Big Data Analytics and Applications
Final project – Ruten identify main products and accessories

- Problem

Ruten is one of the biggest Marketplace in Taiwan. It has more than eleven million members. since 2018, Ruten has provided more e-commerce big data analytic functions and Al services in an open structure and has become Taiwan's central Al e-commerce platform.

As one company with significant influence in e-commerce worldwide, Ruten strives to optimize its e-commerce platform performance by prioritizing customer satisfaction. One of the developments Ruten will carry out is identifying the main product and accessories to leverage customer satisfaction. So when a customer wants to search for a product, the application will also suggest accessories related to the customer's search. For example, many mobile phone back cases show up when customers look for mobile phones.

So our goal is to divide the main product and accessories.

- Solution

We do experiment utilizing CKIP-tagger. Chinese Knowledge and Information Processing (CKIP)-tagger is an open-source library that implements neural CKIP-style Chinese NLP tools. Figure 1 is the flowchart of our solution. This process can be explained as follows:

1) We catch all data with a "keyword" in the category_name and item_name to a group.
2) If the category_name is the same as "keyword", we categorize it into main product; If not, we categorize into the accessories.
3) We used CKIP-tagger to split the item_name and analyze the part of speech for each word to classify it again from the accessories; if item_name contains "keyword", we categorize it into the main product.
4) Finally, we use parts of speech to categorize the main products. We found the two most common parts of speech combinations of accessories. If the item_name in the main product has these two kinds part of speech combinations, we classify them as accessories.

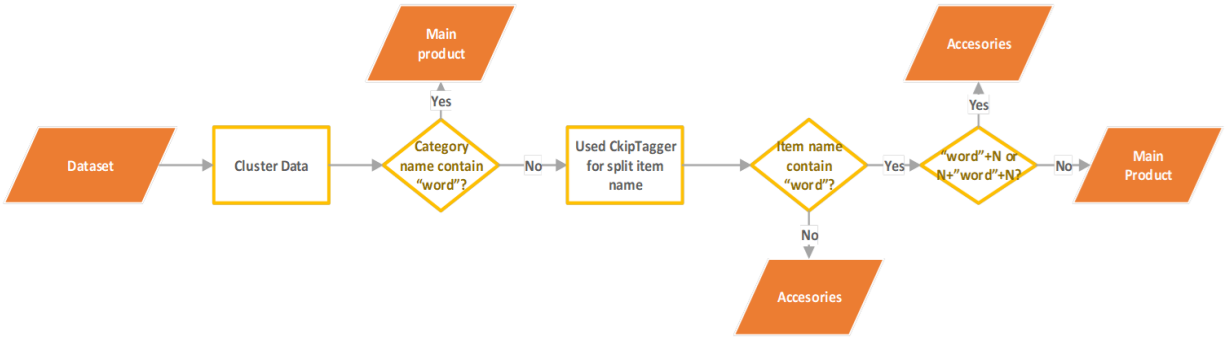


Figure 1. Flowchart experiment using CKIP-Tagger

- Results

In the first step, we collect all the products in "category_name" and "item_name" where "keyword" appears. (Figure 4 uses "相機" as a keyword.)

	index	category_name	category_id	item_id	price	item_name	seller_nickname	seller_id	type	product_type
3917	3917	相機	8000100000000.0	21716900000000.0	3750.0	【新鎂】買就送原廠電池 CANON IXUS 185 小型數位相機 全新台灣佳能公司貨	newmei33	10269860}	text_group	Main_product
9389	9389	相機	8000800000000.0	21952100000000.0	599.0	SanDisk 128GB 128G SDXC Extreme【150MBVs】SDHC ...	top_card	6093153}	text_group	Main_product
11483	11483	相機	8000100000000.0	22028800000000.0	300.0	SONY T300 DSC-T300 1010萬畫素V3.50吋觸控式螢幕 二手盒裝數位相機	crown830826	4429733}	text_group	Main_product
19105	19111	相機	8000100000000.0	22117100000000.0	1600.0	*羅浮工作室=免郵資,功能保固*SONY DSC-T900 數位相機*深棕*	n222227	2822750}	text_group	Main_product
19821	19827	相機	8000100000000.0	22120500000000.0	1200.0	給ywks下單用 Canon IXUS 185 口袋相機 8倍光學 輕巧 1cm微距	kay0606	3721191}	text_group	Main_product
...	...	...	...	...	...	...	...	...	...	...
4831994	9117	相機	8000100000000.0	21948900000000.0	2500.0	[保固一年] [明豐相機] 公司貨 SAMSUNG EX2 EX2F F1.4 大光圈 ...	a5219334	2286593}	text_group	Main_product

Figure 2. Example for where category_name = “相機” → produc_type = Main_product

Next, we will temporarily classify the products that category_name are not the same as "相機" into "text_group".

	index	category_name	category_id	item_id	price	item_name	seller_nickname	seller_id	type
220	220	USB線	1100170003.0	21102100000000.0	20.0	MP3 MP4 USB線 手機 數位相機 讀卡機等專用USB傳輸線 USB轉轉接線 線長60公分	ce6635	1581442}	text_group
247	247	側背、斜肩	8000500000000.0	21104000000000.0	550.0	又敗家(無字酒紅彎型寬版)韓國製Matin馬田減壓相機背帶M-6784相機減壓背帶單眼相機揹...	yoyo55	2527184}	text_group
413	413	相機皮套	800050007.0	21112200000000.0	280.0	Panasonic GF8X GF7X GF6X GX1 GF5X GF3X GF2 GF1...	mingyang65	2252062}	text_group
415	415	RF 連動測距相機	800020001.0	21112200000000.0	800.0	Canon PRIMA Zoom 70F相機	jc4812	2544758}	text_group
717	717	其他	80006.0	21212100000000.0	199.0	含稅附發票】相機攝影外接閃光燈配件 11吋金屬魔術手臂 魔術怪手 萬向360度旋轉支架 轉接...	baohong	2981045}	text_group
...	...	...	...	...	...	...	...	...	...
4846035	23163	其他	90014.0	22133600000000.0	532.0	佳能相機包 單反單肩微單攝影EOS90D70D 200D二代 800D M50M6 胖虎y	leean7245	17528271}	text_group
4847530	24658	其他	90014.0	22141500000000.0	1504.0	【挪美索尼相機配件】caseman C10攝影三角包單反相機包索尼休閒內膽單肩	dqdqw98d3a	18576076}	text_group

Figure 3. Example for, where category_name is not the same as "keyword"

In the text group, we use the CKIP-tagger to segment the item_name and judge the part of speech of each word.

If category_name is not the same "keyword" & item_name = "keyword", classify them in Accessories".

	index	category_name	category_id	item_id	price	item_name	seller_nickname	seller_id	type	product type
415	415	RF 連動測距相機	800020001.0	21112200000000.0	800.0	Canon PRIMA Zoom 70F相機	jc4812	2544758}	text_group	accessories

```
original item name :  Canon PRIMA Zoom 70F相機
word segmentation result: [['Canon PRIMA Zoom 70F', '相機']]
Ports of Speech for each word: [['FW', 'Na']]
```

Figure 4. Example for, where category_name is not the same as "keyword" & item_name is the same as "keyword"

In addition, if the keyword is connected with other words by the word segmentation system in the item name(e.g 相機→相機包), and the category name is not equal to the "相機", it means that it is an accessory.

	index	category_name	category_id	item_id	price	item_name	seller_nickname	seller_id	type	product type
926	926	鏡頭袋	800050032.0	21306200000000.0	210.0	全新@幸運草@Nikon L號-防撞包 保護套 內膽包 單眼相機包 D600\/D610\/...	song321	1615797}	text_group	accessories

```
原始item name :  全新@幸運草@Nikon L號-防撞包 保護套 內膽包 單眼相機包 D600\/D610\/D750 D80 D90..
分詞結果: [['全新', '@', '幸運草@Nikon L號', '-', '防撞包', ' ', '保護套', ' ', '內', '膽包', ' ', '單眼', '相機包', ' D600\\/D610
\\/D750 D80 D90..']]
每個詞的詞性: [['VH', 'FW', 'Nb', 'DASHCATEGORY', 'Na', 'WHITESPACE', 'Na', 'WHITESPACE', 'Ncd', 'Na', 'WHITESPACE', 'A', 'Na',
'Neu']]
```

Figure 5. Example for, where category_name is not the same as "keyword" & item_name is not the same as "keyword"

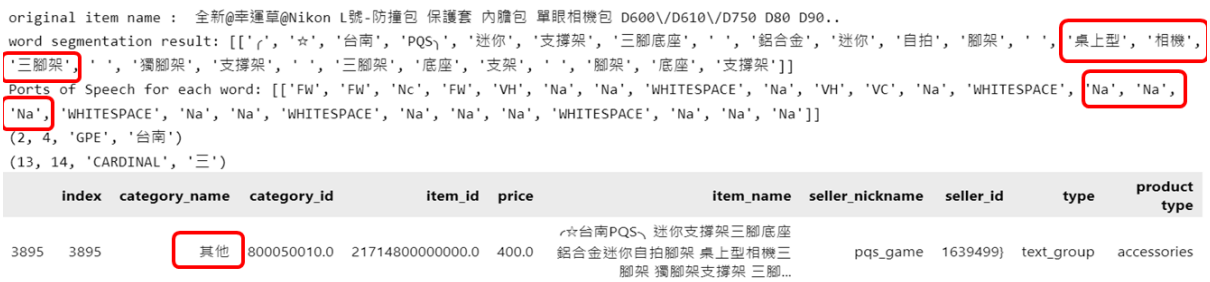
If, after the CKIP-tagger system, the "keyword" is connected with a noun, judge it as an accessory.

```
original item name :  Canon Tripod Grip HG-100TBR 三腳架手把 公司貨 相機手柄 HG100TBR G7XM3適用
word segmentation result: [['Canon Tripod Grip ', 'HG-', '100', 'TBR ', '三腳架', '手把', ' ', '公司貨', ' ', '相機', '手柄', ' HG100', 'TBR
G7', 'XM3', '適用']]
Ports of Speech for each word: [['FW', 'FW', 'Neu', 'FW', 'Na', 'Na', 'WHITESPACE', 'Na', 'WHITESPACE', 'Na', 'Na', 'FW', 'FW', 'FW', 'VJ']]
```

	index	category_name	category_id	item_id	price	item_name	seller_nickname	seller_id	type	product type
11067	11067	電池手把	800050027.0	22023600000000.0	2900.0	Canon Tripod Grip HG-100TBR 三腳架手把 公司貨 相機手柄	yushuang82	1630569}	text_group	accessories

Figure 6. Example for, where category_name is not the same as "keyword" & item_name is the same as "keyword" +N

If after the CkipTagger system, the "keyword" connects nouns before and after, judge it as an accessory too.



```
original item name :  全新@幸運草@Nikon L號-防撞包 保護套 內膽包 單眼相機包 D600\/D610\/D750 D80 D90..
word segmentation result: [['ㄏ', '☆', '台南', 'PQSㄟ', '迷你', '支撐架', '三腳底座', ' ', '鋁合金', '迷你', '自拍', '腳架', ' ', '桌上型', '相機',
'三腳架', ' ', '獨腳架', '支撐架', ' ', '三腳架', '底座', '支架', ' ', '腳架', '底座', '支撐架']]
Ports of Speech for each word: [['FW', 'FW', 'Nc', 'FW', 'VH', 'Na', 'Na', 'WHITESPACE', 'Na', 'VH', 'VC', 'Na', 'WHITESPACE', 'Na', 'Na',
'Na', 'WHITESPACE', 'Na', 'Na', 'WHITESPACE', 'Na', 'Na', 'Na', 'WHITESPACE', 'Na', 'Na', 'Na']]
(2, 4, 'GPE', '台南')
(13, 14, 'CARDINAL', '三')
```

	index	category_name	category_id	item_id	price	item_name	seller_nickname	seller_id	type	product type
3895	3895	其他	800050010.0	21714800000000.0	400.0	ㄏ☆台南PQSㄟ 迷你支撐架三腳底座鋁合金迷你自拍腳架 桌上型相機三腳架 獨腳架支撐架 三腳...	pqs_game	1639499}	text_group	accessories

Figure 7. Example for, where category_name is not the same as "keyword" & item_name is the same as N+ "keyword" +N

Finally, the products that are not marked are all regarded as the main products

- Conclusion
  1. We propose a structure that can be used to classify main products and accessories, and it can apply to the real work problem.
  2. We found that part of speech combinations can be used to determine the product type, which helps us improve judgment accuracy.