TERMProject_201721378

201721378 이유리

2019 6 18

연구문제 도출

- 본 연구는 기업의 새로운 홍보 매체로 평가받는 트위터의 활용에 있어, 기업이 속한 제품군에 따라 트위터 상의 메시지 전달 방식이나 정보 내용에 있어 어떠한 차이가 있는지 알아보고자 작성되었다.

- 연구가설 1: 텍스트의 감정 상태와 소비자의 반응은 상관관계를 보일 것이다.
- 연구가설 2: 제품군에 따라 게시글의 커뮤니케이션 유형에 차이가 있을 것이다.
- 연구가설 3: 저관여 제품군의 정보 탐색 경로는 고관여 제품군과 차이가 있을 것이다.

-이에 따라 고관여 제품군에는 미국의 'american airline'을, 저관여 제품군에는 'starbucks'를 선택, 2019 년 05 월 23 일을 기점으로 제출일 하루 전인 19 년 6 월 22 일까지의 크롤링 데이터를 기반으로 작성하였다.

- 데이터의 크롤링과 csv 형성 과정의 경우 주석 처리하여 아래와 같이 코드를 작성하였으며, 데이터 수집과 응용에 있어 어떤 식으로 형태가 이루어지게 되어있는지에 대한 설명이 부가되어 있다.

트위터 API 접근(1)

consumer_secret<-"KYNZQZZbOHUarr1nVcmLUNnqi28Bv6biu6v0OXW84Z9ErlbM95" consumer_key<-"mbJ2HFxcLbx8KWcDIRDpWkTi5" access_secret<-"hqGxBHGgfMfA2MlI4cslCeCX1DLp24dIBfaNVEpXfE9xz" access_token<-"1127798404455600130-aOadogSGwHmfyh9KIC9RBLAnMWAtB3"

setup_twitter_oauth(consumer_key = consumer_key, consumer_secret = consumer_secret, access_token = access_token, access_secret = access_secret)

starbucks,american airline 키워드 중심 게시글 크롤링

term_starbuck<-“#starbucks” trending_starbuck=searchTwitter(term_starbuck,n=1500,since=“2019-06-11”,until=“2019-06-17”,lang=“en”) term_america<-“#americanair” trending_american=searchTwitter(term_america,n=1500,since=“2019-06-10”,until=“2019-06-17”,lang=“en”) – 위와 같은 코드로 3 주간 크롤링을 지속하였다.

트위터 API 접근(2)

consumer_secret<-“A5b7BpXCuEOGLVh7SRQuREZc9YFlaG0xkKcjpJYoFWyvYLZBYe” consumer_key<-“k8YYvWiaxIYipD6G6WQdgNEUu” access_secret<-“U9c5QpHGwyBsiiqOB4i0A3z9g1L9sAgkOC42xR2UguUSt” access_token<-“1140208861057433600-Q0ePMd8qORLsW8uXkhZT2MmahHwIJ3”

setup_twitter_oauth(consumer_key = consumer_key, consumer_secret = consumer_secret, access_token = access_token, access_secret = access_secret)

starbucks, american airline 계정 크롤링

twitterUser <- getUser(“Starbucks”) tweets_star <- userTimeline(twitterUser, n = 1500, includeRts=FALSE,excludeReplies=FALSE)

twitterUser <- getUser(“AmericanAir”) tweets_american <- userTimeline(twitterUser, n = 1500,includeRts=FALSE,excludeReplies=FALSE)

트위터 데이터 데이터프레임화 함수 선언

listtodf<-function(temp){ trendingTweets.df = twListToDF(temp) trendingTweets.df$text <- -sapply(trendingTweets.df$text,function(x) iconv(enc2utf8(x), sub=“byte”)) return(trendingTweets.df) }

감정분석을 위한 사용자 정의 함수 선언

encodeSentiment <- function(x) { if(x <= -0.5){ “1) very negative” }else if(x > -0.5 & x < 0){ “2) negative” }else if(x > 0 & x < 0.5){ “4) positive” }else if(x >= 0.5){ “5) very positive” }else { “3) neutral” } }

america_account<-listtodf(tweets_american) starbucks_account<-listtodf(tweets_star) review_starbucks<-listtodf(trending_starbuck) review_america<-listtodf(trending_american)

감정분석의 결과를 데이터 프레임에 병합

america_account$text <- as.character(america_account$text) tweetSentiments <- get_sentiment(america_account$text,method = "syuzhet") america_account<- cbind(america_account, tweetSentiments)

starbucks_account$text <- as.character(starbucks_account$text) tweetSentiments <- get_sentiment(starbucks_account$text,method = "syuzhet") starbucks_account<- cbind(starbucks_account, tweetSentiments)

write.csv(america_account,file="american_account_example.csv")
write.csv(starbucks_account,file="starbucks_account_example.csv")
write.csv(review_starbucks,file="review_america_example.csv")
write.csv(review_america,file="review_america_example.csv")

- 위와 같이 데이터를 수집, 크롤링하여 저장하는 방법을 거쳐 데이터를 생성하게 되었다.

연구가설 1

텍스트의 감정 상태와 소비자의 반응은 상관관계를 보일 것이다.

- 위 연구가설에서 텍스트의 감정 상태라 함은, 감정분석의 결과가 절대적으로 큰 값을 보인다는 의미로, 텍스트의 감정 분석 결과가 클수록 고객이 공감하고, 공유하는 횟수가 클 것이라는 가설을 세우게 되었다
- 이에 따라, 두개의 nominal 변수인 감성분석과 리트윗 열을 가지고 진행하므로 상관분석을 실행하게 되었다.

패키지 업로드

```
library(twitteR)
library(tm)

## Loading required package: NLP

library(ggplot2)

## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate

library(stringr)
library(syuzhet)
library(tidyverse)

## -- Attaching packages ---------------------------------------------------
----- tidyverse 1.2.1 --

## √ tibble  2.1.3     √ purrr   0.3.2
## √ tidyr   0.8.3     √ dplyr   0.8.1
## √ readr   1.3.1     √ forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
tidyverse_conflicts() --
## x ggplot2::annotate() masks NLP::annotate()
## x dplyr::filter()     masks stats::filter()
## x dplyr::id()         masks twitteR::id()
## x dplyr::lag()        masks stats::lag()
## x dplyr::location()   masks twitteR::location()

library(dplyr)
library(tidyr)
library(knitr)
library(magrittr)

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

library(ggthemes)
```

단순 상관 분석

- 데이터 업로드

```
starbucks<-read.csv("starbucks_account_example.csv",header = TRUE,stringsAsFactors = FALSE)
america<-read.csv("american_account_example.csv", header=TRUE,stringsAsFactors = FALSE)
#분석을 위해 각각 제품군의 이름으로 데이터를 업로드 하였다.

#데이터의 길이를 맞춰주는 함수
#앞에 넣는 데이터가 반환을 원하는 데이터라 지정하고 코드를 작성하였다.
sameline<-function(a,b){
  if(nrow(a)>nrow(b)){
    a<-a[1:nrow(b),]
  }else{
  }
  return(a)
}

#구조 확인
str(america)
```

```
## 'data.frame':    11500 obs. of  21 variables:
##  $ X.3            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X.2            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X.1            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ text           : chr  "@IMV4 Discrimination has no place at American Airlines. Send us a DM with more info." "@itsaninsider Send us your record locator via DM and we'll take a look." "@itsaninsider We have a deep culture of respect for both our customers and our team members. How can we help?" "@IMV4 We're sorry about the long wait. Send a DM our way with the record locator so  we can look into this." ...
##  $ favorited      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ favoriteCount  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ replyToSN      : chr  "IMV4" "itsaninsider" "itsaninsider" "IMV4" ...
##  $ created        : chr  "2019-06-16 8:03" "2019-06-16 7:52" "2019-06-16 7:47" "2019-06-16 7:46" ...
##  $ truncated      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ replyToSID     : num  1.14e+18 1.14e+18 1.14e+18 1.14e+18 1.14e+18 ...
##  $ id             : num  1.14e+18 1.14e+18 1.14e+18 1.14e+18 1.14e+18 ...
##  $ replyToUID     : num  4.85e+08 2.29e+09 2.29e+09 4.85e+08 2.21e+07 ...
##  $ statusSource   : chr  "<a href=\"http://ivoucher.aa.com/snap\" rel=\"nofollow\">SNAP101</a>" "<a href=\"http://ivoucher.aa.com/snap\" rel=\"nofollow\">SNAP101</a>" "<a href=\"http://ivoucher.aa.com/snap\" rel=\"nofollow\">SNAP101</a>" "<a href=\"http://ivoucher.aa.com/snap\" rel=\"nofollow\">SNAP101</a>" ...
```

```
##  $ screenName     : chr  "AmericanAir" "AmericanAir" "AmericanAir" "Americ
anAir" ...
##  $ retweetCount   : int  3 6 10 18 56 66 76 28 53 13 ...
##  $ isRetweet      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ retweeted      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ longitude      : logi  NA NA NA NA NA NA ...
##  $ latitude       : logi  NA NA NA NA NA NA ...
##  $ tweetSentiments: num  -0.1 0.4 0.6 -0.35 0.4 0.25 -0.25 -0.25 -0.3 0.4
 ...
```

```
#구조 확인
str(starbucks)
```

```
## 'data.frame':    17554 obs. of  20 variables:
##  $ X.2            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X.1            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ text           : chr  "@Randomic_Queen Our Dragon Drink looks good and
tastes good!" "@KHANHFIDENT Our Mango Dragonfruit Starbucks Refreshers is a g
reat afternoon choice!" "@mariahmarielove *insert not so evil genius laugh*"
"@ilisamarie94 That makes us happy to hear!" ...
##  $ favorited      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ favoriteCount  : int  2 1 1 1 1 0 0 0 0 1 ...
##  $ replyToSN      : chr  "Randomic_Queen" "KHANHFIDENT" "mariahmarielove"
"ilisamarie94" ...
##  $ created        : chr  "2019-06-15 1:10" "2019-06-15 1:03" "2019-06-15
0:31" "2019-06-15 0:20" ...
##  $ truncated      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ replyToSID     : num  1.14e+18 1.14e+18 1.14e+18 1.14e+18 1.14e+18 ...
##  $ id             : num  1.14e+18 1.14e+18 1.14e+18 1.14e+18 1.14e+18 ...
##  $ replyToUID     : num  3.43e+08 2.39e+09 4.32e+08 5.08e+08 1.84e+08 ...
##  $ statusSource   : chr  "<a href=\"http://www.lithium.com\" rel=\"nofollo
w\">Lithium Tech</a>" "<a href=\"http://www.lithium.com\" rel=\"nofollow\">Li
thium Tech</a>" "<a href=\"http://www.lithium.com\" rel=\"nofollow\">Lithium
Tech</a>" "<a href=\"http://www.lithium.com\" rel=\"nofollow\">Lithium Tech</
a>" ...
##  $ screenName     : chr  "Starbucks" "Starbucks" "Starbucks" "Starbucks"
 ...
##  $ retweetCount   : int  170 344 59 177 92 239 936 417 10 113 ...
##  $ isRetweet      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ retweeted      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ longitude      : logi  NA NA NA NA NA NA ...
##  $ latitude       : logi  NA NA NA NA NA NA ...
##  $ X              : logi  NA NA NA NA NA NA ...
##  $ tweetSentiments: num  0.75 0.9 0.25 0.75 1 0.6 2 -1.4 0 0.75 ...
```

```
america<-sameline(america,starbucks)
starbucks<-sameline(starbucks,america)
#sameline 함수를 통해 두 변수간 행을 맞추어 다시 저장하였다.
```

- 상관계수 확인

```
cor.test(america$retweetCount,america$tweetSentiments)
```

```
##
##  Pearson's product-moment correlation
##
## data:  america$retweetCount and america$tweetSentiments
## t = 32.128, df = 11498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2701590 0.3037028
## sample estimates:
##       cor
## 0.2870189
```

- 고관여 제품군의 상관계수를 확인했을 때, 0.2538786 로 항공사 계정의 텍스트의 감정 분석 결과와 retweet 되는 개수가 상관이 있다는 결론을 내릴 수 있다.

```
cor.test(starbucks$retweetCount,starbucks$tweetSentiments)
```

```
##
##  Pearson's product-moment correlation
##
## data:  starbucks$retweetCount and starbucks$tweetSentiments
## t = 61.604, df = 11498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4842864 0.5117717
## sample estimates:
##       cor
## 0.4981542
```

- 상관계수를 확인했을 때, 0.4981542 로 스타벅스 계정과 텍스트의 감정 분석 결과와 retweet 되는 개수가 상관이 있다는 결론을 내릴 수 있다.

- 단순 상관분석에 이어, 단순 회귀 분석으로 두 변수간에 어떤 인과관계가 있는지 알아보고자 아래와 같은 코드를 추가적으로 진행하게 되었다.

##단순 회귀 분석

```
#저관여 제품군의 감정분석과 리트윗간의 인과관계 분석
regression1_1 <- lm(retweetCount~tweetSentiments ,data=starbucks)
summary(regression1_1)
```

```
##
## Call:
```

```
## lm(formula = retweetCount ~ tweetSentiments, data = starbucks)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -825.3  -85.0  -45.2   61.6 8956.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       54.221      2.881   18.82   <2e-16 ***
## tweetSentiments  179.359      2.911   61.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 232.2 on 11498 degrees of freedom
## Multiple R-squared:  0.2482, Adjusted R-squared:  0.2481
## F-statistic:  3795 on 1 and 11498 DF,  p-value: < 2.2e-16
```

- 분석 결과, f 값이 3795, p 값이 0.05 보다 낮아 영가설을 기각하며, 독립변수의 기울기값에 대한 추정량은 54.221
- 결과적으로 감정적인 텍스트는 리트윗에 영향을 미친다는 연구가설을 채택한다.

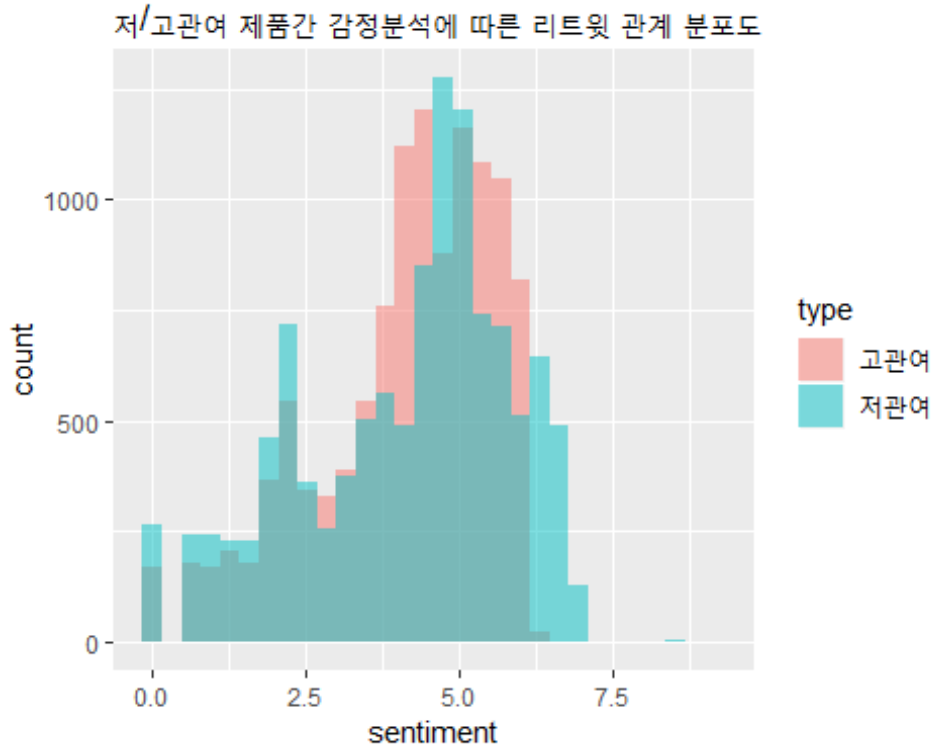
```
#고관여 제품군의 감정분석과 리트윗간의 인과관계 분석
regression1_2 <- lm(retweetCount~tweetSentiments ,data=america)
summary(regression1_2)

##
## Call:
## lm(formula = retweetCount ~ tweetSentiments, data = america)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.20  -84.38  -36.38   61.51 3009.86
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       97.896      1.298   75.41   <2e-16 ***
## tweetSentiments   48.972      1.524   32.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115.7 on 11498 degrees of freedom
## Multiple R-squared:  0.08238,    Adjusted R-squared:  0.0823
## F-statistic:  1032 on 1 and 11498 DF,  p-value: < 2.2e-16
```

- 분석결과, f 값이 792.2, 유의도가 2.2e-16 으로 영가설을 기각하며 독립변수의 기울기값에 대한 추정량은 107.746.
- 결과적으로 감정적인 텍스트는 리트윗에 영향을 미친다는 연구가설을 채택한다.

저관여 제품군과 고관여 제품군간의 소통유형에 따른 민감도 분석

- 위의 결과를 토대로, 고관여 제품군이 저관여 제품군에 비해 감정분석의 결과(정보중심의 커뮤니케이션 유형)에 따른 고객의 반응이 적음을 알 수 있었다.
- 이를 가시적으로 확인하기 위해, 아래와 같은 코드를 작성하였다.

```
#저/고관여 제품군의 텍스트민감도를 확인하기 위한 변수 생성

test_a<-america%>%select(tweetSentiments,retweetCount)
test_s<-starbucks%>%select(tweetSentiments,retweetCount)
#각 데이터에서 감정분석의 결과와 리트윗 개수를 추출


test_a$type[1:nrow(test_a)]<-"고관여"
test_s$type[1:nrow(test_s)]<-"저관여"
#독립변수로 사용하기위한 열 생성

beta_anl<-rbind(test_a,test_s)
#분포도를 위한 변수 선언

beta_anl$type<-as.factor(beta_anl$type)
#새로생성된 데이터의 type 을 명목형 선언

library(ggplot2)
ggplot(beta_anl, aes(log(retweetCount+1), fill=type)) +geom_histogram(alpha=
0.5, position="identity")+ggtitle("저/고관여 제품간 감정분석에 따른 리트윗 관계
분포도") + xlab("sentiment") +ylab("count")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#ggplot 실행
```

- 그래프로 확인하였을때 상대적으로 저관여 제품군이 감정 분석에 따른 리트윗 반응이 더 큰것을 확인할 수 있다.
- 상관관계 확인과 별개로, 저관여 제품군의 경우 평균적으로 리트윗 개수가 고르게 높은 반면 고관여 제품군의 경우 특정 감정 분석의 결과에 민감하게 반응함을 알 수 있다.
- 이러한 결과는 제품군의 특성에 달려있음을 알 수 있다. 고관여 제품군인 america airline 의 경우 감정분석의 결과가 큰 텍스트의 경우 주로 항공 서비스에서의 사건이나, 비행 추락등의 사건을 다루어 고객이 민감하게 반응을 보이나 기본적인 정보전달의 텍스트엔 반응을 잘 보이지 않는 것으로 해석할 수 있다.
- 더불어, 항공사의 상관분석 결과와 비교했을 때 상대적으로 높은 상관계수를 보여 저관여 제품군의 경우 텍스트의 감정 상태와 리트윗과의 관계가 고관여 제품군보다 밀접하게 연관되어 있음을 알 수 있었다.

- 이를 미루어 보아, 저관여 제품군의 경우 고관여 제품군보다 커뮤니케이션형(감정분석의 결과값이 큰) 텍스트에 고객층이 민감하게 반응한다는 것을 알 수 있다.

연구가설 2

제품군에 따라 게시글의 커뮤니케이션 유형에 차이가 있을 것이다.

- '제품군과 문화에 따른 기업 트위터의 커뮤니케이션 유형과 정보내용의 차이'(권택주,조창환 2012)를 바탕으로, 제품군에 따라 게시글의 커뮤니케이션 유형에 차이가 있을 것이라는 가설을 세우게 되었다.
- 앞서 설명한 논문의 경우, 커뮤니케이션 유형에 관한 4 가지 모델을 설립, spss 를 통해 분석하였으나 강의 목적에 맞추어 텍스트의 감정을 분석, 결과를 도출하여 고관여 제품군의 경우 저관여 제품군보다 게시글의 감정분석 결과가 러프하게 나올 것이라는 가정하에 아래와 같은 코드를 작성하게 되었다.

분산 분석을 위한 데이터 전처리

- 분산 분석에 앞서, 저관여 제품군인 starbucks 와 고관여 제품군인 american airline 을 비교하기 위해 새로운 변수를 선언하였다.
- 각 데이터셋에서 감정분석열을 추출, 저관여/고관여 제품군 열을 생성해 분산 분석을 실행하려는 목적하에 아래와 같은 코드를 작성하게 되었다.

```
library(dplyr)
new_star<-starbucks%>%select(tweetSentiments)
new_am<-america%>%select(tweetSentiments)
#데이터 셋에서 감정열 추출
new_star$category[1:nrow(new_star)]<-"1.저관여 제품군"
new_am$category[1:nrow(new_am)]<-"2.고관여 제품군"
#분산 분석을 위한 독립변수 열 생성
new_star<-sameline(new_star,new_am)
new_am<-sameline(new_am,new_star)
#행 맞춤 함수 실행

analysis_sentiment<-rbind(new_star,new_am)
#분산 분석을 위한 데이터 선언
```

```r
colnames(analysis_sentiment)<-c("sentiment","category")
#이름 새로 생성
analysis_sentiment$category<-as.factor(analysis_sentiment$category)
#팩터형 변환

analysis_sentiment$sentiment<-analysis_sentiment$sentiment*10
#계산을 용이하게 하기 위해 각 감정열에 10 을 곱함
```

- 정규성 검정
- 분산분석에 앞서 정규성을 검정하기위해 levene test 를 진행하였다.

```r
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

leveneTest(sentiment ~ category, data=analysis_sentiment)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1   145.7 < 2.2e-16 ***
##       22998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

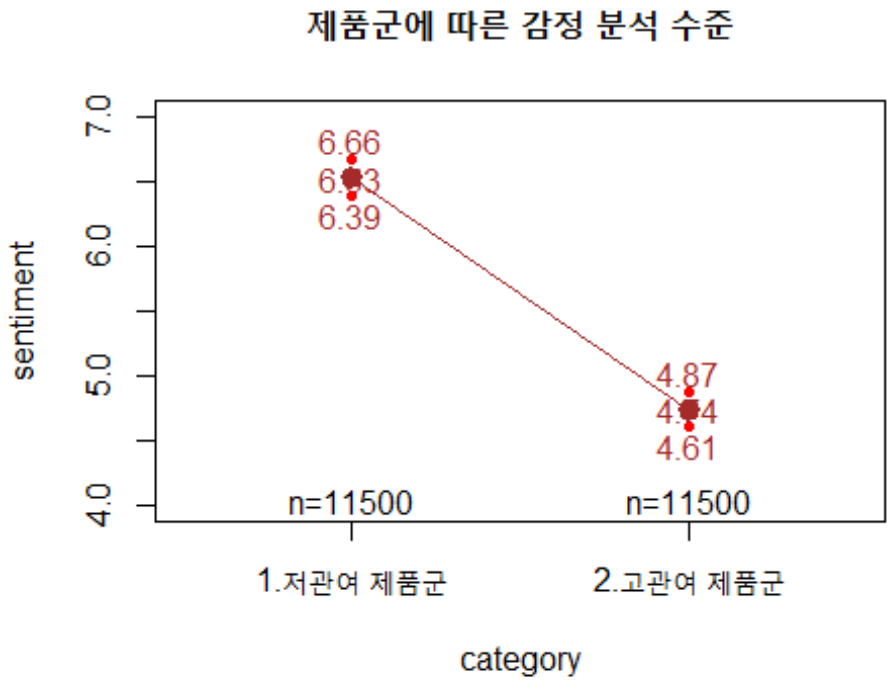
- 정규성 가정을 위한 검사로서 레빈테스트를 실시한 결과, 저관여 제품군과 고관여 제품군간의 등분산성을 검증할 p 값이 0.05 보다 작아 검증되지 않아 분산 분석 실행시 var equal 을 false 로 두고 해야한다는 결론을 내릴 수 있다.

```r
#상관 관계 분석
ano1<-aov(sentiment~category,data=analysis_sentiment)
anova(ano1)

## Analysis of Variance Table
##
## Response: sentiment
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## category      1   18425 18425.4   349.6 < 2.2e-16 ***
```

```
## Residuals 22998 1212087    52.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- p 값이 0.05 보다 낮게 나타나므로 제품군에 따라 커뮤니케이션의 유형이 다르다는 연구가설을 채택한다.

```
library(gplots)
plotmeans(sentiment ~ category, data=analysis_sentiment,
xlab="category", ylab="sentiment", ci.label=T,
mean.label=T, barwidth=5, digits=2,
col="brown", pch=1, barcol="red", ylim=c(4,7),
main="제품군에 따른 감정 분석 수준")

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, :
## zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, :
## zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, :
## zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, :
## zero-length arrow is of indeterminate angle and so skipped
```



가설 3 저관여 제품군의 정보 탐색 경로는 고관여 제품군과 차이가 있을 것이다

- 저관여 제품군의 정보 탐색의 경우 데스크톱보단 모바일을 주로 활용하는 반면, 고관여 제품군의 경우 테스크톱을 활용하여 정보 탐색을 한다는 가설하에 아래와 같은 코드를 작성하게 되었다.

분석을 위한 접속경로 단순화 함수 선언

- 분석에 앞서, android, facebook, ifttt, ipad, iphone, others, Web 등으로 나뉘어져 있는 접속 경로를 가설을 위해 단순화하기 위해 아래와 같은 함수를 선언하였다.
- 안드로이드나 아이폰, 아이패드의 경우 모바일로, 페이스북, ifttt, web 의 경우 데스크톱으로 분류하여 분석 이전 데이터 전처리를 실행하였다.

```
encodeSource <- function(x) {
if(grepl(">Twitter for iPhone</a>", x)){ "mobile"
}else if(grepl(">Twitter for iPad</a>", x)){ "mobile"
}else if(grepl(">Twitter for Android</a>", x)){ "mobile"
} else if(grepl(">Twitter Web Client</a>", x)){ "web"
} else if(grepl(">Twitter for Windows Phone</a>", x)){ "web"
}else if(grepl(">dlvr.it</a>", x)){ "web"
}else if(grepl(">IFTTT</a>", x)){ "web"
}else if(grepl(">Twitter Web App</a>", x)){ "web"
}else if(grepl(">TweetDeck</a>", x)){ "web"
}else if(grepl(">Twitter for iPhone</a>", x)){ "mobile"
}else if(grepl(">Facebook</a>", x)){ "web"
}else { "others"
}
}

#제품군간의 정보탐색경로를 확인하기위한 csv 업로드
review_am<-read.csv("review_america_example.csv",header=TRUE)
review_st<-read.csv("review_starbucks_example.csv",header=TRUE)

#고관여 제품군 구조확인
str(review_am)

## 'data.frame':    9265 obs. of  17 variables:
##  $ X            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ text         : Factor w/ 124 levels "\"On top of it all everyone is so
rude. As if THEY are the ones inconvenienced by this somehow. So furious.\"<e
"| __truncated__,..: 50 3 95 122 62 19 116 2 105 111 ...
##  $ favorited    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
##  $ favoriteCount: int  1 0 0 0 0 0 1 0 0 0 ...
##  $ replyToSN    : Factor w/ 19 levels "AmericanAir",..: 3 NA NA NA NA NA N
A NA NA NA ...
##  $ created      : Factor w/ 127 levels "2019-06-09 18:11",..: 118 117 116
115 114 113 122 121 120 119 ...
##  $ truncated    : logi  TRUE FALSE FALSE TRUE TRUE FALSE ...
##  $ replyToSID   : num  1.14e+18 NA NA NA NA ...
##  $ id           : num  1.14e+18 1.14e+18 1.14e+18 1.14e+18 1.14e+18 ...
##  $ replyToUID   : num  59621062 NA NA NA NA ...
##  $ statusSource : Factor w/ 14 levels "<a href=\"http://instagram.com\" re
l=\"nofollow\">Instagram</a>",..: 3 12 12 12 12 12 12 12 12 12 ...
##  $ screenName   : Factor w/ 103 levels "__sandeepgupta",..: 14 90 87 19 38
 41 81 101 77 84 ...
##  $ retweetCount : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ isRetweet    : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
##  $ retweeted    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ longitude    : logi  NA NA NA NA NA NA ...
##  $ latitude     : logi  NA NA NA NA NA NA ...
```

```
#저관여 제품군 구조확인
str(review_st)
```

```
## 'data.frame':    8031 obs. of  16 variables:
##  $ text         : Factor w/ 2531 levels "'S latest post\n<e2>씩截<8f>\nLuc
as till + Starbucks \n<e2>씩截<8f>\nT.A.G.S.\n@lucastill\n#lucastill\n#2019\n
#"| __truncated__,..: 1352 2137 1281 629 1276 272 219 1606 1763 1529 ...
##  $ favorited    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ favoriteCount: int  0 0 0 0 0 2 2 0 0 0 ...
##  $ replyToSN    : Factor w/ 139 levels "absolutelyChike",..: NA NA NA NA N
A 27 1 NA NA NA ...
##  $ created      : Factor w/ 2629 levels "2019-06-11 0:00",..: 1921 1920 19
19 1918 1917 1916 1914 1913 1912 1911 ...
##  $ truncated    : logi  TRUE FALSE TRUE TRUE TRUE TRUE ...
##  $ replyToSID   : num  NA NA NA NA NA ...
##  $ id           : num  1.14e+18 1.14e+18 1.14e+18 1.14e+18 1.14e+18 ...
##  $ replyToUID   : num  NA NA NA NA NA ...
##  $ statusSource : Factor w/ 42 levels "<a href=\"http://alaskafilmmakers.c
om/\" rel=\"nofollow\">TwitterBot907</a>",..: 16 16 16 16 16 16 16 16 16 16
 ...
##  $ screenName   : Factor w/ 2643 levels "___Admsyn","__Mcl0vinn",..: 412 5
51 1269 1090 1815 2292 844 1399 1139 744 ...
##  $ retweetCount : int  0 0 0 1 0 0 0 2 2 0 ...
##  $ isRetweet    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ retweeted    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ longitude    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ latitude     : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
#기존에 선언한 encode source 로 sapply
review_am$tweetSource = sapply(review_am$statusSource, encodeSource)
review_st$tweetSource = sapply(review_st$statusSource, encodeSource)

#행을 맞추기 위해 sameline 실행
review_am<-sameline(review_am,review_st)
review_st<-sameline(review_st,review_am)

#정보탐색경로만을 저장하는 contact 변수선언
contact_st<-review_st%>%select(tweetSource)
contact_am<-review_am%>%select(tweetSource)

contact_st$category[1:nrow(contact_st)]<-"저관여"
contact_am$category[1:nrow(contact_am)]<-"고관여"
#독립변수를 위한 category 열 선언

analysis_contact<-rbind(contact_st,contact_am)
#분석을 위한 데이터 프레임 생성

colnames(analysis_contact)<-c("contact","category")
#열이름 지정

analysis_contact$category<-as.factor(analysis_contact$category)
analysis_contact$contact<-as.factor(analysis_contact$contact)
#명목형변수 변환

analysis_contact<-na.omit(analysis_contact)
#NA 값 제거

analysis_contact_table<-table(analysis_contact$contact,analysis_contact$categ
ory)
#교차분석을 위한 table 변수 선언
```

교차 분석 실행

```
chisq.test(analysis_contact_table, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  analysis_contact_table
## X-squared = 729.98, df = 2, p-value < 2.2e-16
```

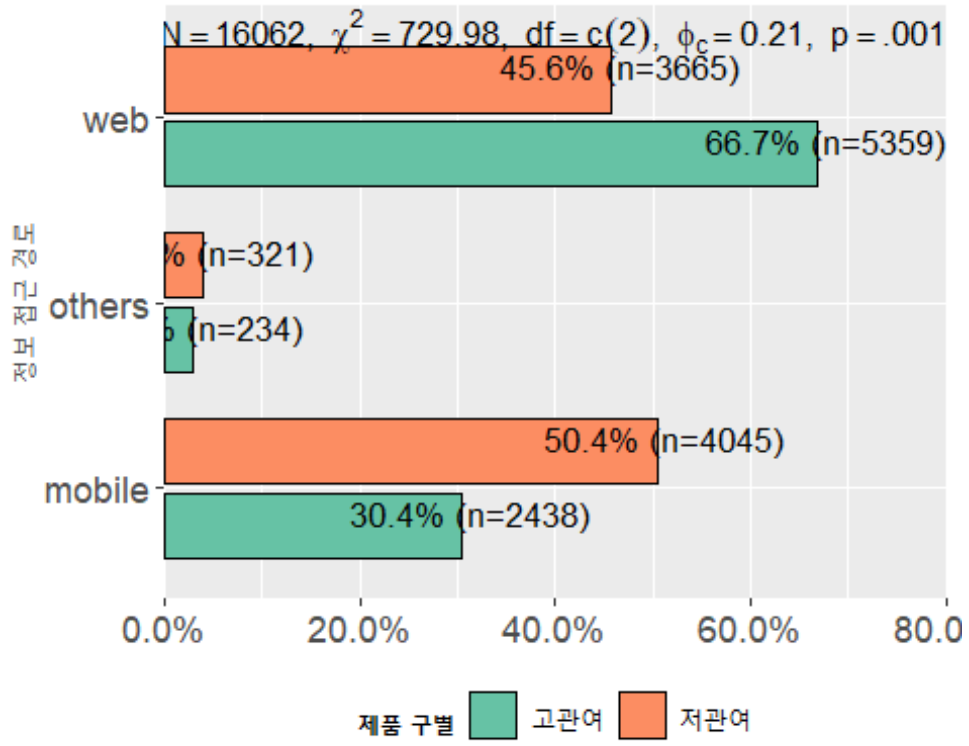
- 카이스퀘어 값이 747.52, 유의 확률이 2.2e-16.

- 카이스퀘어 테스트를 통해 p 값을 확인한 결과, p 값이 0.05 아래로, 제품군에 따라 정보접근의 경로가 다르다는 연구가설을 채택한다.

```
library(sjPlot)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
set_theme(geom.label.size = 4.5, axis.textsize = 1.1,
legend.pos="bottom")
sjp.xtab(analysis_contact$contact, analysis_contact$category,type="bar",
y.offset = 0.01,margin = "col", coord.flip = T, wrap.labels = 7,
geom.colors = "Set2", show.summary = T, show.total = F,
axis.titles = "정보 접근 경로",
legend.title = "제품 구별")
```



- 확인 결과, 고관여 제품군이 web 으로 접속하는 경우가 더 많은 것으로 나타났다. 즉, 소비자가 관심도가 높고 잘못된 구매의사 결정을 내렸을 경우의 지각될 위험이 클 때 모바일보다는 데스크 톱을 통해 정보에 접근한다는 결론을 내릴 수 있다.

연구결과 및 시사점

- 위와 같은 연구 가설을 통해, 전반적으로 저관여 제품군보다 고관여 제품의 경우 고객과 소통하는 방식이 정보전달에 가까운 것을 확인할 수 있었다.(연구가설 2)
- 또한 텍스트의 감정 분석의 결과와 소비자와의 반응이 상관관계가 있음을 확인 할 수있었으며, 저관여 제품군의 경우 고관여 제품군에 비해 상대적으로 고객층이 제품군의 텍스트의 감정 상태에 민감하게 반응함을 알 수 있다.(연구가설 1)
- 또한 고관여 제품군의 경우 정보에 접근하는 경로가 저관여 제품군에 비해 데스크톱으로 접근하는 경우가 많음을 확인할 수 있었다.(연구가설 3)
- 그러나 트위터 데이터 크롤링 기간을 한 달 가량으로 설정했기 때문에, 위와 같은 연구결과를 일반화하기엔 다소 어려울 수 있다.
- 또한 정보접근 경로의 경우 페이스북, 인스타그램등의 출처가 불분명한 경로를 임의로 웹과 모바일로 나누어 그 연구결과의 신뢰성이 다소 떨어질 수 있다.