NLP 商品詞向量 (同義詞,相似詞, 商品分類) 實作

實作步驟

1 數據處理:

1.1 取出 db_goods 不重複商品名稱數據

1.2 語料預處理()

1.3 過濾無須訓練及無用的樣本(例如:全英文商品、商品名稱長度小於 10 個字)

1.4 準備自定義詞典(詞匯表) db_goods_relate_key69, db_goods_relate_key89 中文部分，長度大於 3

1.5 準備停用自詞表(提高斷詞準確率)

2 語料樣本斷詞處裡(db_goods.g_name→db_goods_relate.g_name)

2.1 因需要商品名稱多特徵性則使用斷詞工具 jieba.analyse.extract_tags (取名詞'N'、動名詞'Vn',形容詞 Vi)等詞性篩選

2.2 產生語料樣本檔 goods_texts.txt 共取樣本數 480 萬筆/1700 萬筆，約 30hr

3 word2vec 詞向量處裡(訓練模型)

3.1 屬性:

#size:這表示的是訓練出的詞向量維度
#alpha:機器學習中的學習率，這東西會逐漸收斂到 min_alpha

#sg: sg=1表示採用skip-gram,sg=0 表示採用cbow默认sg=1是skip-gram算法，对低频词敏感；不过这里因为是计算近似词所以要选择CBOW(sg=0)；

#window:能往左往右看幾個字的意思;考慮前5個詞或後5個詞。
窗口是前後看詞的單位，3表示在目標詞前看3-B個詞，後面看b個詞（B在0-3之間隨#機）;這裡因為語料的句子太短，設置過大的窗口會導致結果並不那麼理想，最後CBOW用的3。
#min_count:若這個詞出現的次數小於min_count，那他就不會被視為訓練對象,小於設定值就會被丟棄

#sample:值越小單詞被保留下來的概率越小，預設 sample=0.001

negative 和 sample 根據訓練結果微調即可，樣品採樣雖然根據官網介紹設置為 1e-5，不能降低太多高頻詞的採樣率

3.2 470 萬商品 ==>花費 40min 訓練檔名*.model，依據各屬性訓練出合適的模型。存取模型，日後分析使用(以調整 model() 的參數，持續訓練找出合適的模型比方窗口大小、維度、學習率)

4 找出同義詞、相似詞、兩個詞彙間 Cosine 相似度、選出集合中不同類的詞語(反義詞)

4.1 同義詞、相似詞:分析 品牌、品項

4.2 抽取出結果商品

4.3 參考實作後結果: NLP 商品詞向量實作結果_不同維度.xlsx

樸素貝葉斯分類算法演算預測出商品分類

1. 貝葉斯分類算法以樣本可能屬於某類的概率來作為分類依據
P(xy|z)=p(xyz)/p(z)=p(xz)/p(z)*p(yz)/p(z)

P(類別 | 特徵)=P(特徵 | 類別)*P(類別) / P(特徵)

2. 樸素貝葉斯機率模型簡介簡單貝氏模型直接假設所有的隨機變數之間具有條件獨立的情況，因此可以直接利用條件機率相乘的方法，計算出聯合機率，監督式學習的樣本集中能取得得非常好的分類效果。

參考: https://www.jb51.net/article/142622.htm

參考: http://mropengate.blogspot.com/2015/06/ai-ch14-3-naive-bayes-classifier.html

計算方式詳解: https://www.jb51.net/article/143146.htm

3. 算法步驟:

① :分解樣本數據中的特徵(分詞)

② :計算算各類據中，各特徵的條件概率
（比如：特徵 1 出現的情況下，屬於 A 類的概率 p(A|特徵 1)，屬於 B 類的概率 p(B|特徵 1)，屬於 C 類的概率 p(C|特徵 1).....）

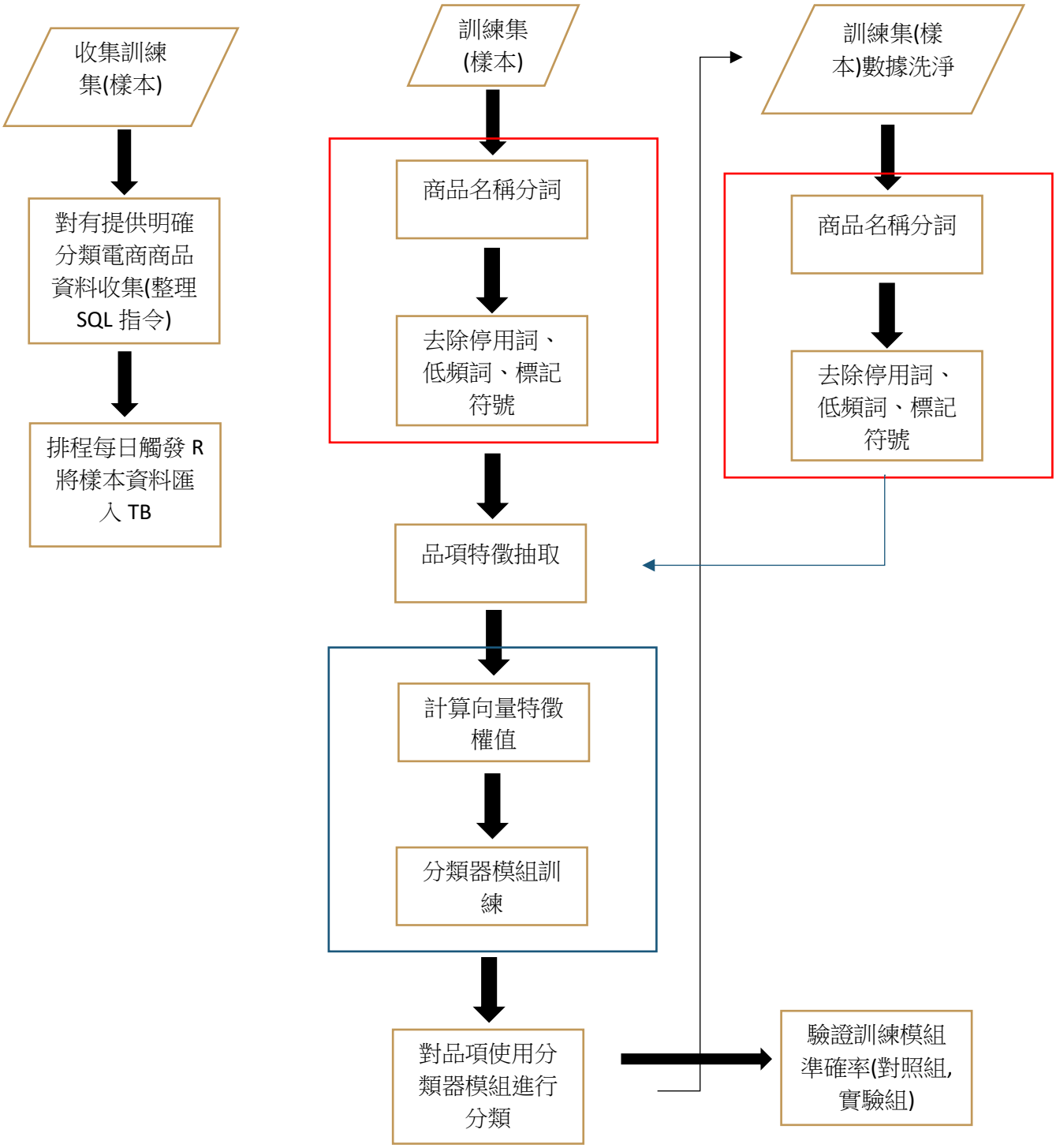
③ : 分解待分類數據中的特徵（特徵 1、特徵 2、特徵 3、特徵 4.....）

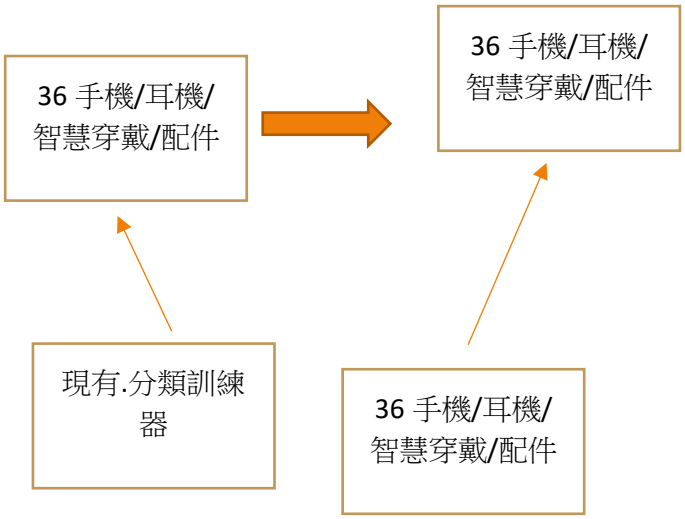
④ : 計算各特徵的各條件概率的乘積，如下所示：

判斷為 A 類的概率：p(A|特徵 1)*p(A|特徵 2)*p(A|特徵 3)*p(A|特徵 4).....
判斷為 B 類的概率：p(B|特徵 1)*p(B|特徵 2)*p(B|特徵 3)*p(B|特徵 4).....

判斷為 C 類的概率：p(C|特徵 1)*p(C|特徵 2)*p(C|特徵 3)*p(C|特徵 4).....

- 參考: https://www.imooc.com/article/36694
- 參考:https://www.jb51.net/article/142622.htm
- 參考: http://www.ruanyifeng.com/blog/2013/12/naive_bayes_classifier.html




36 手機/耳機/
智慧穿戴/配件
36 手機/耳機/
智慧穿戴/配件
現有.分類訓練
器
36 手機/耳機/
智慧穿戴/配件

文字探勘 WORD2VEC 的介紹

- Word2vec 是 Google 公司在 2013 年開放的一款用於訓練詞向量的軟體工具。基於非監督學習的 word2vec。
- 屬於類神經網路概率語言模型，具有良好的語義特性，是表示詞語特徵的常用方式。詞向量的每一維的值代表一個具有一定的語義和語法上解釋的特徵。故可以將詞向量的每一維稱為一個詞語特徵。

  例如: 商品名稱: 原廠公司貨 國際牌 窗型冷氣機 排水彎頭/L 型彎管

  斷詞分析: 排水彎頭,原廠,窗型冷氣機,國際牌,彎管

  每個詞用一個很長的向量表示，向量的維度表示詞表大小，絕大多數是 0，只有一個維度是 1，代表當前詞。 假設:「國際牌」表示為 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 ...] 即從 0 開始「國際牌」記為 3，「窗型冷氣機」[1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 ...] 記為 0 ....
- 根據給定的語料庫，通過優化後的訓練模型快速有效的將一個詞語表達成向量形式，其核心架構包括 CBOW 和 Skip-gram。
- 訓練語料其值越大模型就越準確。
- 參考: https://zhuanlan.zhihu.com/p/27234078
- 參考: http://zake7749.github.io/2016/08/28/word2vec-with-gensim/
- 參考 Word2Vec 訓練同義詞模型: https://blog.csdn.net/chunyun0716/article/details/60465806
- 參考: http://cpmarkchang.logdown.com/posts/773062-neural-network-word2vec-part-1-overview

WORD2VEC 應用

聊天機器人、機器人語言感知、客戶語情分析(評語分析)、語音交互應用(小米語音助理)、新聞分類、詞雲、本文分類、同義詞挖掘.....

- 參考: https://buzzorange.com/techorange/tag/artificialintelligence/
- 參考: https://hk.wxwenku.com/d/102306480
- 參考: https://www.zhihu.com/question/25269336

- 參考: 自己動手做聊天機器人教程 : https://github.com/warmheartli/ChatBotCourse
- IT 邦幫 word2vec 介紹 產品標籤分類: https://ithelp.ithome.com.tw/articles/10194369