

# 고객 리뷰 활용 **Voc**

Voice of the customers

# 논문 연구 배경

pRank을 활용하여 제품의 리뷰를 분석하고  
특정기능에 기반한 품질랭킹을 만들어  
사용자의 니즈를 만족시키는 지표를 제공한다

pRank이란?

Google의 검색엔진 알고리즘(PageRank)을 리뷰데이터 분석에 적합하게 변형시킨 알고리즘

# 프로젝트 목적

- 고객 리뷰를 분석하여 제품의 기능별 랭킹을 만든다

1. 아마존 연구 데이터를 활용하여 리뷰 데이터 수집
2. 제품 간에 비교 할 '기능'을 10가지 선정
3. 리뷰의 성향을 분석하여 분류 (CS/SS분류 / 극성분류)
4. pRank알고리즘을 해서 기능별 순위를 선정

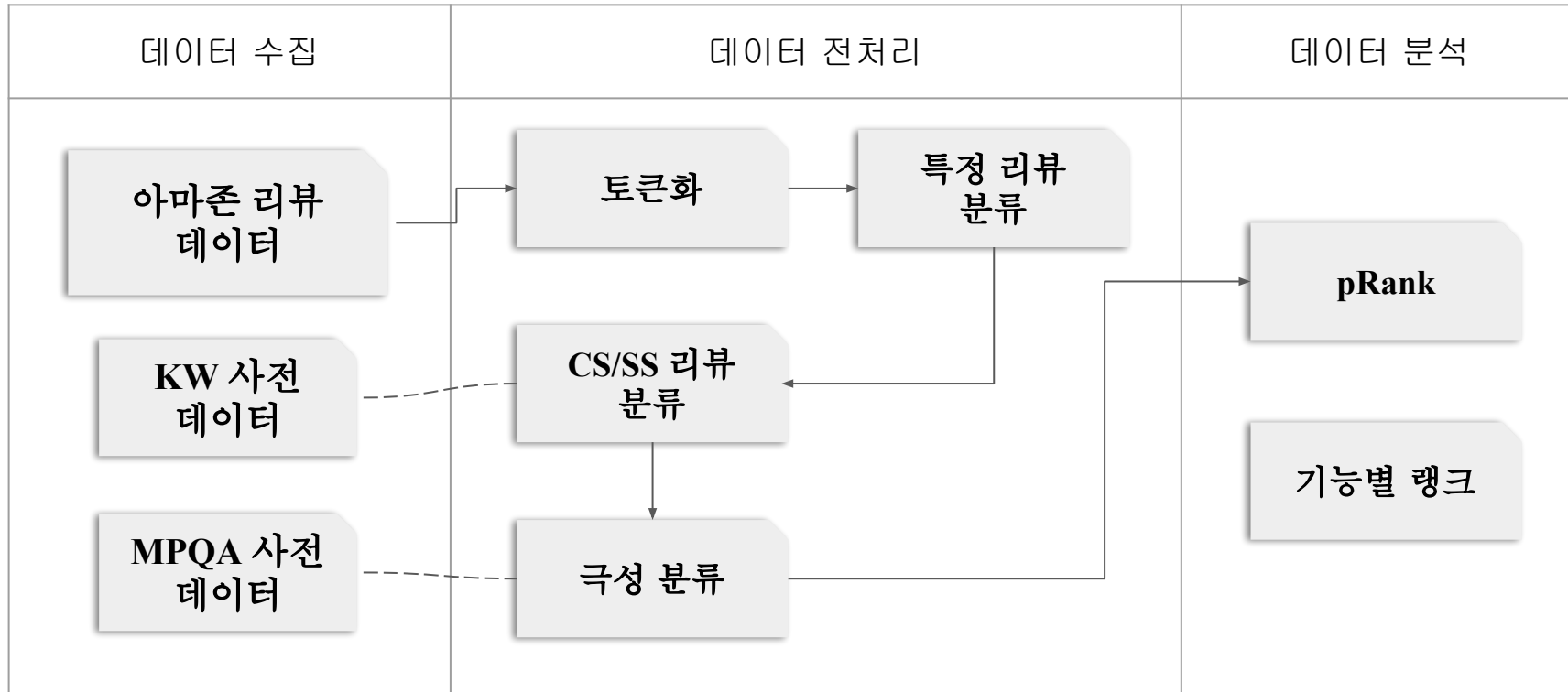
CS / SS 분류이란?

리뷰를 다른 제품과 비교한 리뷰(CS)와 해당 제품만 평가한 리뷰(SS)로 분류하는 것

극성분류이란?

리뷰의 어조를 분석하여 제품에 대해 우호적인 문장과 비판적인 문장으로 분류하는 것

# 논문 구현 요약



# 역할분담



강동훈

- 자체 데이터 추출
- MPQA 사전 추출
- pRank 결과값 출력



이동재

- 데이터 정의 사전 작성
- 주요 함수 설계
- 프로젝트 발표



김요한

- 제품 모델명 리스트 전처리
- KW 사전 추출
- PPT 작성



정윤성

- 주요 함수 설계
- 제품 모델명 리스트 작성
- KW 사전 분류
- pRank 결과값 출력

# 환경구축

## 구글 드라이브

데이터 저장 - 입력 데이터 csv, 출력 데이터 csv, 참조 데이터 csv(or xlsx) 저장  
데이터 정의 - 설계한 함수 및 csv 칼럼 설명

## 트렐로

일정 확인 - elegantt를 이용하여 진행 날짜/Deadline 확인  
To Do List - To Do List 작성, 할 일 분배 및 한 일 확인

## 깃헙

코드 공유 - 단계별로 폴더를 만들어 코드 관리  
- readme를 통해 버전 관리

# 프로세스



- 아마존 연구 데이터를 활용하여 메타 데이터와 리뷰 데이터를 수집.

메타 데이터 : 498,196 개  
리뷰 데이터 : 7,824,482개  
카메라 리뷰 데이터 : 146,893개  
KW사전 : 61개  
MPQA사전 :  
긍정 : 2,005개, 부정 : 4,783개

- feature(특정기능)를 정의.
- Review 내의 feature명을 통일.
- feature가 언급된 리뷰만 sentence 추출

feature(특정기능) 사전 : 10가지  
feature가 언급된 리뷰 : 367,553개

- 특정 표현, 혹은 특정 품사에 해당하는 리뷰들에 CS 라벨링.
- CS 분류된 리뷰에서 타겟 제품이 언급된 데이터를 타겟 제품명으로 라벨링.

CS Sentence : 15,644개  
SS Sentence : 148,954개

- MPQA 사전을 기반으로, 리뷰의 긍정, 부정 감성을 분류하여 1(긍정), 0(중립), -1(부정)으로 라벨링.

긍정 리뷰 : 75,300개  
부정 리뷰 : 29,474 개  
중립 리뷰 : 59,824개

- 극성 분류를 토대로 가중치를 매긴 매트릭스를 만든다
- pRank알고리즘을 통해 순위를 추출한다.

# 1 데이터 수집

API에서 메타 데이터와 리뷰데이터 추출하여 병합하는 과정을 진행합니다

## 데이터 추출

아마존 연구 데이터에서  
메타 데이터와 리뷰데이터를  
가져온다



## 특정 데이터 분류

아마존 연구 데이터에서  
카메라에 해당하는 카테고리의  
데이터만을 선택하여 가져온다.  
(단, 리뷰가 10개이하인 제품은  
제외시킨다)



## 데이터 병합

asin코드를 기준으로  
메타데이터와 리뷰데이터를  
머지하여 하나의  
데이터프레임으로 출력한다.



# 아마존 연구 데이터에서 전자기기 리뷰 & 메타 데이터 가져오기

```
1 #전자기기 리뷰 읽기
2 review_df = pd.read_csv('reviews_Electronics.csv', engine='python')
3 review_df
```

7824475	A2R6Q6KJCYSH7	BT008UKTMW	Patrick	[2, 2]	brief Like Keyboard tray has a ...	3.0	mousepad is too small	1343520000	07 29, 2012
7824476	A2IGIABRZ5LAB	BT008UKTMW	PetOwner	[0, 0]	MY keyboard is just what I needed and with my ...	5.0	Underdesk or Table Keyboard	1356739200	12 29, 2012
7824477	A2YZI3C9MOHCOL	BT008UKTMW	Tim Church	[0, 0]	The included template made it easy to install...	5.0	Well Made	1396569600	04 4, 2014
7824478	A322MDK0M89RHN	BT008UKTMW	Tom Lawrence	[2, 4]	The item came sooner than expected and was in ...	5.0	Keyboard Drawer	1413368400	08 15, 2014
7824479	A1MH90R0ADMiko	BT008UKTMW	Tommy A. Bird "wordyotm"	[0, 0]	It's a great little device, especially when yo...				
7824480	A10M2KEFFEQDHN	BT008UKTMW	Wendy	[5, 5]	I have a small desk and this fits perfectly. ...				
7824481	A2G81TMIOIDEQQ	BT008V9J9U	Your Future Ex Husband "Vader Was Framed!"	[0, 0]	There is nothing wrong with the mount, you're ...				
7824482 rows × 9 columns									

← 전자기기 리뷰 데이터(csv)

```
1 #메타 데이터 읽기
2 meta_df = pd.read_csv('meta_ele.csv', engine = 'python')
3 meta_df
```

		GPS Syst...	NOTICE...						
498192	amazon.com/images/I/31oF9oNv...	[[['Electronics', 'Computers & Accessories', 'C...]]	Quatech - 1 Port PCMCIA to DB-25 Parallel Adap...	BT008SXQ4C	Parallel PCMCIA Card 1PORT Epp	{'also_bought': ['B000SR2H4W', 'B001Q7X0W6'], ...}	NaN	23.99	
498193	amazon.com/images/I/21WlrX5f...	[[['Electronics', 'Computers & Accessories', 'C...]]	C2G - 5m Ultima USB 2.0 A Mini B Cble	BT008G3W52	C2G / Cables to Go 5M Ultima USB 2.0 Cable	{'bought_together': ['B0002D6QJO'], 'buy_after...	NaN	18.91	
498194	amazon.com/images/I/41TNAvmf...	[[['Electronics', 'Computers & Accessories', 'C...]]	Keyboard drawer.	BT008UKTMW	Underdesk Keyboard Drawer	{'also_viewed': ['B0002LD0ZY', 'B0002LCZP0'], ...}	NaN	25.54	Felk
498195	amazon.com/images/I/41x-15fR...	[[['Electronics', 'Computers & Accessories', 'C...]]	Garmin USB to R232 Converter CableUSB to RS232...	BT008T2BGK	USB To R232 Converter Cable	{'also_viewed': ['B0007T27H8', 'B00425S1H8'], ...}	NaN	62.31	Ge
498196 rows × 9 columns									

메타 데이터(csv) →

## 특정 데이터만 선택하여 추출한다

- 메타 데이터에서 해당 카테고리에 속하는 데이터만 분류 (디지털 카메라)
- 리뷰가 10개 미만인 제품 제거 (10개 미만은 pRank에 큰 영향을 미치지 않을 것이라 판단하여 제거)

```
#디지털 카메라 카테고리 걸러내는 함수
def cat_dslr(text):
    #해당 카테고리에 속하면 카테고리명을 수정하기 편하게 바꿈. 카테고리를 수정하고 싶으면 이곳에서 업데이트
    if text in ["[['Electronics', 'Camera & Photo', 'Digital Cameras', 'Point & Shoot Digital Cameras']]",
               "[['Electronics', 'Camera & Photo', 'Digital Cameras', 'Digital SLR Cameras']]",
               "[['Electronics', 'Camera & Photo', 'Digital Cameras']]",
               "[['Electronics', 'Camera & Photo', 'Digital Cameras', 'Compact System Cameras']]",
               "[['Electronics', 'Camera & Photo', 'Digital Cameras', 'Medium Format Digital Cameras']]"]:
        return 99999 #99999도 없을 바꾸어 수정하기 쉽게 만들
    return text
```

#리뷰가 10개 미만인 제품일 경우 제거한다->asin 코드 기준으로 슬라이싱 및 리뷰 nan 값 제거

```
def slicing(df):
    cnt_df = pd.DataFrame(df[['asin'], value_counts()]).reset_index().rename(columns={'index': 'asin', 'asin': 'review_count'})
    cnt_df = cnt_df[cnt_df['review_count'] >= 10]
    new_df = df.merge(cnt_df, on='asin')
    del new_df['review_count']
    |
    #nan 값 제거
    new_df=new_df.fillna(999)
    new_df=new_df[new_df['review']!=999]
    return new_df
```

## 제품명을 모델명으로 변환

모델명은 같으나 세세한 차이(바디 칼라 등)가 있는 제품 통일이 목적

- 모델 명의 경우 전체 title에서 띄어쓰기로 구분되어 있음을 확인, 단순히 모델명 양 끝에 띄어쓰기를 넣은 후 탐색

*#제품명을 모델명으로 줄임. 제품 리스트에 존재하지 않을 경우 그 데이터는 그냥 삭제 할 것*

```
def product_modelize(text):  
  
    for model in prod_ls:  
        if " "+model+" " in text.lower():  
            return model  
  
    return '999'
```

'Polaroid PDC-640 0.3MP Digital Camera Creative Kit' → pdc-640

'Agfa ePhoto SMILE 0.2MP Digital Camera'

→ 999



제품명에 모델명이  
없는 리뷰는 삭제

## 데이터 입력

### 전자기기 리뷰 데이터

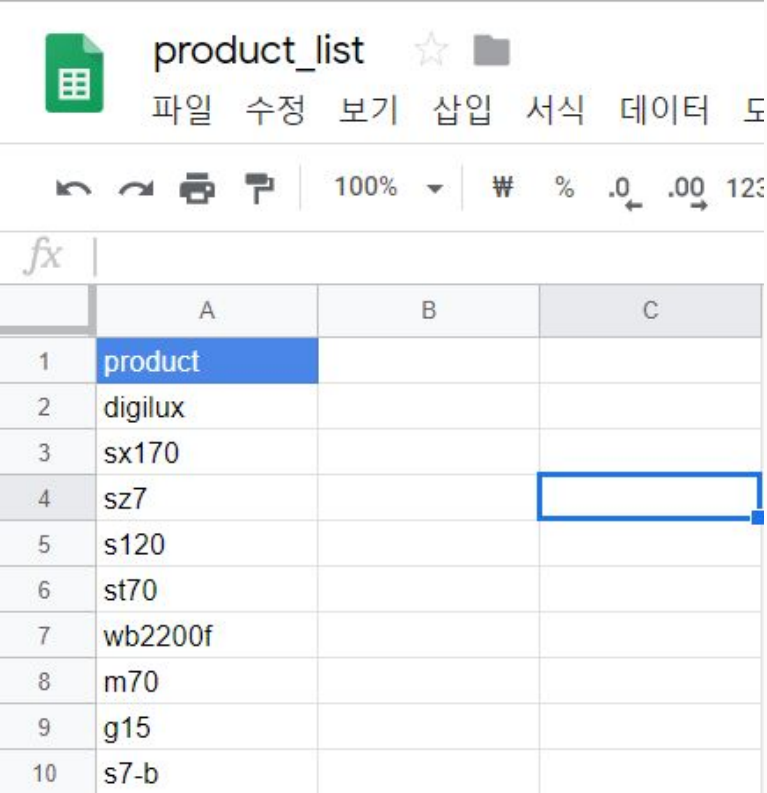
- 7,824,482개의 리뷰

### 전자기기 메타 데이터

- 498,196개의 메타데이터

### 제품 모델명 데이터

- 2,346개의 제품 모델명



product\_list

파일 수정 보기 삽입 서식 데이터 도

100%

fx

	A	B	C
1	product		
2	digilux		
3	sx170		
4	sz7		
5	s120		
6	st70		
7	wb2200f		
8	m70		
9	g15		
10	s7-b		

## 데이터 추출

### 디지털 카메라 리뷰 데이터

- 147,580 개의 리뷰

Unique 제품 개수 1,228개

-> 논문 제품 개수 1,350개

Unique asin코드 갯수 1,605개

Out[13]:

	product_name	asin	review	review_date	score	user
0	cdc-640	B00000J48G	I have really enjoyed using this camera, for a...	07 15, 2000	4.0	"111jbd"
1	cdc-640	B00000J48G	I've had my camera for two months and I love i...	11 18, 1999	4.0	A.J.Kirchoff (xgi@ametro.net)
2	cdc-640	B00000J48G	I bought this camera a few months ago and love...	06 10, 2002	4.0	Andrea Rowland "Andrea"
3	cdc-640	B00000J48G	I bought this as a first time digital camera p...	04 26, 2000	4.0	"denisey65"
4	cdc-640	B00000J48G	This is great value for the price. I was look...	10 11, 2000	5.0	Dennis Frank "djarchivist"
5	cdc-640	B00000J48G	You probobly wont like this review because its...	02 21, 2000	4.0	d
6	cdc-640	B00000J48G	The camera does not work and it did not come w...	07 9, 2011	1.0	Dissatisfied
7	cdc-640	B00000J48G	I recieved this camera for Christmas and was d...	05 18, 2001	2.0	D. Krovitz
8	cdc-640	B00000J48G	I think this is a great digital camera. It is ...	12 4, 1999	5.0	Drew Smith
9	cdc-640	B00000J48G	I have had this camera for about a year and I ...	02 26, 2000	5.0	Edna Tollison "Mama T"
10	cdc-640	B00000J48G	This was the second digital camera I picked up...	10 18, 2003	4.0	Eric McCann
11	cdc-640	B00000J48G	When I got this camera it is nice looking and	08 5, 2000	2.0	Edna Choi

## 2 리뷰 데이터 전처리

feature(특정기능) 사전에 기반하여 Review 내의 feature명을 표준화한 후, feature가 언급된 sentence 추출

### 리뷰 토큰화

디지털 카메라 리뷰만 추출한  
리뷰 데이터를 문장단위로  
토큰화한다



### feature 표준화

정규표현식을 활용하여 각  
feature별 동의어를 사전에  
결정한 단어로 표준화



### Sentence 추출

feature별 랭킹을 산출하는  
논문의 목적에 비추어 분석에  
필요한 feature가 언급된  
sentence만을 사용데이터로 추출

## 리뷰의 문장 토큰화

디지털 카메라 리뷰만 추출한 데이터, 리뷰 개수 147,580개 → 1,120,719개 문장

	product_name	asin	review	review_date	score	user
0	pd-c-640	B00000J48G	I have really enjoyed using this camera, for a...	07 15, 2000	4.0	"111jbd"
1	pd-c-640	B00000J48G	I've had my camera for two months and I love i...	11 18, 1999	4.0	A.J.Kirchoff (xgi@ametro.net)
2	pd-c-640	B00000J48G	I bought this camera a few months ago and love...	06 10, 2002	4.0	Andrea Rowland "Andrea"
3	pd-c-640	B00000J48G	I bought this as a first time digital camera p...	04 26, 2000	4.0	"denisey65"
4	pd-c-640	B00000J48G	This is great value for the price. I was look...	10 11, 2000	5.0	Dennis Frank "djarchivist"

## 리뷰 내에 feature의 동의어를 대표 feature명으로 통일

```
def make_same_feature(text):  
    text=text.lower()  
    # 각각의 feature 별 동의어를 대표명으로 정리  
  
    text=re.sub('resolution|pixel|megapixel', 'pixel', str(text))  
    text=re.sub('lens|wide#sangle|normal#srange', 'lens', str(text))  
    text=re.sub('optical|zoom|optical#szoom|digital#szoom', 'optical', str(text))  
    text=re.sub('memory|megabytes|mb', 'memory', str(text))
```

Many people don't understand sensor size and obsess with megapixels that they'll affectively never use/need.


Many people don't understand sensor size and obsess with pixel that they'll affectively never use/need.



## feature가 언급된 리뷰만 추출

1,120,719개 문장 → 370,068개의 feature가 언급된 문장

```
# feature가 들어있는 문장만 별도로 출력  
# str은 앞의 시리즈 데이터를 문자열로 처리하기 위한 변환식  
# contains는 데이터 내에 패턴이 존재하기만 하면 위치에 상관없이 True 값 출력  
df = df[df.review.str.contains('pixel|lens|optical|memory|burst|battery|focus|lcd|compression|flash')] == True]  
df = df.reset_index(drop = True)  
return df
```



if compact size is your highest priority in a 2-pixel camera, the canon batteryshot s100 digital elph is a ...  
the lens is great, the optical is great, the software is great....i've used the olympus 3000 and a kodak ...  
the d-150 is is one of the few digital cameras in it's price class that has optical optical (3x) which is far ...  
the s400 will probably suit most users fine, but if you want to have focus control while still retaining a ...

# 입력 데이터

리뷰 개수 147,580개

	product_name	asin	review	review_date	score	user
0	pdc-640	B00000J48G	I have really enjoyed using this camera, for a...	07 15, 2000	4.0	"111jbd"
1	pdc-640	B00000J48G	I've had my camera for two months and I love i...	11 18, 1999	4.0	A.J.Kirchoff (xgi@ametro.net)
2	pdc-640	B00000J48G	I bought this camera a few months ago and love...	06 10, 2002	4.0	Andrea Rowland "Andrea"
3	pdc-640	B00000J48G	I bought this as a first time digital camera p...	04 26, 2000	4.0	"denisey65"
4	pdc-640	B00000J48G	This is great value for the price. I was look...	10 11, 2000	5.0	Dennis Frank "djarchivist"

## 출력 데이터

각각 sentence feature가 언급된 리뷰 데이터 - 370,068개

370056	B00K7O2DJU	06 20, 2014	dsc-rx100m	i will discuss this more later but by going wi...	5.0	Walt Kurtz
370057	B00K7O2DJU	06 20, 2014	dsc-rx100m	"optical - as mentioned above the optical rang...	5.0	Walt Kurtz
370058	B00K7O2DJU	06 20, 2014	dsc-rx100m	the optical speed is quite good.	5.0	Walt Kurtz
370059	B00K7O2DJU	06 20, 2014	dsc-rx100m	i can see using it a great deal with this came...	5.0	Walt Kurtz
370060	B00K7O2DJU	06 20, 2014	dsc-rx100m	i really like the burst mode on this camera wi...	5.0	Walt Kurtz
370061	B00K7O2DJU	06 20, 2014	dsc-rx100m	when carrying and using the camera the extra w...	5.0	Walt Kurtz
370062	B00K7O2DJU	06 20, 2014	dsc-rx100m	i have a very small lowepro belt case for my o...	5.0	Walt Kurtz
370063	B00K7O2DJU	06 20, 2014	dsc-rx100m	despite 30+ years of photography i am still am...	5.0	Walt Kurtz
370064	B00K7O2DJU	06 20, 2014	dsc-rx100m	if you are an ultraflash traveller who likes g...	5.0	Walt Kurtz
370065	B00K7O2DJU	06 25, 2014	dsc-rx100m	it was the addition of the amazing evf pop-up ...	5.0	Yogi Moore "YogiM"
370066	B00K7O2DJU	06 25, 2014	dsc-rx100m	cool &#34;we-fie&#34; flip-up lcd lcd, a pop-u...	5.0	Yogi Moore "YogiM"
370067	B00K7O2DJU	06 25, 2014	dsc-rx100m	handheld twiflash, anti-burst blur, are all here.	5.0	Yogi Moore "YogiM"

370068 rows × 6 columns

# 3 CS/SS 분류

특정 표현, 혹은 특정 품사가 포함된 리뷰는 CS, 남은 리뷰는 SS 라벨링, 모든 제품명과 리뷰에 대하여 제품명 표준화하여 리뷰에 언급된 제품명은 타겟으로 라벨링하고, 타겟이 없는 CS데이터는 제거함

## CS/SS 분류

kw사전 제작

kw사전의 단어 혹은 비교구문(JJR, RBR, JJS, RBS)가 포함된 문장은 Comparative Sentence(CS)로 라벨링하고, 나머지 문장은 Subjective Sentence(SS)로 분류함



## 제품명 표준화

제품타이틀 데이터를 토큰화하여 일정 갯수(10) 이하의 단어 선별

선별된 단어는 수작업을 통해 제품명 데이터 분류



## 타겟 라벨링

표준화한 제품명이 리뷰에 포함되어 있으면, 타겟칼럼에 해당 제품명 라벨링

타겟이 없는 데이터는 불필요한 데이터로 인식하여 제거

## KW사전 제작

kw - 문장에 비교 키워드가 있는지 확인 할 수 표현

nlTK 모듈에 comparative\_sentences에서 추출

그 중 애매한 표현은 제외시킴

	A	B
1	to_use	not_use
2	beat	POS tag JJR //adjective, comparative
3	inferior	POS tag RBR //adverb, comparative
4	outstrip	POS tag JJS //adjective, superlative
5	Choice	POS tag RBS //adjective, superlative
6	choose	as <word> as //e.g. as good as
7	prefer	both
8	recommend	on par with
9	outperform	one of few
10	superior	same
11	all	either
12	favor	similar
13	defeat	identical
14	twice	equal
15	thrice	equivalent
16	Number one	together
17	more	match
18	like	rival
19	Versus	alternate
20	first	near
21	outdistance	
22	before	
23	double	
24	outsell	
25	nobody	

## CS, SS 라벨링

Comparative Sentence(CS) - 다른 제품과 비교한 리뷰

Subjective Sentence(SS) - 해당 제품만 평가한 리뷰

- KW 사전에 존재할 경우, CS
- 혹은 주요 품사(JJR, JJS, RBS, RBR)가 존재할 경우, CS
- 그 외에는 SS 라벨링(해당 제품을 평가하는 리뷰 일 것이라 판단)

```
#CS / SS 라벨링
kw_list=call_kw()
def tokenizedNtagging(text):
    #kw로 분류
    for kw in kw_list:
        if kw.lower() in text:
            return 'CS'

    #pos로 분류
    from nltk import word_tokenize, pos_tag
    # nltk로 토큰화, 포스태깅
    tokens = word_tokenize(text)
    pos_tokens = pos_tag(tokens)

    count = 0
    for tok, pos in pos_tokens:
        if pos in ['JJR', 'JJS', 'RBS', 'RBR']:
            count+=1

    if count != 0:
        return 'CS'
    else:
        return 'SS'
```

## CS 리뷰에서 타겟 제품이 언급된 리뷰 탐색

리뷰에서 언급된 모든 모델들을 쉼표 형태로 구분된 문자열로 전환

*#리뷰에서 언급된 모델명 탐색*

```
import re
import numpy as np

model_reg = call_model()
def targetmodel_read(text):
    read_ls = re.findall(model_reg, text)
    read_ls = list(set(read_ls))
    if read_ls :
        return ','.join(read_ls).replace(' ', '')
    else:
        return np.nan
```

## 제품명 리스트 제작

- 데이터 타이틀에서 추출한 데이터의 '제품명' 칼럼에서 전처리 과정을 거쳐 제품명 리스트를 만든다
- 수작업으로 상위 브랜드의 제품(모델명)을 인터넷에서 검색하여 제품명 리스트에 추가한다

canon	nikon	fujifilm	sony	Olympus	coopix	JK IMAGING LTD	coleman	Panasonic	Konica-Minolta
300D	F5	X-H1	α33	C-400	SQ	Kodak Pixpro S-1	CVD600	GH4AGC	vectis s 1
350D	F6	X-Pro2	α35	C-400L	S9900	Kodak Pixpro AZ651	CDV700GW	GH4GC	vectis s 100
400D	F-801S	X-T3	α37	C-420L	S9700	Kodak Pixpro AZ526	CDV500HDG	S1	vectis 2000
450D	F90	X-T2	α55	C-500L	S9600	Kodak Pixpro AZ525	CDV400	S1M	vectis weathermatic
500D	F90X	X-T30	α57	C-600L	S9300	Kodak Pixpro AZ522	CDV200	S1R	DiMAGE RD 3000
550D	F80	X-T20	α58	C-620L	S9200	Kodak Pixpro AZ521	CDV100	S1RM	DiMAGE EX
600D	F100	X-T100	α65	C-800L	S9100	Kodak Pixpro AZ501	C9WP	GH5S	DiMAGE 5
650D	F-601	X-E3	α68	C-820L	S9	Kodak Pixpro AZ422	C6WP	G95M	DiMAGE 7
700D	F70	X-E2S	α77	C-830L	S8200	Kodak Pixpro AZ365	C5WP	G95H	DiMAGE 7i
750D	F75	X-A5	α99	C-840L	S8100	Kodak Pixpro AZ362	C3WP	GH5GA	DiMAGE 7Hi
760D	F-401	X100F	α100	C-860L	S800c	Kodak Pixpro AZ361	C30wpz	GH5LGA	DiMAGE A1
800D	F-401S	XF10	α200	D 400	S8000	Kodak Pixpro AZ251	C20WP	G85KGC	DiMAGE A2
100D	F-401X	xp140	α230	C-920	S8000	Kodak Pixpro FZ41	C12WP	G7K	DiMAGE A200
200D	F50	xp130	α290	C-960	S80	Kodak Pixpro FZ43	2V9WP	GH4AGC	DiMAGE Z1
250D	F60	gfx100	α300	C-990	S8	Kodak Pixpro FZ51	2V8WP	GH4GC	DiMAGE Z2
1000D	F65	gfx 50s	α330	C-1000L	S7c	Kodak Pixpro FZ53	2V7WP	FZ2500GA	DiMAGE Z3
1100D	F55	gfx 50r	α350	C-1400L	S710	Kodak Pixpro FZ151		FZ1000GA	DiMAGE Z5
1200D	S2		α380	C-1400XL	S7000	Kodak Pixpro FZ201			DiMAGE Z6
1300D	SP		α390	C-2500L	S700	Kodak Pixpro SL5			DiMAGE Z10
1500D	S3		α450	C-2000	S70	Kodak Pixpro SL10			DiMAGE Z20
4000D	S4		α500	C-2020	S7	Kodak Pixpro SL25			DiMAGE X
D30	S3M		α550	C-2040	S6900	Kodak Pixpro SPZ1			DiMAGE Xi
D60	S3		α560	C-2100	S6400	Kodak Pixpro SP1			DiMAGE Xt
10D	S6000		α580	C-211	S640	Kodak Pixpro SP350			DiMAGE Xs



## 타겟 제품을 검색하기 위한 제품 모델명 리스트 제작

- ('모델명1'|'모델명2') 형태의 단순 정규표현식으로 리스트 변환 (제품 모델 2,346개)
- 리뷰에서 모델명이 띄어쓰기 형태로(' dsc-rx100m ') 언급되는 것을 확인,  
띄어쓰기를 넣은 ('모델명1' | '모델명2') 형태로 수정

## #모델 사전 호출 함수

```
def call_model():
    df=pd.read_csv('product_list.csv', engine='python')
    model_ls=list(df['product'])
    model_reg='( '+' | '.join(model_ls)+' )'
    return model_reg
```

```

04 | p5000 | sd940 | stylus 400 | vx209 | d400 | c-1 | dmc-2r1 | stylus
sp-100ee | fe-25 | d100 | df | x300z | dsc-h9 | vivitar mini | i100
5 | dylax-7x1 | digimax 420 | e30 | xz1 | s1000pj | 7d mark ii | f
e-110 | dmc-ts10 | ai456w-bk | ex-23 | dimage rd 3000 | d-2 | sp | ps3
2 | a7 | vluiu p160 | vluiu s85 | 100d | f100 | ex-s10b | k-1 | mobis
actioncam | ek-gc1002wakoo | sd1400 | dimage v | x700 | d-150 | c20 |
gopro hero3 | v80 | digimax 201 | 750d | c-21 | s52 | s2500h | coolpi
x 5900 | tg-820 | m6 | f50 | dmc-ghik | s810rf | dsc-h50 | wb850f | a8
| a2000i | gx7 | q25wide | h50ex | 6x00 | zr1000 | a5000 | c-5000 | s
t68 | dimage e201 | i14 | a610 | s440 | 28612is | s80 | light field |
sd3500 | f7-12 | g95m | vluiu nv9 | 70d | 80d | d2h | s2600 | jl100 | f9
0 | 1300d | ze4 | e60 | dmc-fs1 | s9900 | sd1 | ex-26 | dsira850 | e
-450 | j1 | v2 | dsc-hx5v | ex-s1 | sz30 | s110 | v20 | dimage z10 | c
-7070 | f800ex | fd92 | stylus 710 | c-800l | kenox ux4 | pixie | d3300
0 | sx150 | s140 | c143 | t51 | sd750 | dimage e323 | vluiu 12 | dscu30
| 200vp3 | d7000 | c12wp | dmc-fh20 | 2v9wp | c-220 | dsc-p5 | nx1100
| st66 | sd900 | dsc-t77 | i100 | dx49 | 2v8wp | jx200 | dscn1 | x100s
| dmc-is8 | jx250 | stylus 9000 | cdl1500 | kenox ux33 | 520 | s4150 | l
100 | g1 | x | kodak pixpro az651 | i840 | dx6340 | fe-5000 | d200 | nv
15 | f1 | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 | 2000

```

## 입력 데이터

## 처리된 리뷰 데이터

- 370,068개의 리뷰 문장

## KW 사전

- 불필요한 단어 제거한 61개 KW  
(논문 KW 126개의 절반 미만)

## 제품 모델명 데이터

- 2,346개의 제품 모델명

product	
0	digilux
1	sx170
2	sz7
3	s120
4	st70
5	wb2200f
6	m70
7	g15
8	s7-b
9	vluu st10
10	l5

	A	B
1	to_use	not_use
2	beat	POS tag JJR //adjective, comparative
3	inferior	POS tag RBR //adverb, comparative
4	outstrip	POS tag JJS //adjective, superlative
5	Choice	POS tag RBS //adjective, superlative
6	choose	as <word> as //e.g. as good as
7	prefer	both
8	recommend	on par with
9	outperform	one of few
10	superior	same
11	all	either
12	favor	similar
13	defeat	identical
14	twice	equal
15	thrice	equivalent
16	Number one	together
17	more	match
18	like	rival
19	Versus	alternate
20	first	near
21	outdistance	
22	before	
23	double	
24	outsell	
25	nobody	

## 출력 데이터

- CS 문장 15,798개
- SS 문장 149,924개
- CSSS 문장 165,722개 (총 문장 기준 14.8%)

```
1 len(df_t[df_t.CS_SS=='CS'])
```

15798

```
1 len(df_t[df_t.CS_SS=='SS'])
```

149924

# 4 극성 분류

리뷰의 성향을 분석하여 긍정, 부정감성을 분류하고 1(긍정), 0(중립), -1(부정)으로 라벨링

## MPQA 사전 제작

온라인상에 업로드 되어있는  
MPQA사전을 활용

## 극성 분석

극성 분석 후 리뷰에  
긍정(1) / 부정(-1) / 중립(0) 분류

## 라벨링 작업

극성 분석한 리뷰에 대해 라벨링  
실시

# MPQA 사전 제작

온라인 상에 업로드 되어있는 MPQA 사전 활용

- POS 단어 2,005개
- NEG 단어 4,782개

In [6]:	1	pos
Out [6]:	<b>pos_MPQA</b>	
	0	abound
	1	abounds
	2	abundance
	3	abundant
	4	accessable
	5	accessible
	6	acclaim
	7	acclaimed
	8	acclamation
	9	accolade
	10	accolades
	11	accommodative

In [7]:	1	neg
Out [7]:	<b>neg_MPQA</b>	
	0	2-faces
	1	abnormal
	2	abolish
	3	abominable
	4	abominably
	5	abominate
	6	abomination
	7	abort
	8	aborted
	9	aborts
	10	abrade
	11	abusive

## 극성 분석

극성 분석 후 리뷰에 긍정(1) / 부정(-1) / 중립(0) 라벨링

```
for token in tokens:
    if token in pos_list:
        pos_count += 1
    elif token in neg_list:
        neg_count += 1

if pos_count > neg_count:
    df.loc[i, 'polarity'] = 1
elif pos_count < neg_count:
    df.loc[i, 'polarity'] = -1
```

Robbert Patrison	SS	NaN	lcd	0
Robbert Patrison	SS	NaN	lens,optical,focus	1
Robbert Patrison	SS	NaN	battery,focus	1
Robbert Patrison	SS	NaN	battery,focus	-1
Ryan Matthews	SS	NaN	focus,flash	0
Ryan Matthews	SS	NaN	flash	-1
Sam M	SS	NaN	lcd	-1
Scott	SS	NaN	lens	0
ScottvBear	SS	NaN	flash	-1

# 입력 데이터

## 처리된 리뷰 데이터

- 총 문장 165,722개

165718	B00K7O2DJU	06 20, 2014
165719	B00K7O2DJU	06 20, 2014
165720	B00K7O2DJU	06 20, 2014
165721	B00K7O2DJU	06 20, 2014

165722 rows × 9 columns

## MPQA 사전

- POS 단어 2,005개
- NEG 단어 4,782개

Out[6]:

pos_MPQA	
0	abound
1	abounds
2	abundance
3	abundant
4	accessible
5	accessible
6	acclaim
7	acclaimed
8	acclamation
9	accolade
10	accolades
11	accommodative

Out[7]:

neg_MPQA	
0	2-faces
1	abnormal
2	abolish
3	abominable
4	abominably
5	abominate
6	abomination
7	abort
8	aborted
9	aborts
10	abrade
11	abusive

# 출력 데이터

문장 극성 분석 데이터

- 긍정 문장 75,828개
- 부정 문장 29,657개
- 중립 문장 60,237개

```
In [11]: 1 len(polarity_data[polarity_data['polarity'] == 1])
```

```
Out[11]: 75828
```

```
In [12]: 1 len(polarity_data[polarity_data['polarity'] == -1])
```

```
Out[12]: 29657
```

```
In [13]: 1 len(polarity_data[polarity_data['polarity'] == 0])
```

```
Out[13]: 60237
```



# 5 pRank

인덱스는 product, 칼럼은 target product인 매트릭스의 각 셀에 가중치를 계산하여 입력하고, 완성된 매트릭스를 토대로 pRank 계산

## 매트릭스 계산

극성분류된 데이터를 product와 target을 기준으로 분류하여 극성에 기반한 가중치 계산

계산된 가중치를 셀에 채워 매트릭스 계산

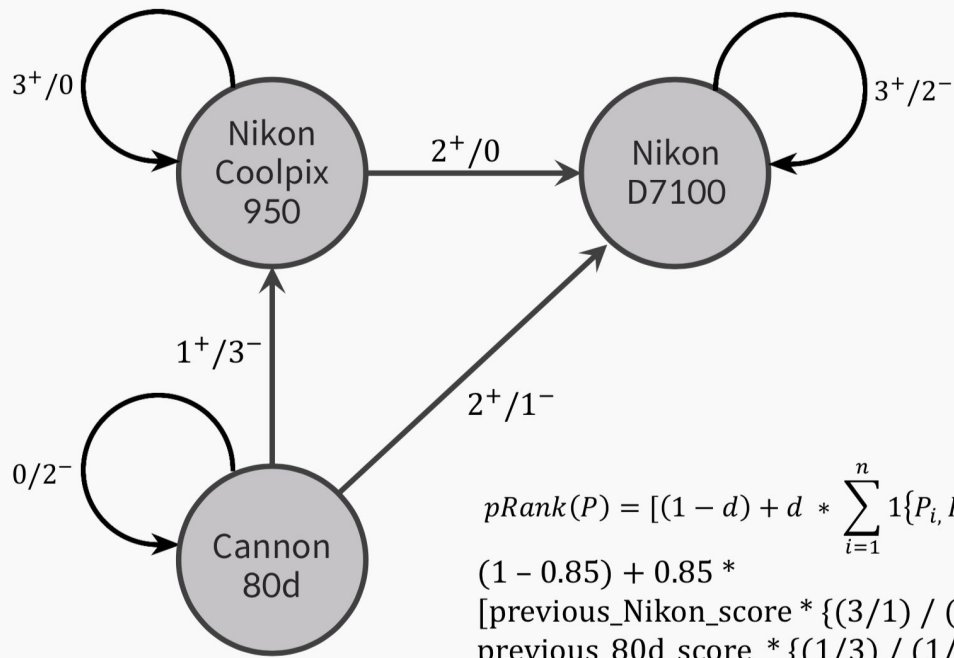


## pRank 계산

zero division error를 회피하기 위해  $\text{sum} = 0$  인 열을 삭제

완성된 매트릭스를 기반으로 pRank 계산

# pRank



## pRank

- google PageRank 기반 알고리즘
- 상품 : 노드, 상품간 비교 : 엣지로 표현
- 리뷰 출처에 따른 엣지의 방향성
- CS : from 비교상품 to 리뷰상품
- SS : from 리뷰상품 to 리뷰상품
- 가중치 = 긍정리뷰 수 / 부정리뷰 수
- 만약부정리뷰가 0개인 경우  
zero division error 회피목적으로 1  
입력

$$pRank(P) = [(1 - d) + d * \sum_{i=1}^n 1\{P_i, P\} * pRank(P_i) * C_e(P_i)] * C_v(P)$$

$$(1 - 0.85) + 0.85 *$$

$$[\text{previous\_Nikon\_score} * \{(3/1) / (3/1 + 2/1)\} +$$

$$\text{previous\_80d\_score} * \{(1/3) / (1/3 + 0/2 + 2/1)\}] *$$

$$(\text{Coolpix\_950\_score}) / (\text{Coolpix\_950\_score} + 80d\_score + D7100\_score)$$

# 유니크한 모델 리스트 기반으로 매트릭스 생성

데이터에서 df['product'].unique() 사용 (1,227개의 제품)

	c-2000	pdc-700	dsc-f505	d-360l	d-460	c3030	qv3000ex	s10	pdr-m60
c-2000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
pdc-700	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
dsc-f505	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
d-360l	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
d-460	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
c3030	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
qv3000ex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
s10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
pdr-m60	0.0	0.0	0.0	0.0					
s20	0.0	0.0	0.0	0.0					
cds4100	0.0	0.0	0.0	0.0					
4700	0.0	0.0	0.0	0.0					
c-3000	0.0	0.0	0.0	0.0					



가중치 입력

	c-2000	pdc-700	dsc-f505	d-360l	d-460	c3030	qv3000ex	s10	pdr-m70	pdr-m60
c-2000	2.0	0.000	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.0	0.000
pdc-700	0.0	2.375	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.0	0.000
dsc-f505	0.0	0.000	9.0	0.000000	0.000000	0.000000	0.000000	0.00	0.0	0.000
d-360l	0.0	0.000	0.0	2.888889	1.000000	0.000000	0.000000	0.00	0.0	0.000
d-460	0.0	0.000	0.0	0.000000	2.666667	0.000000	0.000000	0.00	0.0	0.000
c3030	0.0	0.000	0.0	0.000000	0.000000	2.333333	0.000000	0.00	0.0	0.000
qv3000ex	0.0	0.000	0.0	0.000000	0.000000	0.000000	4.571429	0.00	0.0	0.000
s10	0.0	0.000	0.0	0.000000	0.000000	0.000000	0.000000	2.16	0.0	0.000
pdr-m70									0.9	0.000
pdr-m60										1.125
s20										0.000
cds4100										0.000
4700										0.000
c-3000										0.000

```
# os에 의한 웨이트 계산
for i, t in enumerate(product_list):
    cs_data = df_cs[df_cs['target'] == t]['polarity'].value_counts()
    if len(cs_data.index) == 2:
        matrix.loc[p, t] = int(cs_data[cs_data.index == 1]) / int(cs_data[cs_data.index == -1])
    elif 1 not in cs_data.index:
        matrix.loc[p, t] = 0
    else:
        matrix.loc[p, t] = int(cs_data[cs_data.index == 1])
```

자율 차장

<

## 결과 분석

### 랭크별 제품 개수

- 최대 : flash 1,119개
- 최소 : compression 287개

Rank	pixel	optical	memory	lens	lcd	focus	flash	compression	burst	battery
1	z990	dsc-wx300/r	dmc-fz100	g1 x	sx150	dsc-rx100/b	5d mark iii	dmc-g3	5d mark iii	p510
2	ds1ra350k	sx150	5d mark iii	dmc-fz100	g1 x	nex-6/b	k-01	t4i	dmc-lx5	tl240
3	a810	dmc-fz60	dmc-fh25k	nex-6/b	nex-6/b	5d mark iii	dmc-g5kk	slt-a57	dmc-fz200	dsc-wx300/r
4	dmc-zr3	g1 x	sx150	d-lux5	gr	dmc-g5kk	dmc-fz200	p7700	dsc-w730	dmc-lx5
5	x-a1	dmc-fz200	dmc-lx5	dmc-fz200	h200	dmc-fz200	nex-6/b	nex-6/b	s7000	g1 x
6	dmc-tz4s	z90	dsc-rx100/b	dsc-rx100/b	z90	a7s	f600exr	e-m1	fx100k	elph
7	a3100is	sl1000	dmc-lx7w	sx150	70d	k-01	dmc-fz100	pen e-pl1	z990	sx210is
8	a7s	slt-a57	s6000fd	5d mark iii	p7100	s9700	sx150	70d	slt-a99v	z1015is
9	sd780is	dsc-rx100/b	d600	dmc-lx7w	pen e-pl1	sx120is	dsc-hx9v	g1 x	dsc-wx300/r	c79900
10	gr	nex-6/b	dsc-v1	x-t1	k-01	sd4500	dsc-wx150	dmc-lx5	dmc-zs7	gr
11	f40fd	dmc-lx7w	dmc-zs7	s6000	5d mark iii	dmc-gf1	dsc-rx100/b	x-pro	p310	a1400
12	a495	s7000	sx210is	sd980is	dsc-rx100/b	e-pm2	dmc-lx7w	d5000	slt-a65v	dmc-f2k
13	ex-fh100	h200	nex-6/b	e-m10	sd3500is	dsc-wx150	f40fd	a7r	x-t1	z950
14	p520	dmc-g1	a630	dmc-lx5	z990	dmc-zr3	pen e-pl1	d7100	dmc-fz100	sx170
15	elph	s6	slt-a65v	pen e-pl1	dsc-hx9v	g1 x	a1400	d800	sx120is	dsc-rx100m
16	xf1	dmc-fz100	sx170	z990	sx260	sx210is	sl1000	p5100	dmc-lx7w	x-t1
17	z90	sx160	a620	a530	ex-tr15we	s700	x-t1	dmc-gf5kk	sd4500	z990
18	dmc-fz200	hz30w	a1300	dsc-bx55	j2	dsc-hx300/b	sd4000is	p5000	dmc-g5kk	x-5
19	dsc-rx100/b	sd4500	k-r	dmc-gh1k	l28	dmc-lx7w	g1 x	s1	dmc-g1	ex-zr200
20	p7700	s52	x20	dsc-wx300/r	xc-tl500zbpbu	f850exr	dsc-wx300/r	k-5	p7700	l24
21	dsc-wx300/r	l1	dmc-g1	a1400	s6	dmc-gh4kbody	hz30w	xf1	p7100	x-pro
22	s9300	dsc-rx100m	e-m10	s6	ex-fh100	s6	elph	slt-a99v	sd1400	p310
23	v1233	dmc-zs10	a1400	ec-tl500zbpbus	dsc-rx100m	pen e-pl1	nx300	a400	d5300	s9500
24	sd990is	dsc-w330	dscs85	sx700	a3000	slt-a65v	p7700	nv10	sl1000	s700

## 결과 분석

### Flash 전체 리뷰 중, 17.33%


	결과	품질기반 순위 (전문가 사이트)
1	Canon 5D MIII	Pentax K-01
2	Pentax K-01	Panasonic FZ200
3	Panasonic FZ200	Canon 5D MIII
4	Panasonic FZ100	Panasonic FZ100
5	Canon SX150 IS	Sony HX9V
6	Sony HX9V	Canon SX150 IS
7	Sony WX150	Sony WX150

71.43% 유사

### LCD 전체 리뷰 중, 7.82%

	결과	품질기반 순위 (전문가 사이트)
1	Canon SX150 IS	Canon 70D
2	Canon G1 X II	Fujifilm Z900EXR
3	Ricoh GR	Ricoh GR
4	Sony H200	Canon G1 X II
5	Fujifilm Z900EXR	Nikon P7100
6	Canon 70D	Pentax K-01
7	Nikon P7100	Sony H200

28.57% 유사



## 결과 분석

### 결과 검증

- 리뷰 언급 정도  
Flash : 17.33% , LCD : 7.82%
- 랭킹 순위 정확도  
Flash : 71.43%, LCD : 28.57%

### 정확도 차이의 이유

- 전문적인 사진사들은 별도의 Lens와 Flash 구매  
-> Flash는 사진의 품질에 큰 영향을 미친다.
- LCD : 사진의 구도를 잡는 용도로 사용, 실제 사진 품질에 기여하는 영향이 적다.

### 분석

- Flash는 품질에 기반한 소비자평가가 많아 리뷰기반/품질기반 순위 정확도가 높고,  
LCD는 틸트(화면 각도 조절 기능)와 같은 부가 기능이 구매에 영향을 끼치기에  
LCD자체에 대한 평가가 적어 순위 정확도가 낮다.

# Insight

## 연구 한계

1. 수집한 리뷰데이터에 존재하는 제품명만을 추출하여 사용함. 신제품에 대한 분석이 불가능하다.
2. 데이터 수의 부족으로 약 16년의 데이터를 일괄적으로 처리하여, 시계열에 따른 데이터 분석이 불가능하다.
3. Domain knowledge의 부족으로 모델명에 대한 동의어 처리를 하지 못함. 리뷰데이터의 분석 성능 저하.

## 후속 연구방향

1. 브랜드별로 모델명을 작성하는 방식이 동일한 점을 활용하여 브랜드별 정규표현식을 활용하여 모델명을 추출하는 코드를 추가 / 신제품에 대한 분석 가능성
2. 데이터의 수집 경로를 확대 / 시계열에 따른 데이터 변화를 분석함으로써 회사의 마케팅 방향 설정에 사용 가능성
3. 모델명의 동의어 처리를 통해 분석 성능 향상



# 참고자료 및 분석도구

## 참고 자료

- Kunpeng Zhang. Voice of the Customers- Mining Online Customer Reviews for Product Feature-based Ranking

- Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine

## 분석 도구

언어 : Python

개발환경 : Jupyter Notebook, VS code

라이브러리 : Pandas, Numpy, Nltk, re

## 사이트 자료

아마존 연구 데이터 : <http://jmcauley.ucsd.edu/data/amazon/>

MPQA 긍정/부정 사전 : <https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107>

검증 비교 사이트 : <https://cameradecision.com>

감사합니다