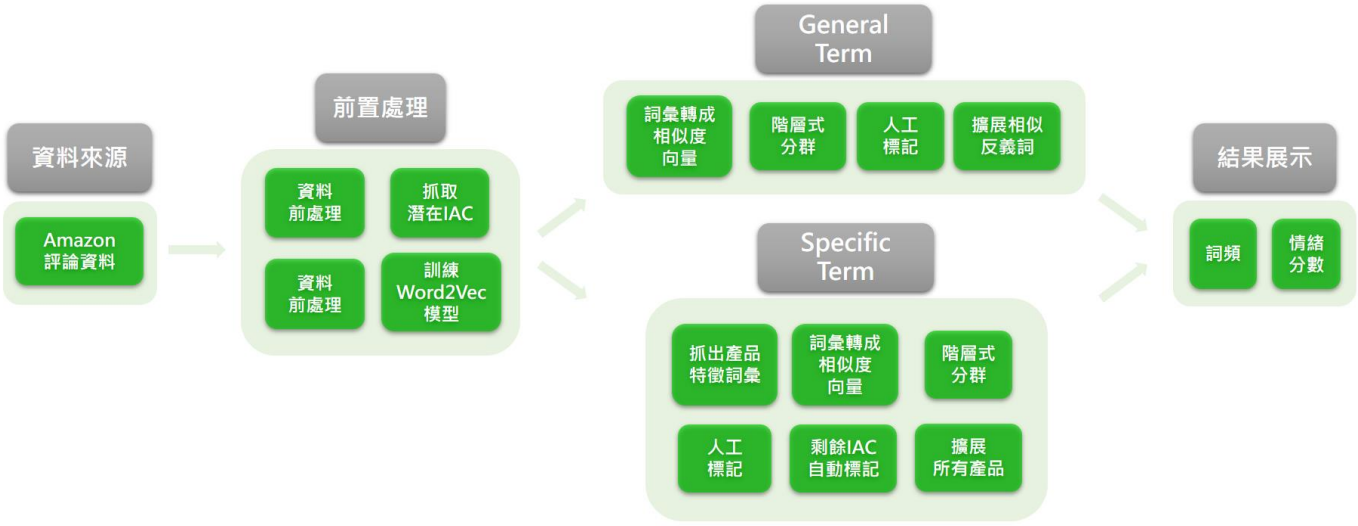
Amazon Reviewers' Aspect Detection 說明文件



圖一、系統流程圖

一、資料來源

運用 Octoparse 來爬 Amazon 的評論資料，只針對 TreeMall 中有交易的 3C 家電產品，共有 35 項產品，每項商品的評論資料至多爬取 200 則，共計爬了 232,849 則評論，資料型態欄位如下：

Name	Brand	Price	Title	Score	Time	Text	Product
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Perfect for my 2(	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 t	AirConditioner
Frigidaire FFRA0511R1 5,	Frigidaire	$139.99	Frigidaire 5000 B	5.0 out of 5 stars	5-Jun-17	This 5000 b.t.u. p	AirConditioner

二、前置處理

2.1 潛在 IAC(Get_all_IAC.py)

1. 資料前處理：

將每則評論切成句子，並使用 StanfordCoreNLP 的套件將句子拆解成 Dependency parse tree，解構每個句子的文法以及詞性，之後再進行 Lemmatization，將所有詞彙還原成字根的型態。

2. 抓出潛在 IAC：

參考 Poria(2014)[1]論文中列出的 13 個文法規則，分為三大類別：Subject Noun Rule、No Subject Noun Relation、Additional Rules

2.2 訓練 Word2Vec 模型(Train_Word2Vec.ipynb)

1. 資料前處理：

將評論中非英文的符號去除，只留下英文字母，並將所有字母統一以小寫表示，最後用 StanfordCoreNLP 將評論切成詞彙。

2. 訓練 Word2Vec 模型：

使用 gensim 中的 word2vec 套件來進行訓練，所有詞彙將以 500 維向量表示，並只留下詞頻至少為 10 的詞彙，window size 為 10。

三、General Term(General_term.ipynb)

3.1 相似度向量

從 2-1 抓出的所有潛在 IAC 將以 Word2Vec 模型訓練好的 500 維向量表示，但因 Word2Vec 有設定詞彙詞頻至少為 10，因此會有一些 IAC 不存在 Word2Vec 模型中，所以在計算相似度向量之前，先將不存在在 Word2Vec 的 IAC 刪除，並存成 valid_IAC.pkl。

若 valid_IAC 中共有 N 個詞彙，則建立 N^2 的矩陣，計算每個詞彙跟剩餘 N-1 個詞彙的不相似程度，此處每個詞彙為 500 維向量，相似度則是計算詞彙間的 cosine similarity，因為後面階層式分群的演算法是由小到大來連結，因此矩陣存成詞彙間不相似的程度，以(1-cosine similarity)表示，而對角線為 0，因為自己跟自己的不相似程度為 0。

3.2 階層式分群(Hierarchical Clustering)

計算完不相似矩陣後，就能直接使用 hierarchy 套件，將越相似的詞彙圈在一起(因為 hierarchy 套件是由小到選取要連結的詞彙，因此越相似的詞彙，其不相似程度越低，則會被優先選取)，這裡

的連結方式是使用 Ward’s Minimum Variance，群聚過程中會最小化各群內的變異加總，而門檻值為 0.4，代表當詞彙間的不相似程度高於 0.4，亦即相似度低於 0.6 時，就不再將詞彙連結起來。

3.3 人工標記

將相似的詞彙圈在一起後，就要用人工標記的方式為每種概念的群類命名，此處我們先統整出 8 大方面，分別為 Performance、Appearance、Quality、Functionality、Size、Feeling、Service 以及 Price，要將每個群類納入其中一種方面，倘若都不屬於，就將其丟掉，而有些群類可能屬於 2 種以上方面。

而標記的流程是兩位實習生獨立的各自標記，期望降低主觀的偏誤，而後再確認標記的結果是否一致，倘若不一致則經過討論決定最後歸屬。

3.4 擴展相似反義詞

因標記完的詞彙數量不多，因此透過 Wordnet 納入詞彙的同義詞和反義詞，以此來擴增字典，但英文會有一詞多義的問題，為了避免納入雜訊，故而只取 Wordnet 第一層的同義反義詞。

四、Specific Term(Specific_term.ipynb)

4.1 產品特徵詞彙

由於不同的產品會用不同的詞彙來進行描述，因此這裡要先找出專屬於各產品的特徵詞彙，透過計算所有詞彙在各個產品的 MI*TF-IDF 以及 Lift 的分數，再根據分數的大小轉換成排名，將 MI*TF-IDF 及 Lift 的名次平均，取前 500 名的詞作為屬於此產品的特徵詞彙。

4.2 相似度向量

這裡 3-1 一樣建立不相似程度的矩陣，但因只有各產品的特徵詞彙，因此為 500x500 的矩陣。

4.3 階層式分群

如同 3-2 將特徵詞彙分群，門檻值一樣為 0.4。

4.4 人工標記

如同 3-3 的方式進行人工標記，但從產品的特徵詞彙中出現很多描述產品使用對象、地點等的詞彙，因此增設一個類別 Target，最終共有 9 大方面。

4.5 剩餘 IAC 自動標記

不論在通用的詞彙上或是產品的專屬詞彙，都只有標記一小部分的 IAC 做為種子，在各個產品中仍有大量的 IAC 是沒有進行標記的，因此這邊還會對各個產品中尚未標記的 IAC 進行自動標記。自動標記的方法為：欲標記的詞彙 t 將分別與各類別中的所有詞彙計算相似度(cosine similarity)，若相似度高於 0.6，則該類別 count 加 1，最後回傳 count 最大的類別，代表此類別中含有最多跟 t 相似的詞彙，最後設下一個門檻值，若跟 t 相似的詞彙至少有 5 個，則 t 就歸屬於該類別。

4.6 擴展所有產品

因所有產品共有 35 項，我們只標記其中的 6 項產品，剩餘 29 項產品都是使用自動標記的方式擴展。

五、結果展示(Compare.ipynb)

結果展示的部分是透過計算某產品的評論中詞彙的詞頻，輸出詞頻最高的 10 個詞彙，代表評論中最常提到關鍵字，但因輸出的僅有詞彙而已，會無法分辨其情緒是否為否定的意思，因此額外使用 nltk 的套件算出一個情緒分數，表示含有該詞彙的句子平均來說是正向亦是負向的情緒。