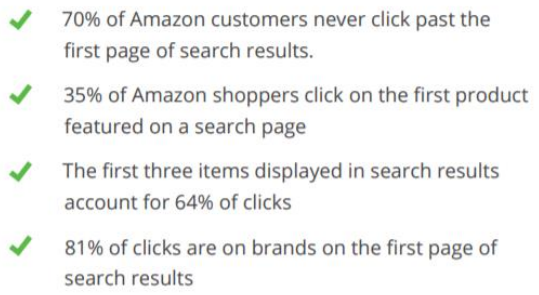
# AMAZON PRODUCT LISTING OPTIMIZATION

## 아마존 리스팅 최적화를 위한 텍스트 분석

SNU 4차산업혁명 아카데미
빅데이터 애널리틱스 양서윤

# 주제 선정 배경

- 초국가적인 온라인 전자상거래가 발달하면서 국내 기업의 아마존 진출이 활발해지고 있다.
- 아마존 소비자들의 70%가 검색 결과 1 페이지에서 구매 결정을 내리는 만큼, 자사 제품이 검색 결과 상위에 노출될 수 있는 방안을 마련하는 것은 아마존에 진출하는 신규 브랜드에게 매우 중요한 과제다.
- 상품의 리스팅 최적화, 즉 소비자가 아마존 검색 결과를 통해 원하는 정보를 얻고 구매하는 데 최적화된 리스팅(이름, 가격, 설명, 리뷰 등)을 작성하는 것은 검색 결과 상위 노출을 위한 제 1 전략으로 알려져 있다.




▲ 아마존 고객의 상품 검색 후 클릭에 관한 조사 결과

출처 : http://learn.cpcstrategy.com/rs/006-GWW-889/images/2018-Amazon-Shopper-Behavior-Study.pdf

*키워드 항목 좋은 기입의 예*

| 상품 | 사용 가능한 키워드 | 이유 |
|---|---|---|
| Lord of the Rings | lotr | 유명한 영화, 책의 경우 고객들은 해당 영화, 책 이름의 약어를 검색할 가능성이 높습니다. |
| TON 3057 20-oz. Jacketed Fiberglass Claw Hammer | nail pounder, nail puller, ripping tool | 고객들은 Claw Hammer(장도리)의 동의어인 Nail pounder, nail puller, ripping tool 등으로 검색할 가능성이 높습니다. |
| Harry Potter and the Deathly Hallows | Hermione Granger, Ron Weasley | 고객들은 책이나 영화에 나오는 등장인물의 이름을 검색할 가능성이 높습니다. |
| 2015 Fox Racing 180 Drezden Pants | Racewear, motocross gear | 고객들은 Racing Pants 의 대체어가 될 수 있는 Racewear, motocross gear 등을 검색할 가능성이 높습니다. |
| The Shadow Beast (An Adventure in Time) | Time travel | 고객은 찾고자 하는 책의 주제(=장르)를 검색할 가능성이 높습니다. |

▲ 아마존 코리아에서 제공하는 리스팅 최적화 예시

출처 : Amazon Services Korea

# 주제 설명과 선정 이유

- 본 프로젝트의 목적은 특정 상품의 리스팅 최적화를 위한 실행 방안을 제안하는 것이다. 따라서 본 프로젝트에서는 하나의 판매 상품을 가정하고, 해당 상품의 리스팅 최적화를 위한 텍스트 분석을 실시하고 구체적인 실행 방안을 제안하는 것을 목표로 한다.

- 이를 위해 상품이 속한 카테고리의 검색 결과 1~10페이지에 노출된 상품들의 상품명, 가격, 리뷰 텍스트를 분석한다. 이를 통해 상위 노출 상품들의 리스팅 특징들을 파악하고, 실제 실현 가능한 리스팅 전략을 제공하고자 한다.

- 이 프로젝트는 작성자 본인이 뷰티 스타트업 근무 당시 아마존 상품 리스팅 최적화를 경험한 내용을 바탕으로 설계되었다.


Earth's Recipe Waterful Sun Gel 50ml SPF 50+ PA+++ Facial Sunscreen with Harrogate Sparkling Water, Rich Watery Essence Broad Spectrum Moisturizing Sunblock Oil-Free Light Aqua Fragrance Non-Tinted

by Earth's Recipe

$25.00 ($14.79/Fl Oz)

prime

FREE Shipping on eligible orders

More options available:

$24.00 Other Sellers

66

Product Features

... ✔ EVERYTHING YOU WISHED IN A *SUNSCREEN* - Formulated with Britain's ...

▲ 작성자가 근무한 스타트업에서 판매하는 자외선 차단제, 현재 아마존 sunscreen 검색 결과 15 page에 리스팅 되어 있다.

# 분석 내용 및 소스 설명

- 판매 상품을 자외선 차단제로 가정한다. 자외선 차단제 검색 키워드인 sunscreen의 검색 결과 1~10 페이지에 노출된 상품명, 가격, 리뷰를 크롤링해 텍스트 분석을 실시한다.

- 10개 페이지 상위 노출 상품들의 상품명을 단어 단위로 쪼개 빈도수를 분석하고, 워드클라우드와 막대 그래프로 단어 출현 빈도 수를 시각화 한다.

- 10개 페이지 상위 노출 상품들의 가격대를 수집하여 평균, 분산, 중앙값, 분위수, 이상치 등의 통계량을 확인하고, 상자 그림을 통해 주 가격대의 범위를 시각화하고, 가격대가 어떠한 형태로 분포로 구성되어 있는지 히스토그램과 밀도 함수로 확인한다.

- 상위 노출 10개 상품을 선정하여(동일 브랜드 제품 제외) 각 상품의 10개 페이지의 리뷰를 수집, 출현 빈도수가 높은 단어를 워드 클라우드로 시각화 한다.

- 분석 도구로 R을 사용하였으며, 미국 아마존 웹사이트(https://www.amazon.com)에서 데이터를 수집했다.

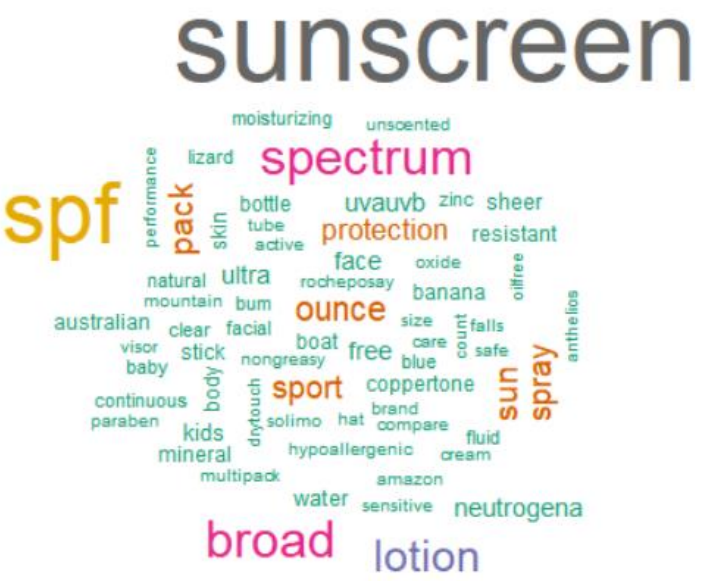
# 1. 상품명(Title) 분석 – code

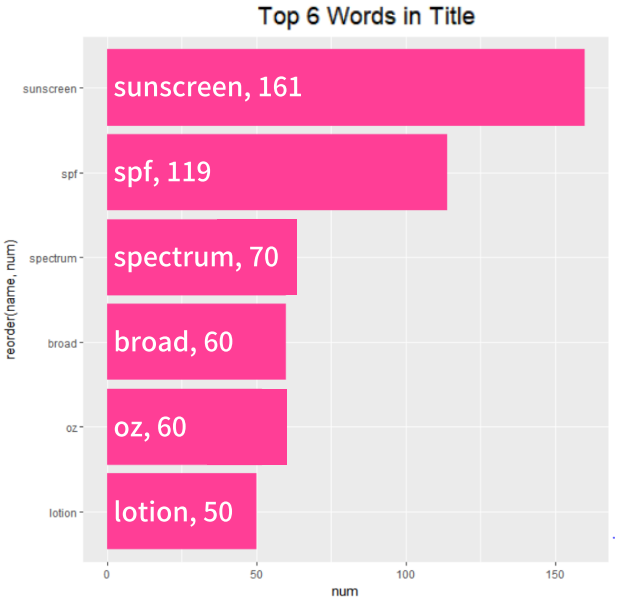
```r
# 패키지 가져오기
library(rvest)

# 크롤링으로 1~10 페이지 가져오기
base_url_1 = "https://www.amazon.com/s/ref=sr_pg_"
base_url_2 = "?rh=i%3Aaps%2Ck%3Asunscreen&page="
base_url_3 = "&keywords=sunscreen&ie=UTF8&qid=1532086589"
total_name <- data.frame(name=character(), stringsAsFactors=FALSE)
total_price <- data.frame(price=character(), stringsAsFactors=FALSE)
total_link <- data.frame(link=character(), stringsAsFactors=FALSE)

for (i in 1:10){
  url = paste0(base_url_1, i, base_url_2, i, base_url_3)
  print(url)
  html_sun = read_html(x = url, encoding = 'UTF-8')
  name_sun <- html_nodes(html_sun, ".a-link-normal") %>% html_attr("title")
  name_sun <- na.omit(name_sun)
  price_sun <- html_nodes(html_sun, ".a-offscreen") %>% html_text()
  price_sun <- gsub(pattern="[Sponsored]", replacement = NA, price_sun)
  price_sun <- na.omit(price_sun)
  link_sun <- html_nodes(html_sun, ".a-size-small.a-link-normal.a-text-normal")
  %>%  html_attr("href")
  total_name <- rbind(total_name, name_sun, stringsAsFactors=FALSE)
  total_price <- rbind(total_price, price_sun, stringsAsFactors=FALSE)
  total_link <- rbind(total_link, link_sun, stringsAsFactors=FALSE)
  cat(i, "\n")
}
```

```r
# total_name의 벡터 변환, unique 값 추출, 파싱
unlist_name <- tolower(unlist(total_name))
unq_unlist_name <- unique(unlist_name)
split_name <- strsplit(unq_unlist_name, split = ' ')
split_name_vec <- unlist(split_name)
split_name_vec <- gsub("[.|!|,|(|)|&|-]", "", split_name_vec)

# total_name의 워드클라우드
library(wordcloud)
library(RColorBrewer)
wordcloud(split_name_vec, min.freq=4,
          colors=brewer.pal(8, "Dark2"))

# total_name의 barplot
split_name_vec_sort <- sort(table(split_name_vec), decreasing=T)
split_name_vec_bar <- unclass(split_name_vec_sort)
split_name_vec_bar_1 <- split_name_vec_bar[8:19]
barplot(split_name_vec_bar_1,  main = "7th~18th Words in Title",
        xlab="words", ylab = "numbers", col="springgreen3")
split_name_vec_bar_2 <- split_name_vec_bar[20:31]
barplot(split_name_vec_bar_2, main = "19th~30th Words in Title",
        xlab="words", ylab = "numbers", ylim=c(0,40),
        col="springgreen3")
```

- rvest 패키지를 사용하여 아마존 사이트 sunscreen 검색 결과 상위 10개 페이지의 상품명을 수집했다.
- 상품명(특히 브랜드명)이 중복될 가능성이 있으므로 unique 값을 뽑아낸 후, space 단위로 상품명을 분리하여 단어 형태로 추출, gsub 함수로 특수 문자를 정리했다.
- wordcloud와 RColorBrewer 패키지를 사용하여 단어 별 출현 빈도를 단어의 크기와 색으로 시각화 했다.
- ggplot2 패키지를 사용하여 출현 빈도 순서대로 1위~6위, 7위~18위, 19위~30위 데이터를 각각 추출해 막대 그래프로 시각화 했다.

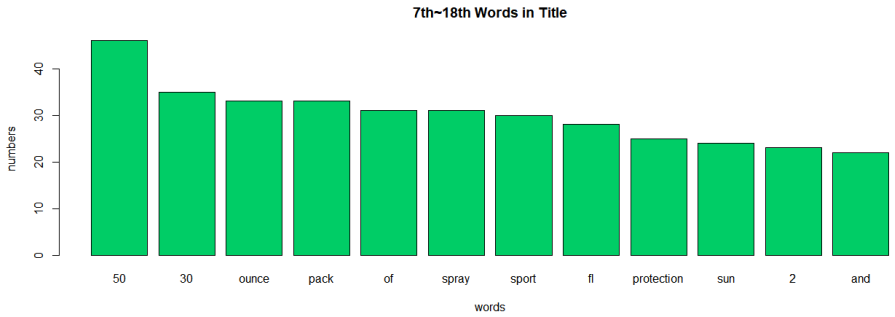
# 1. 상품명(Title) 분석 – wordcloud

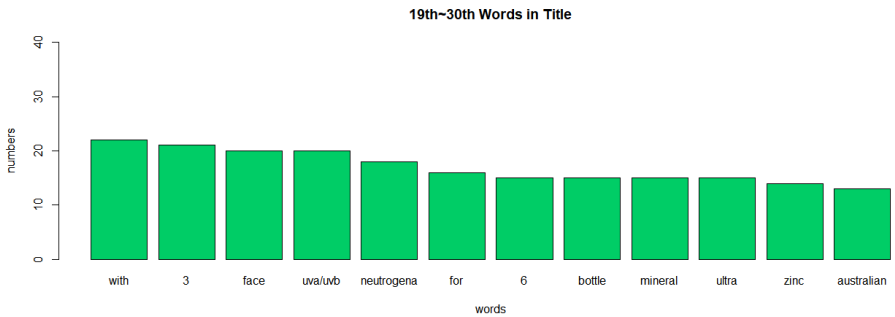



- 총 158개의 상품명의 2232개 단어 중 출현 빈도가 높은 상위 6개 단어는 sunscreen, spf, spectrum, broad, oz, lotion 이다. 검색 키워드인 sunscreen은 상품명 개수만큼 빈도 수가 확인되는 것으로 보아, 거의 모든 자외선 차단제 상품명에 반드시 들어가는 키워드라는 추측도 가능하다. 검색 결과에 노출되기 위해서는 상품이 검색 키워드를 포함하고 있어야 한다는 상식을 위 데이터로도 한 번 더 확인할 수 있다.
- 상위 2~4위 키워드를 차지하고 있는 단어들은 제품의 주요 기능인 자외선 차단 수치를 나타내는 키워드다. 자외선 차단제의 경우 검색 키워드 만큼이나 차단 지수인 spf에 대한 명시도 중요하다고 할 수 있다. 특히 3, 4위인 spectrum과 broad는 “broad spectrum”이라는 하나의 문구로 많이 사용되며, 차단 범위가 넓은 차단제가 인기가 있음을 알 수 있다.
- 5위까지는 거의 대부분의 자외선 차단제에 해당하는 키워드였다면, 6위는 특정한 제형을 나타내는 lotion이라는 점이 흥미롭다. 키워드에 상위 노출되는 제품들은 lotion 제형이 많으니, 판매하고자 하는 제품이 lotion 형태라면 상품명에 명시하는 것이 좋겠다.

# 1. 상품명(Title) 분석 – barplot




- 상위 7위와 8위는 각각 숫자 50과 30을 차지하고 있는데, 이는 가장 일반적인 자외선 차단 지수 숫자로, 단어 “spf”와 함께 쓰인다. 그 외에도 앞에서 살펴본 키워드와 관련하여, 중량을 나타내는 ounce와 fl의 출현 빈도 수가 높은 것을 알 수 있다. 카테고리명인 “sunscreen”, 차단 지수, 중량 이 세 키워드는 대다수의 상위 노출 상품들이 포함하고 있는 중요 키워드라고 할 수 있다.
- 특정한 기능을 명시하는 키워드(sport)도 상위 15위권 안에 등장했다. “sport” 단어가 붙은 제품은 차단 효과가 강력한 야외 활동용 자외선 차단제를 말하며, spf 50을 초과하는 강력한 차단 제품도 전체 카테고리 안에서 인기가 높다는 것을 알 수 있다.
- 제품의 형태와 구성에 대한 키워드(spray, bottle, pack)도 30위권 안에서 발견할 수 있다. 제품의 구성과 형태를 상품명에 명시할수록 해당 제품을 원하는 소비자들이 상품을 클릭할 것이고 더불어 판매에도 긍정적인 영향을 미칠 것이라 유추해볼 수 있다. 특히 단어 pack의 출현 빈도가 높은 것으로 보아, 세트로 구성된 상품이 꽤 많고 잘 팔리고 있음을 추측해볼 수 있으며, 세트 상품일 경우 세트 개수 등의 정보 등을 상품명에 명시하는 것이 좋겠다.
- 자외선 차단제의 성분 키워드 또한 30위권 안에서 발견되었다. 피부 방어력을 강화하는 미네랄(mineral) 성분과, UVA/UVB를 둘 다 차단하며 차단력이 우수한 성분으로 알려진 징크옥사이드(zinc)도 상품 클릭과 구매에 유의미한 영향을 미치는 단어라고 볼 수 있다.

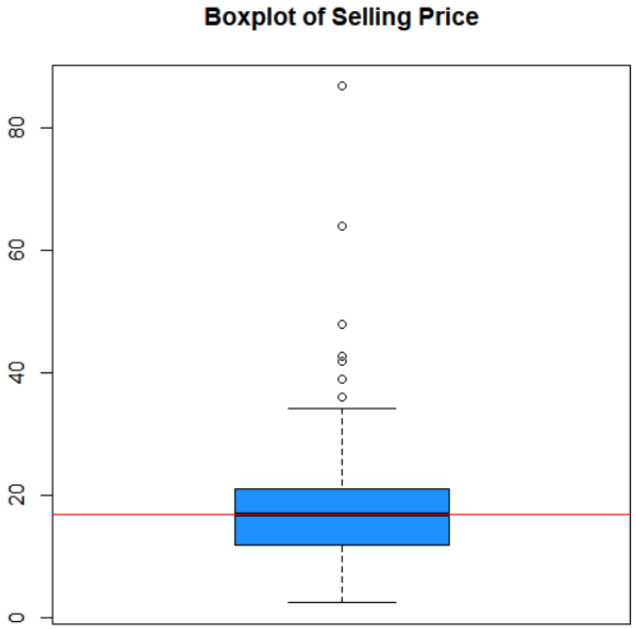
# 2. 가격대(Price) 분석 – code

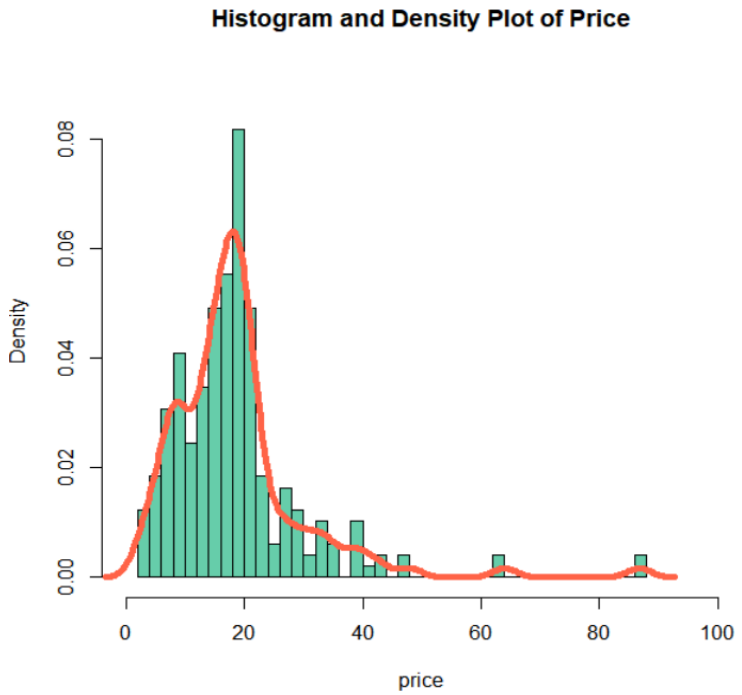
```
# total_price의 벡터 변환, 파싱, 단일 가격 추출, 기술 통계량, plots
unlist_price <- unlist(total_price)
unlist_price <- gsub('\\$','',unlist_price)
unlist_price <- Filter(function(x){nchar(x)<=5}, unlist_price)
unlist_price <- as.numeric(unlist_price)
summary(unlist_price)
med = median(unlist_price)
boxplot(unlist_price, main="Boxplot of Selling Price", col="dodgerblue")
abline(h=med, col="red")
hist(unlist_price, main="Histogram and Density Plot of Price",
     xlim=c(0,100), ylim=c(0,0.09), freq=FALSE, xlab="price",
     breaks=50, col="mediumaquamarine")
?hist
lines(density(unlist_price), col="tomato1", lty=1, lwd=5)
```

```
> summary(unlist_price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.49   10.11   16.99   20.77   26.87   86.75
```

- 아마존 웹사이트 검색창에 sunscreen 검색 후 결과로 나온 상위 10개 페이지의 상품 가격을 크롤링했다.
- 가져온 데이터에서 gsub 함수로 $ 달러 표시를 없애고, filter 함수를 사용해 단일 가격을 추출했다.
- (원활한 계산을 위해 14.99 – 16.99와 같이 range가 있는 데이터는 제외시켰다)
- summary 함수를 사용해 가격(price)의 기술 통계량을 파악하고, boxplot을 그려 이를 시각화 했다.
- hist 함수를 사용해 가격(price)을 x축으로 하는 히스토그램을 그리고 lines 함수로 밀도함수를 그렸다.

# 2. 가격대(Price) 분석 – plots




- sunscreen 검색 결과 상위 10페이지에 노출된 상품들의 가격의 평균은 20.77달러다.
- 전체 가격 데이터의 50%(IQR, 사분위수범위)가 10.11달러와 26.87달러 사이에 분포한다.
- 히스토그램과 밀도함수로 보아 20달러 근방으로 가격이 집중되어 있음을 알 수 있다.
- 제품의 밸류 포인트에 따라 가격은 달라질 수 있지만, 많이 팔리는 상위 노출 제품의 경우 20달러 전, 후로 판매되고 있다는 사실을 염두에 두고 가격을 책정하는 것이 좋겠다.

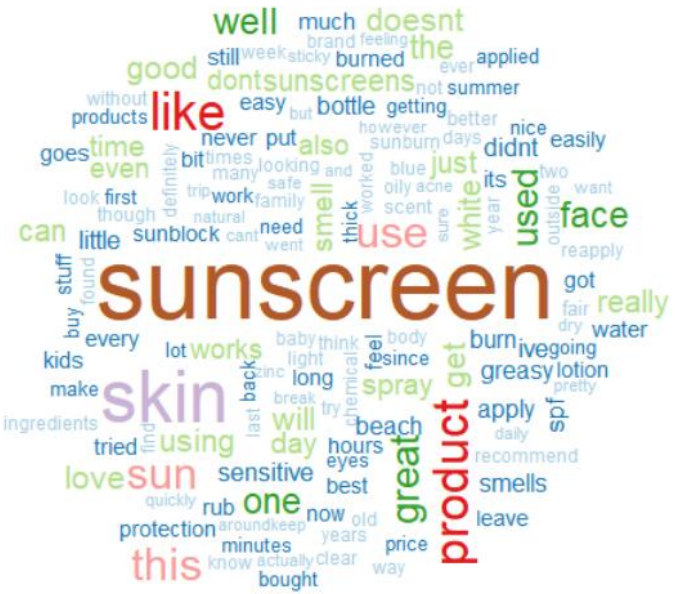
# 3. 상품리뷰(Review) 분석 – code

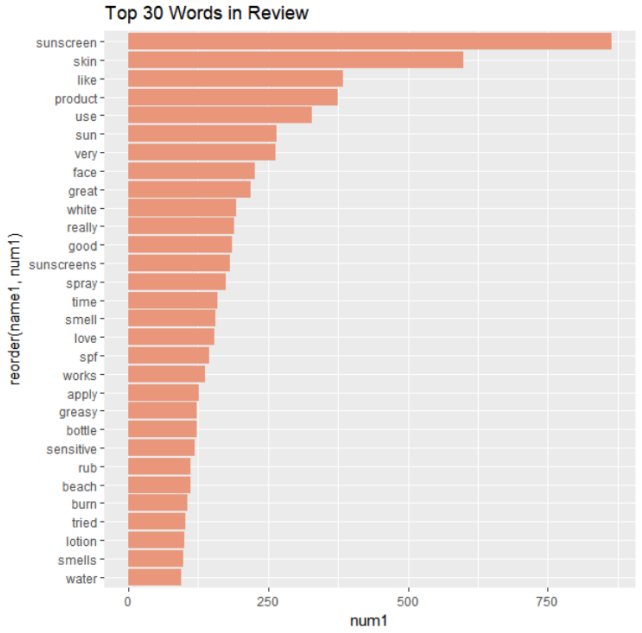
```
# 1 : Sun-Bum-Moisturizing-SPF-Hypoallergenic
base_url_1 = "https://www.amazon.com/Sun-Bum-Moisturizing-SPF-Hypoallergenic/
product-reviews/B004XGPMFA/ref=cm_cr_arp_d_paging_btm_"
base_url_2 = "?ie=UTF8&reviewerType=all_reviews&pageNumber="
total_rev <- data.frame(review=character(), stringsAsFactors=FALSE)

for (i in 1:10){
  url = paste0(base_url_1, i, base_url_2, i)
  print(url)
  html_rev = read_html(x = url, encoding = 'UTF-8')
  review <- html_nodes(html_rev, "span.a-size-base.review-text")
  %>% html_text()
  total_rev <- rbind(total_rev, review, stringsAsFactors=FALSE)
  cat(i, "\n")
}
total_rev
names(total_rev) = c(1:10)
unlist_rev <- tolower(unlist_rev)
unlist_rev <- unlist(total_rev)
split_rev_vec <- strsplit(unlist_rev, " ", fixed=TRUE)
final_rev_1 <- unlist(split_rev_vec)
final_rev_1 <- gsub("[.|!|,|-|(|)|:|&|;|!|@|^|+]", "", final_rev_1)
```

```
# 10개 제품의 리뷰들 하나의 벡터로 합치고 sort
total_review <- c(final_rev_1, final_rev_2, final_rev_3, final_rev_4,
                  final_rev_5, final_rev_6, final_rev_7, final_rev_8,
                  final_rev_9, final_rev_10)
sorted_total_review <- sort(table(total_review), decreasing=T)
sorted_total_review[1:100]
sorted_total_review[101:200]

# total_name의 워드클라우드
library(wordcloud); library(RColorBrewer)
wordcloud(split_name_vec, min.freq=5,
          colors=brewer.pal(8, "Dark2"))
```

- 아마존 웹사이트 검색창에 sunscreen 검색 후 결과로 나온 상위 10개 상품의 리뷰 10 페이지, 총 100 페이지를 크롤링 했다. (1위~10위 상품 선택 시 브랜드 등이 중복될 가능성이 있어 서로 다른 상위 10개 상품의 데이터를 사용하기로 한다)
- 가져 온 각각의 리뷰 데이터를 space 단위로 단어를 분리하여 단어 추출하고, 문자를 소문자화 하고, gsub 함수로 특수 문자를 정리했다. 각 상품의 리뷰 10개 페이지를 크롤링한 결과를 하나의 벡터로 결합하고, 내림차순으로 정렬했다.
- wordcloud 패키지를 사용하여 단어 별 출현 빈도를 시각화하고, RColorBrewer와 ggplot2를 사용하여 barplot을 그렸다. 인덱싱을 사용해 리뷰에서 출현 빈도 수가 높은 상위 200개 단어를 추출했다.

# 3. 상품리뷰(Review) 분석 – wordcloud



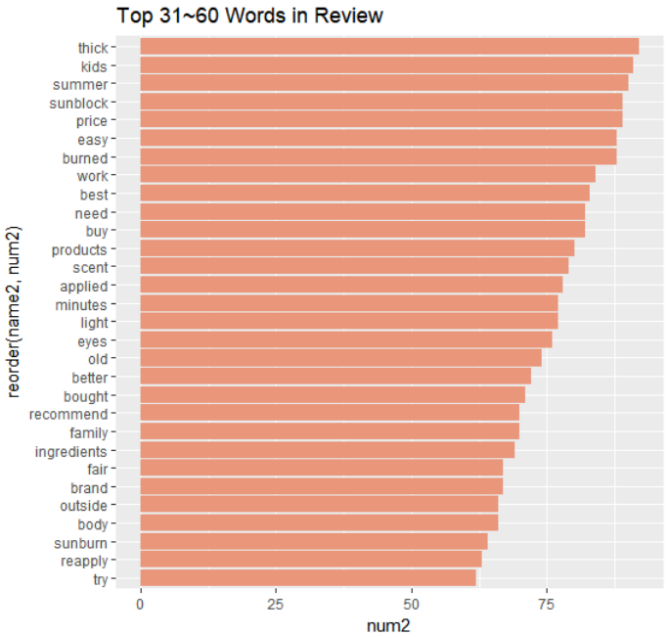




- sunscreen 검색 결과 상위 10개 제품의 리뷰 10 페이지, 총 100페이지의 리뷰 텍스트를 워드클라우드로 표현했다. 그 중에서 I, the, a, this, is와 같은 대명사, 조사 등은 생략하고 유의미한 단어만 추출해 출현 빈도가 높은 30개 단어를 막대 그래프로 시각화했다.

- 유의미한 단어만 추출해 정렬한 우측 막대 그래프에서 상위 5개 단어는 sunscreen, skin, like, product, use로 나타났다. 흥미로웠던 점은 이 상위 5개 단어로 sunscreen 리뷰의 주제(topic)를 표현할 수 있다는 것이다. 한 마디로 주제는 “How I like sunscreen(product) when using on skin”이라고 할 수 있다.

- 유의미한 단어만 추출해 정렬한 상위 60개의 키워드를 살펴 본 결과, 앞에서 추출한 상품명(title) 키워드와 중복으로 출현하는 단어들을 발견할 수 있었다. 상품명과 리뷰 모두 검색 키워드인 sunscreen의 노출 빈도 수가 압도적으로 높았으며, 상품명 상위 30개 안에 들었던 단어 spf, lotion, spray, sun, face, bottle이 리뷰 텍스트에도 동일하게 등장했다. 해당 단어들은 실제 소비자들 사이에서 많이 회자되는 단어들인 만큼 브랜드에서 상품 리스팅 곳곳에 효과적으로 사용하면 좋겠다.

# 3. 상품리뷰(Review) 분석 – plots




**Key buying factor** : price, product, brand, ingredients

**Feature** : white, sensitive, scent, thick, light, greasy

**Response** : like, very, good, great, love, best, better, fair

**People** : kids, family

**Environment** : sun, beach, summer, outside

**Act** : smell, recommend, try, need, work, apply, buy

- 좌측 막대 그래프는 리뷰 텍스트 상위 30~60위까지의 키워드를 정렬한 것으로, 상위 1~30위 키워드보다 좀 더 구체적이고 다양한 단어들이 많이 발견되었다. 총 60개의 키워드를 확인한 결과, 비슷한 키워드들을 추려 여러가지 주제들로 묶을 수 있었고 그 결과는 우측의 그림과 같다. 총 6개의 주제 key buying factor, feature, response, people, environment, act로 분류해 보았다.
- 상위 키워드들을 하나의 주제로 묶어보니, 소비자들이 무엇을 말하고 있는지 그들의 목소리가 좀 더 명확해졌다. feature 카테고리를 보면 소비자들이 자외선 차단제의 주요 특징으로 민감성, 백탁 현상, 향 등을 언급하고 있으며, 해당 특징들을 상품 설명과 같은 리스팅에 명시한다면 그들이 원하는 제품임을 강조하고 어필할 수 있기 때문에 판매에 있어 긍정적인 효과가 예상된다.
- 리뷰는 소비자의 니즈를 확인할 수 있는 중요한 데이터다. 분류 결과를 보면, 제품을 사용하는 사람으로 어린이와 가족이 주로 언급되고 사용 환경으로는 여름, 해변, 외출 등의 키워드가 언급되는 것을 알 수 있다. people이나 environment의 키워드들을 소비자들이 가장 먼저 접하는 상품명에 넣어주는 것도 하나의 날카로운 전략이 될 수도 있겠다. 예를 들어, 상품명에 다음과 같은 단어들을 추가할 수 있다(“for kids”, “family sunscreen”, “sunscreen for summer beach”).

# 결론 및 제안

- 본 프로젝트에서는 자외선 차단제를 하나의 판매 상품으로 가정하고, r을 이용한 텍스트 분석을 통해 아마존 리스팅 최적화 방안을 살펴보았다. 아마존 sunscreen 검색 결과 상위 10개 페이지에 노출된 제품들의 리스팅을 분석하여 자사 제품에 적용할 수 있는 방법들을 알아보았다. 정리한 결론은 다음과 같다.

- **1. 상품명** : 검색 결과 상위 10개 페이지에 노출된 상품명의 단어 중 출현 빈도가 높은 키워드를 살펴본 결과, 검색 키워드인 “sunscreen”, 차단 지수, 중량은 기본적으로 상품명에 기재하는 것을 권장한다. 그리고 상품명 상위 키워드 중 자사 제품에 해당하는 구체적인 키워드가 있다면(예를 들어 제형을 나타내는 lotion), 상품명에 명시하는 것도 좋은 전략이 될 수 있다.

- **2. 가격** : 상위 10개 페이지에 리스팅 된 제품 가격의 평균은 20.77 달러로, 평균 전후로 가격대가 밀집된 분포를 보인다. 20달러를 기준으로 저렴한 제품을 선호하는 고객을 타겟 하려면 이보다 조금 낮게, 프리미엄 제품으로 중고가의 제품을 선호하는 고객을 타겟 하려면 그보다 조금 높은 가격을 책정하는 것이 좋겠다.

- **3. 리뷰** : 리뷰는 소비자의 니즈를 확인하여 제품 성장의 인사이트를 얻을 수 있는 중요한 자산이다. 상위 노출되는 제품의 리뷰에서 소비자들이 자주 언급하는 키워드들은 브랜드 측이 상품 페이지에서 한 번 더 짚어 줌으로써 그들이 원하는 제품임을 강조하고 어필할 수 있다. 더불어 자사 구매 고객에게 리뷰를 요청할 때, 상위 노출 제품의 리뷰 키워드를 함께 제공하여 해당 제품과 동일한 효과를 기대할 수 있을 것이라 생각한다.

- 아마존은 리스팅 최적화를 위한 팁을 일부 제시할 뿐, 정확히 어떤 기준과 알고리즘으로 검색 키워드 상위 노출을 결정하는지는 밝히지 않는다. 다만 판매자들과 대행 업체들이 키워드 상위 노출된 제품들의 리스팅 특징들을 분석하고, 그 결과를 자신의 제품에 적용하는 방법으로 리스팅 최적화를 진행하고 있다. 본 프로젝트에서 분석한 내용도 정확한 상위 노출 알고리즘을 알지 못해 추측에 불과하다는 지적을 받을 수 있다. 그러나 작성자 본인의 경험에 비추어 봤을 때 위와 같은 리스팅 최적화는 효과가 있으며, 해당 자료는 트렌드 분석과 신제품 개발 시 유용한 자료로도 활용될 수 있다.