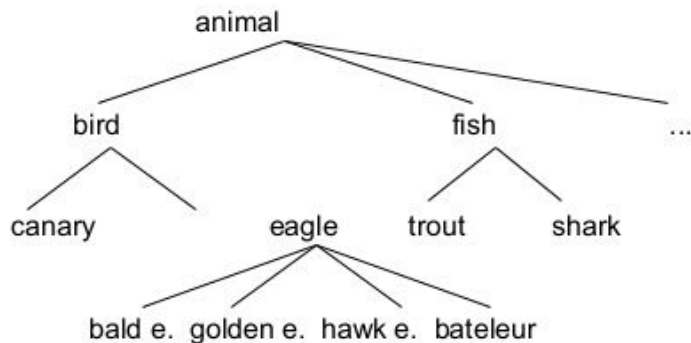

```
  1   wn.synsets('small')
```

```
[Synset('small.n.01'),
 Synset('small.n.02'),
 Synset('small.a.01'),
 Synset('minor.s.10'),
 Synset('little.s.03'),
 Synset('small.s.04'),
 Synset('humble.s.01'),
 Synset('little.s.07'),
 Synset('little.s.05'),
 Synset('small.s.08'),
 Synset('modest.s.02'),
 Synset('belittled.s.01'),
 Synset('small.r.01')]
```

研究方法

3-4

將相似概念的詞聚類在一起
— Specific term

Feature Selection

- Mutual Information(MI) * TF-IDF

$$MI = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N} \frac{f(y)}{N}}$$

x : term t 出現的評論數(DF) ex. light

y : 某產品的評論數(DF) ex. Table Lamp

- LIFT

$$Lift = \frac{\frac{f(x,y)}{f(y)}}{\frac{f(x)}{N}}$$

x : term t 出現的評論數(DF) ex. light

y : 某產品的評論數(DF) ex. Table Lamp

Feature Selection

計算各詞彙在
各產品的
MI*TFIDF & Lift

→

分別對
MI*TFIDF &
Lift進行排序
轉換成名次

→

將MI*TFIDF &
Lift的名次平均
取前500個詞彙

Our process is easy

Word2Vec → **Hierarchical Clustering** → **Labeling**

📌 Labeling

- 先人工標6個產品，剩餘29個產品採自動標記的方式

  - ['TableLamp', 'HairDryer', 'Refrigerator',

    'ElectronicPot', 'InstantPot', 'MicrodermabrasionDevices']

- 興增一個類別 — **Target**

  - 產品使用對象、地點、部位等

```
1  dictionary_specific['target']
```

```
['desktop',
 'nightstand',
 'keyboard',
 'laptop',
 'monitor',
 'computer',
 'granddaughter',
 'teen',
 'vanity',
 'table',
 'headboard',
 'desk',
 'newborn',
 'office']
```

Automatic labeling

◉ 作法：將要標類別的詞彙與已經標好的類別裡的每個詞分別比較相似度，相似度若大於 threshold (0.6) 則該類別 count+1，回傳 count 最大的類別，若 count > 4，則此詞彙歸納為該類別

Size

Price

Feeling

Term t → Return max(count) → If count > 4

Similarity >= 0.6 → count += 1
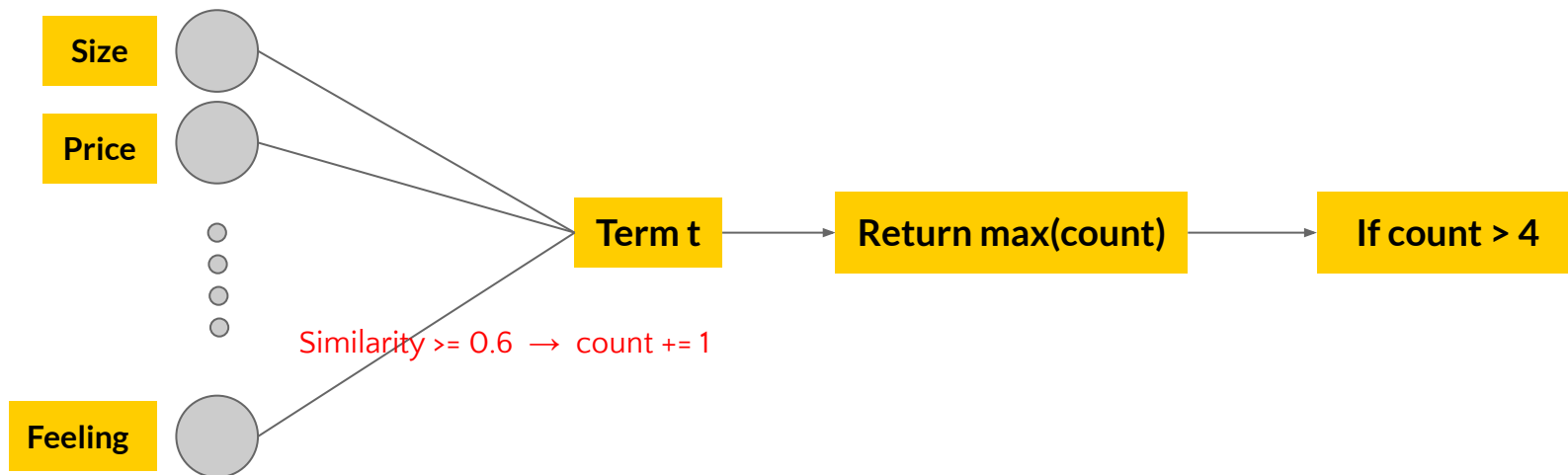
4 研究成果

比較兩商品

以 Table Lamp 為例

Appearance

term	sent	tf
nice	0.463028	23
color	0.361980	14
compact	0.297411	10
sleek	0.548160	7
modern	0.406475	6
white	0.384217	6
pink	0.483150	5
stylish	0.624900	3
yellow	0.000000	2
classy	0.796400	2

term	sent	tf
nice	0.448450	8
white	0.187800	6
sleek	0.000000	4
stylish	0.279733	4
black	-0.318200	4
metal	0.290600	3
modern	0.359100	3
attractive	0.466550	3
neutral	0.177700	3
color	0.028800	3

Functionality

term	sent	tf
light	0.286137	186
lamp	0.283986	150
battery	0.071275	57
usb	-0.022062	32
charge	0.094052	31
lighting	0.340652	26
led	0.238518	19
brightness	0.524384	16
charged	0.062785	12
warm	0.487582	10

term	sent	tf
lamp	0.143445	139
light	0.162377	76
bulb	-0.013254	48
halogen	0.209246	27
arm	0.118659	15
led	-0.095033	15
brightness	0.488933	10
hot	0.116250	7
illumination	0.175233	6
adjust	0.180959	5

以 Shaver 為例

Performance

Price

term	sent	tf
irritation	-0.315360	43
closer	0.036279	29
burn	-0.051939	19
excellent	0.588387	16
irritated	-0.327886	14
faster	0.155431	13
hot	0.000942	8
downside	-0.177078	7
warm	0.309000	6
irritate	-0.242955	6

term	sent	tf
excellent	0.546725	19
closer	-0.106223	14
irritation	-0.234507	12
hot	0.163600	10
burn	-0.001571	9
faster	0.082829	8
outstanding	0.612400	4
downside	-0.073650	2
quicker	-0.148000	2
burning	-0.226300	1

term	sent	tf
expensive	0.133291	31
cheap	0.006537	11
inexpensive	0.146800	2
afford	0.261225	2
affordable	0.091067	2
worthless	-0.129800	1
economically	0.000000	0
expensively	0.000000	0
economical	0.000000	0
valuable	0.000000	0

term	sent	tf
expensive	0.132937	17
cheap	0.038444	4
afford	0.329850	2
worthless	-0.440400	2
costly	-0.102700	1
economically	0.000000	0
inexpensive	0.000000	0
expensively	0.000000	0
economical	0.000000	0
valuable	0.000000	0

以 Shaver 為例

['Replacement blades only got more and more expensive.',
 "I have never bought one of their expensive models until now and don't regret the extra money.",
 "But they're not expensive.",
 'Which is a little short-lived for such an expensive shaver.',
 'This was a great buy -- expensive.',
 'I thought it was a bit expensive.',
 'What an expensive rip-off !!',
 'This one is overpriced and uses expensive refill cartridges for cleaning.',
 'Very expensive and so are all replacement parts .',
 'Just seems expensive.',
 'Too expensive for such a quality shaving.',
 'No razor burn) that I am now going to switch back to electric shaving after years of buying expensive straight razor refills.',
 'And I use expensive brand name straight razors.',
 "9700 Shaver I wondered what could possibly be so much better in this shaver that made it 3 or 4 times more expensive than other good quality shavers I've owned (including others from Norelco).",
 'The only reasons the 9700 was more expensive were that the Smart Clean is "Pro" and there is a digital readout on the handle.',
 '-It is very expensive for not much improvement over the 9300 series shaver.']

Expensive

['This is definitely not the cheapest configuration of this razor.',
 'Replacement heads are just the blades and screens but easy enough to replace and actually a little cheaper than for my previous (also high end) unit.',
 'My box included a 9800 charging base which is cheaper & does not display charging ring.',
 'The heads are already ruined and it is harder than it sounds to clean cheap plastic moving parts without snapping it in half.',
 'The heads on the shaver keep falling apart and I cannot get them to stay in as in the cheaper models of this shaver',
 'No advantage over cheaper versions.',
 "Didn't work any better than cheap Norelco.",

Cheap

37

Reference

- Poria, S., Cambria, E., Ku, L. W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)* (pp. 28-37).

- https://blog.csdn.net/lihaitao000/article/details/51812618 文法關係解釋

- https://corenlp.run/ Dependence Parse Tree Demo

Thanks !

Q & A ?