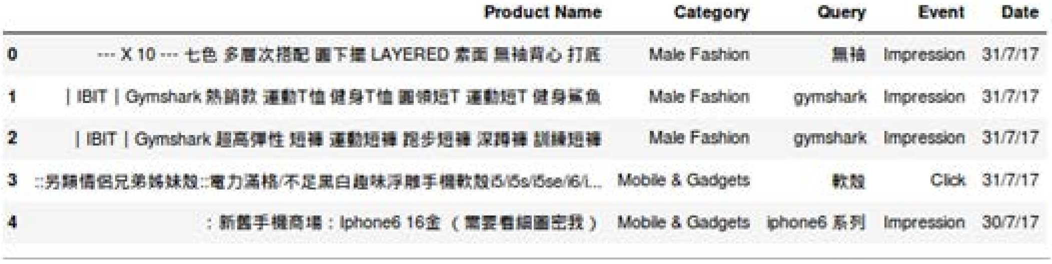
# Procedures for the case study of product keywords extraction

For a given product, to recommend a list of best keywords, the main idea is to find a similar product list from the train dataset, evaluate the queries listed in the similar product list, and then recommend the query with good qualities as the final recommended keywords. Query with good quality means query may bring more clicks and more relevant to the given product.

## Step 1: Clean the train data

Including loading the data, retrieving the basic information of the dataset, adding the header for the dataframe and removing the null data. The train data is represented as follows after processing



| | Product Name | Category | Query | Event | Date |
|---|---|---|---|---|---|
| 0 | --- X 10 --- 七色 多層次搭配 圓下擺 LAYERED 素面 無袖背心 打底 | Male Fashion | 無袖 | Impression | 31/7/17 |
| 1 | ｜IBIT｜Gymshark 熱銷款 運動T恤 健身T恤 圓領短T 運動短T 健身鯊魚 | Male Fashion | gymshark | Impression | 31/7/17 |
| 2 | ｜IBIT｜Gymshark 超高彈性 短褲 運動短褲 跑步短褲 深蹲褲 訓練短褲 | Male Fashion | gymshark | Impression | 31/7/17 |
| 3 | ::另類情侶兄弟姊妹殼::電力滿格/不足黑白趣味浮雕手機軟殼i5/i5s/i5se/i6/i... | Mobile & Gadgets | 軟殼 | Click | 31/7/17 |
| 4 | ：新舊手機商場：Iphone6 16金（需要看細圖密我） | Mobile & Gadgets | iphone6 系列 | Impression | 30/7/17 |

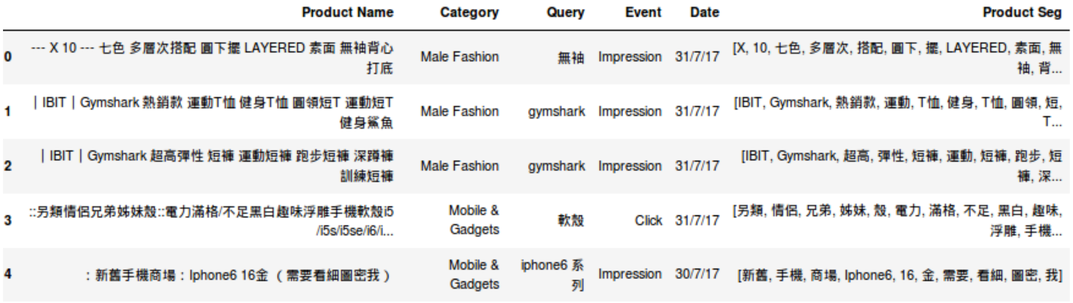
Function and description used in this step

| Function | Description |
|---|---|
| preprocess | to clean the train data |

The following steps (step 2 and step 3) aim to construct a similarity index model

## Step 2: Process the train data

2.1 Segment the product name in the train data using jieba library



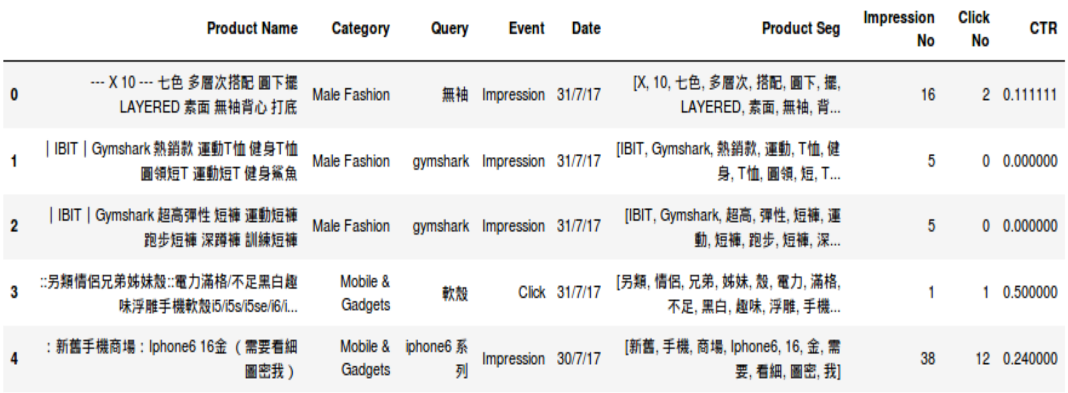
| | Product Name | Category | Query | Event | Date | Product Seg |
|---|---|---|---|---|---|---|
| 0 | --- X 10 --- 七色 多層次搭配 圓下擺 LAYERED 素面 無袖背心 打底 | Male Fashion | 無袖 | Impression | 31/7/17 | [X, 10, 七色, 多層次, 搭配, 圓下, 擺, LAYERED, 素面, 無袖, 背... |
| 1 | ｜IBIT｜Gymshark 熱銷款 運動T恤 健身T恤 圓領短T 運動短T 健身鯊魚 | Male Fashion | gymshark | Impression | 31/7/17 | [IBIT, Gymshark, 熱銷款, 運動, T恤, 健身, T恤, 圓領, 短, T... |
| 2 | ｜IBIT｜Gymshark 超高彈性 短褲 運動短褲 跑步短褲 深蹲褲 訓練短褲 | Male Fashion | gymshark | Impression | 31/7/17 | [IBIT, Gymshark, 超高, 彈性, 短褲, 運動, 短褲, 跑步, 短褲, 深... |
| 3 | ::另類情侶兄弟姊妹殼::電力滿格/不足黑白趣味浮雕手機軟殼i5/i5s/i5se/i6/i... | Mobile & Gadgets | 軟殼 | Click | 31/7/17 | [另類, 情侶, 兄弟, 姊妹, 殼, 電力, 滿格, 不足, 黑白, 趣味, 浮雕, 手機... |
| 4 | ：新舊手機商場：Iphone6 16金（需要看細圖密我） | Mobile & Gadgets | iphone6 系列 | Impression | 30/7/17 | [新舊, 手機, 商場, Iphone6, 16, 金, 需要, 看細, 圖密, 我] |

Function and description

| Function | Description |
|---|---|
| jieba_fenci | to exclude all the special symbols and segment the product name |
| remove_space | to remove null data after split the product name |

2.2 Evaluate the queries

Impression No – frequency of a query appears as ‘Impression’ in train data set.
Click No – frequency of a query that appear as ‘Click’ in the train data set.
CTR – Click No/(Click No+ Impression No)



| | Product Name | Category | Query | Event | Date | Product Seg | Impression No | Click No | CTR |
|---|---|---|---|---|---|---|---|---|---|
| 0 | --- X 10 --- 七色 多層次搭配 圓下擺 LAYERED 素面 無袖背心 打底 | Male Fashion | 無袖 | Impression | 31/7/17 | [X, 10, 七色, 多層次, 搭配, 圓下, 擺, LAYERED, 素面, 無袖, 背... | 16 | 2 | 0.111111 |
| 1 | ｜IBIT｜Gymshark 熱銷款 運動T恤 健身T恤 圓領短T 運動短T 健身鯊魚 | Male Fashion | gymshark | Impression | 31/7/17 | [IBIT, Gymshark, 熱銷款, 運動, T恤, 健身, T恤, 圓領, 短, T... | 5 | 0 | 0.000000 |
| 2 | ｜IBIT｜Gymshark 超高彈性 短褲 運動短褲 跑步短褲 深蹲褲 訓練短褲 | Male Fashion | gymshark | Impression | 31/7/17 | [IBIT, Gymshark, 超高, 彈性, 短褲, 運動, 短褲, 跑步, 短褲, 深... | 5 | 0 | 0.000000 |
| 3 | ::另類情侶兄弟姊妹殼::電力滿格/不足黑白趣味浮雕手機軟殼i5/i5s/i5se/i6/i... | Mobile & Gadgets | 軟殼 | Click | 31/7/17 | [另類, 情侶, 兄弟, 姊妹, 殼, 電力, 滿格, 不足, 黑白, 趣味, 浮雕, 手機... | 1 | 1 | 0.500000 |
| 4 | ：新舊手機商場：Iphone6 16金（需要看細圖密我） | Mobile & Gadgets | iphone6 系列 | Impression | 30/7/17 | [新舊, 手機, 商場, Iphone6, 16, 金, 需要, 看細, 圖密, 我] | 38 | 12 | 0.240000 |

Function and description

| Function | Description |
|---|---|
| event_count | to count the number of 'impression' and 'click' for each query |

## Step 3: Create a similarity model

3.1 Create a ‘dictionary’ of the training corpus’ raw text:

```
dictionary = corpora.Dictionary(texts_cut)
```

3.2 Convert the training corpus to vector space:

```
corpus = dictionary.doc2bow(texts_cuts)
```

3.3 Initialize the TF-IDF model and convert the corpus to vectors in tf-idf space

```
tfidf = models.TfidfModel(corpus)
corpus_tfidf = tfidf[corpus]
```

3.4 Create a similarity index model

```
index =similarities.MatrixSimilarity(corpus_tfidf)
```

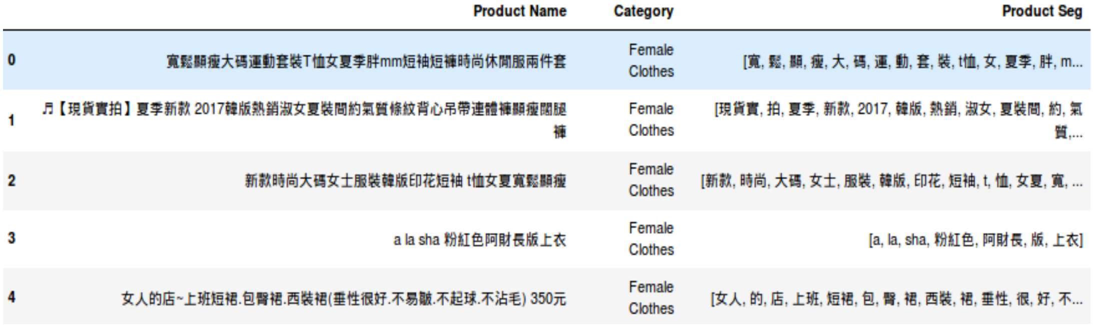
For vectors in the tfidf space, it is shown that the maximum length is 50, and the average length is about 15, so we don't choose the lsi model, which is a topic model and can be used to project the vectors into a low dimensional space, meanwhile lsi model with large num_topic needs long time to train.

Function and description

| Function | Description |
|---|---|
| get_index | to create a dictionary, corpus, tdidf model and a similarity index model |

## Step 4: Get the similar product list for a new product from test data

4.1 Segment the product name in the test data



| | Product Name | Category | Product Seg |
|---|---|---|---|
| 0 | 寬鬆顯瘦大碼運動套裝T恤女夏季胖mm短袖短褲時尚休閒服兩件套 | Female Clothes | [寬, 鬆, 顯, 瘦, 大, 碼, 運, 動, 套, 裝, t恤, 女, 夏季, 胖, m... |
| 1 | ♬【現貨實拍】夏季新款 2017韓版熱銷淑女夏裝簡約氣質條紋背心吊帶連體褲顯瘦闊腿褲 | Female Clothes | [現貨實, 拍, 夏季, 新款, 2017, 韓版, 熱銷, 淑女, 夏裝簡, 約, 氣質,... |
| 2 | 新款時尚大碼女士服裝韓版印花短袖 t恤女夏寬鬆顯瘦 | Female Clothes | [新款, 時尚, 大碼, 女士, 服裝, 韓版, 印花, 短袖, t, 恤, 女夏, 寬, ... |
| 3 | a la sha 粉紅色阿財長版上衣 | Female Clothes | [a, la, sha, 粉紅色, 阿財長, 版, 上衣] |
| 4 | 女人的店~上班短裙.包臀裙.西裝裙(垂性很好.不易皺.不起球.不沾毛) 350元 | Female Clothes | [女人, 的, 店, 上班, 短裙, 包, 臀, 裙, 西裝, 裙, 垂性, 很, 好, 不... |

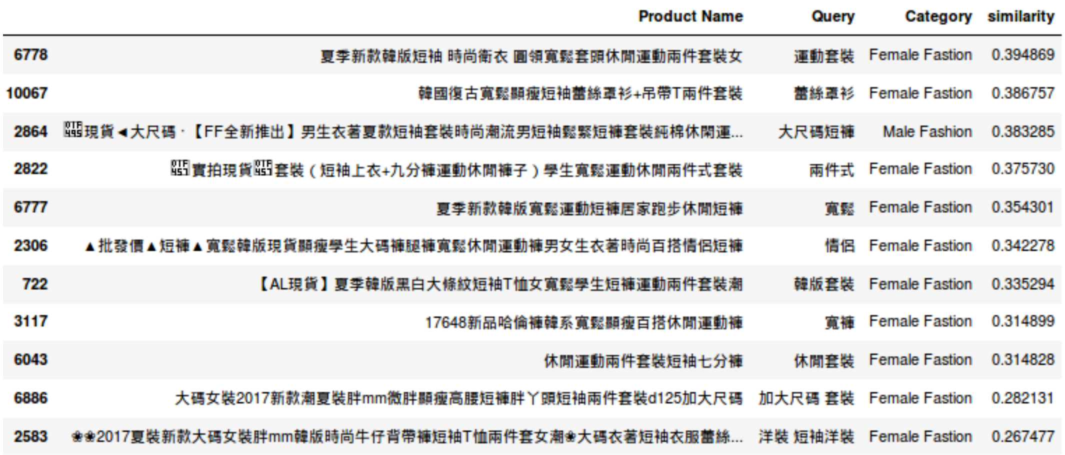
4.2 For a given product, get the similarity product list

```
query_bow = dictioary.doc2bow(query_product_cuts)
query_tfidf = tfidf[query_bow]
sim = index[query_tfidf]
```

For a given product such as

寬鬆顯瘦大碼運動套裝T恤女夏季胖mm短袖短褲時尚休閒服兩件套

Get the similar product list from the train dataset with 'num_similar_product_list' =40, 'similarity' is the similarity score between the given product and the corresponding similar products.



| | Product Name | Query | Category | similarity |
|---|---|---|---|---|
| 6778 | 夏季新款韓版短袖 時尚衛衣 圓領寬鬆套頭休閒運動兩件套裝女 | 運動套裝 | Female Fastion | 0.394869 |
| 10067 | 韓國復古寬鬆顯瘦短袖蕾絲罩衫+吊帶T兩件套裝 | 蕾絲罩衫 | Female Fastion | 0.386757 |
| 2864 | □現貨◄大尺碼·【FF全新推出】男生衣著夏款短袖套裝時尚潮流男短袖鬆緊短褲套裝純棉休閒運... | 大尺碼短褲 | Male Fashion | 0.383285 |
| 2822 | □實拍現貨□套裝（短袖上衣+九分褲運動休閒褲子）學生寬鬆運動休閒兩件式套裝 | 兩件式 | Female Fastion | 0.375730 |
| 6777 | 夏季新款韓版寬鬆運動短褲居家跑步休閒短褲 | 寬鬆 | Female Fastion | 0.354301 |
| 2306 | ▲批發價▲短褲▲寬鬆韓版現貨顯瘦學生大碼褲腿褲寬鬆休閒運動褲男女生衣著時尚百搭情侶短褲 | 情侶 | Female Fastion | 0.342278 |
| 722 | 【AL現貨】夏季韓版黑白大條紋短袖T恤女寬鬆學生短褲運動兩件套裝潮 | 韓版套裝 | Female Fastion | 0.335294 |
| 3117 | 17648新品哈倫褲韓系寬鬆顯瘦百搭休閒運動褲 | 寬褲 | Female Fastion | 0.314899 |
| 6043 | 休閒運動兩件套裝短袖七分褲 | 休閒套裝 | Female Fastion | 0.314828 |
| 6886 | 大碼女裝2017新款潮夏裝胖mm微胖顯瘦高腰短褲胖丫頭短袖兩件套裝d125加大尺碼 | 加大尺碼 套裝 | Female Fastion | 0.282131 |
| 2583 | ❀❀2017夏裝新款大碼女裝胖mm韓版時尚牛仔背帶褲短袖T恤兩件套女潮❀大碼衣著短袖衣服蕾絲... | 洋裝 短袖洋裝 | Female Fastion | 0.267477 |

Function and description

| Function | Description |
|---|---|
| get_similar_product_list | to get a similar product list for product from test data (configurable parameter: num_similar_product_list) |

## Step 5: Further evaluate query quality so as to recommend keywords

5.1 Calculate more evaluators of the query

query_count – frequency of a ‘query’ that appears in the similar product list
relevance_score – relevance between the ‘query’ and the given product

How to compute the relevance-score for the query?
First we need to segment the query if necessary, the number of terms that appears in the given product divided by the total number of terms in the segmented query will be the relevance_score for the particular query.

For a query such as ’大尺码短裤’, after segment becomes [大，尺码，短裤], so number of terms that appears in the given product = 2, relevance_score for this query = 2/3 =0.6667

5.2 Compute the total score for a particular query

To get a total score for a ‘query’, we sum up the evaluator with different weights. Please note that the weights are tuned according the performance of the final keyword recommendation.

*evaluator = [query_count*, click*, CTR, event, similarity, relevance_score]*
*weights = [0.1, 0.1, 0.1, 0.1, 0.2, 0.4]*

To make sure each evaluator is [0,1], we need to standardization the evaluators.

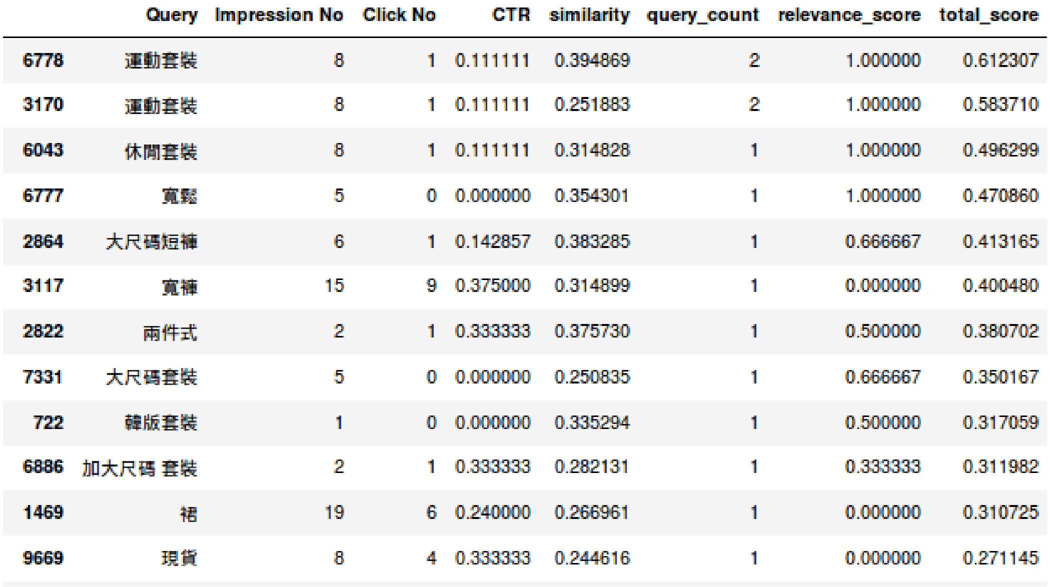
*query_count*= query_count/ query_count.max()*
*click* = Click No/ Click No.max()*

The total score for the query can be evaluated as follows:

*total_score = sum(evaluator.*weights)*

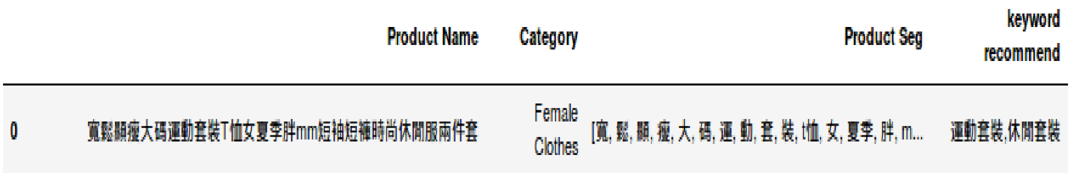
It is observed that ‘total_score’ for each query falls into [0,1].

The performance for the query is listed as follows:



| | Query | Impression No | Click No | CTR | similarity | query_count | relevance_score | total_score |
|---|---|---|---|---|---|---|---|---|
| 6778 | 運動套裝 | 8 | 1 | 0.111111 | 0.394869 | 2 | 1.000000 | 0.612307 |
| 3170 | 運動套裝 | 8 | 1 | 0.111111 | 0.251883 | 2 | 1.000000 | 0.583710 |
| 6043 | 休閒套裝 | 8 | 1 | 0.111111 | 0.314828 | 1 | 1.000000 | 0.496299 |
| 6777 | 寬鬆 | 5 | 0 | 0.000000 | 0.354301 | 1 | 1.000000 | 0.470860 |
| 2864 | 大尺碼短褲 | 6 | 1 | 0.142857 | 0.383285 | 1 | 0.666667 | 0.413165 |
| 3117 | 寬褲 | 15 | 9 | 0.375000 | 0.314899 | 1 | 0.000000 | 0.400480 |
| 2822 | 兩件式 | 2 | 1 | 0.333333 | 0.375730 | 1 | 0.500000 | 0.380702 |
| 7331 | 大尺碼套裝 | 5 | 0 | 0.000000 | 0.250835 | 1 | 0.666667 | 0.350167 |
| 722 | 韓版套裝 | 1 | 0 | 0.000000 | 0.335294 | 1 | 0.500000 | 0.317059 |
| 6886 | 加大尺碼 套裝 | 2 | 1 | 0.333333 | 0.282131 | 1 | 0.333333 | 0.311982 |
| 1469 | 裙 | 19 | 6 | 0.240000 | 0.266961 | 1 | 0.000000 | 0.310725 |
| 9669 | 現貨 | 8 | 4 | 0.333333 | 0.244616 | 1 | 0.000000 | 0.271145 |

Then we recommend the keywords based on the total score of the query. Let num_keywords =2,



| | Product Name | Category | Product Seg | keyword recommend |
|---|---|---|---|---|
| 0 | 寬鬆顯瘦大碼運動套裝T恤女夏季胖mm短袖短褲時尚休閒服兩件套 | Female Clothes | [寬, 鬆, 顯, 瘦, 大, 碼, 運, 動, 套, 裝, t恤, 女, 夏季, 胖, m... | 運動套裝,休閒套裝 |

Function and description

| Function | Description |
|---|---|
| get_relevance | to evaluate the relevance between the recommend query and the given product |
| get_evaluate_product_list | to get a similar product list with more evaluators |
| get_recommed_key_word | to recommend keywords according the overall query performance(configurable parameter: num_keywords) |