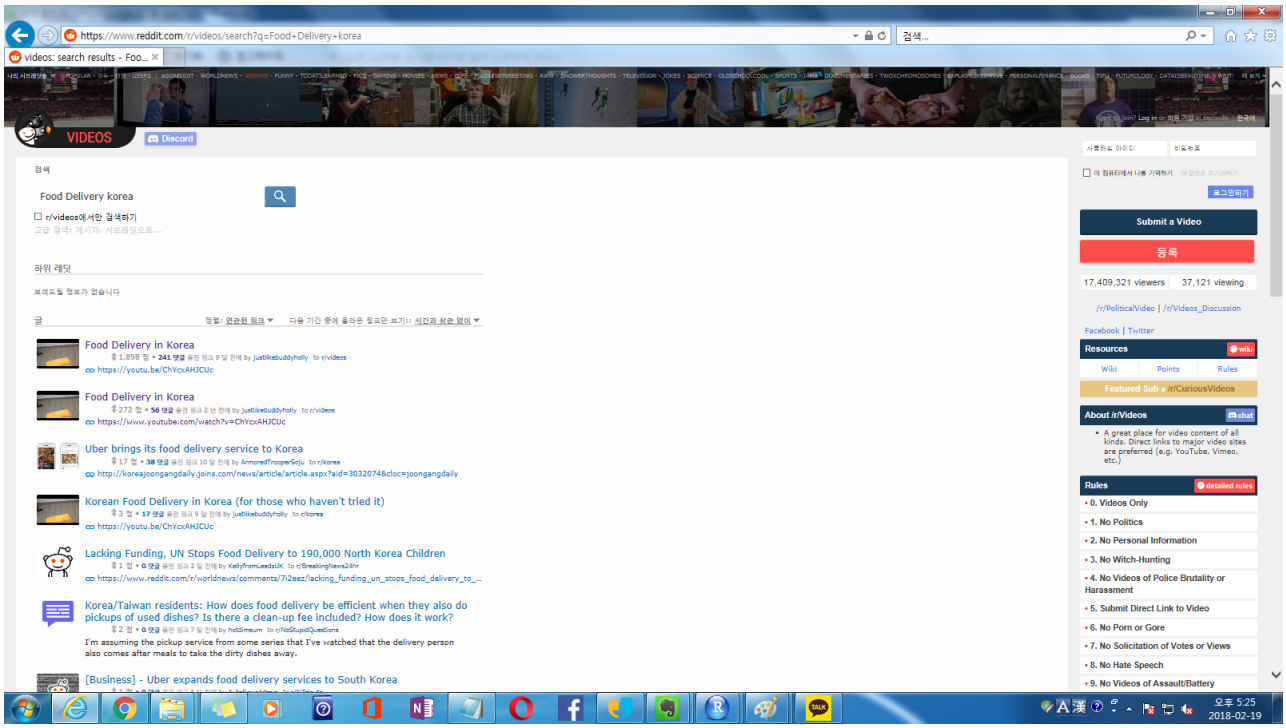
미국에서 바라본 한국 배달문화

서론

우리나라는 세계적으로 배달문화가 발달한 나라로 다양한 종류의 음식들을 배달 가능하다 혼자 사는 1인가구가 늘어나면서 간편하고 편리한 음식에 수요가 증가함에 따라서 기존에 자리잡고 있던 한국의 배달시장이 더욱 상승세가 일어나고 있다. 미국도 전자상거래가 보편화되면서 빠르고 정확한 상품 배송이 중요해져 미국 가정배달(Home Delivery) 서비스 시장 규모가 크게 확대되고 있어 한국의 배달음식과 문화에 대해 어떻게 생각하고 있는지 궁금해 시작하게 되었다.

이 자료는 미국 커뮤니티 사이트 reddit에 Food Delivery in korea 라는 게시물을 보고 여러사람들이 남긴 댓글을 R크롤링해 조사한 자료입니다.



본론

사용된 라이브러리

```
library(rvest)
library(tm)
library(dplyr)
```

```
library(wordcloud2)
library(ggplot2)
library(ggthemes)
library(arules)
library(sna)
library(rgl)
```

게시물 크롤링

첫번째 게시물 주소

```
html<-
read_html("https://www.reddit.com/r/videos/comments/69ix3e/food_delivery_in_korea/")
```

 # html CSS를 이용해 text 저장

```
data1<-html_nodes(html, ".linklisting .md, .commentarea .md")%>%
        html_nodes("p")%>%
        html_text()
```

두번째 게시물 주소

```
html<-
read_html("https://www.reddit.com/r/videos/comments/42fllfq/food_delivery_in_korea/
")

data1<-c(data1,html_nodes(html, ".linklisting .md, .commentarea .md")%>%
          html_nodes("p")%>%
          html_text())
```

세번째 게시물 주소

```
html<-
read_html("https://www.reddit.com/r/korea/comments/64n4on/uber_brings_its_food_delivery_service_to_korea/")

data1<-c(data1,html_nodes(html, ".linklisting .md, .commentarea .md")%>%
          html_nodes("p")%>%
          html_text())
```

네번째 게시물 주소

```
html<-
read_html("https://www.reddit.com/r/korea/comments/69ixqc/korean_food_delivery_in_korea_for_those_who/")

data1<-c(data1,html_nodes(html, ".linklisting .md, .commentarea .md")%>%
           html_nodes("p")%>%
           html_text())


# 크롤링한 데이터를 txt 파일 저장
write.csv(data1, file="c:/r/reddit.txt", row.names=FALSE)


# 데이터 파일 불러옴
setwd("c:/r")
data1<-read.csv("reddit.txt",header=F)
```

text mining

```
library(tm)

# 벡터형을 -> VCorpus 변환후 작업 시작
corp1 <- VCorpus(VectorSource(data1))
```

데이터 정제

```
# 대문자를 소문자 변환
corp2 <- tm_map(corp1,tolower)

# 특수 문자 제거
corp2 <- tm_map(corp2,removePunctuation)

# 숫자 제거
corp2 <- tm_map(corp2,removeNumbers)

# 마감 / 형변환
corp2 <- tm_map(corp2,PlainTextDocument)
```

```
# 불필요한 전치사와 단어들을 word 변수에 담아준다
word1 <- c(stopwords(´en´),
"and","but","not","the","point","for","just","are","you","that","not","your","like",
          "dont","can","one","get","will","use","much","cant","even","maybe","place","many","d
ay","ive","usd","used","makes",
          "arent","find","yeah","thing","doesnt","youre","pretty","deleted","put","stuff","floor","
eat","lot","dont","eat","live")

word2 <- c(stopwords("SMART"),
          "whats","video","sounds","put","sit","generally","didnt","understand","shit","theyre"
,"fuck",
          "basically","compared","south","korea","food","delivery","people")

# 불필요한 단어 제거
corp2 <- tm_map(corp2,removeWords,word1)
corp2 <- tm_map(corp2,removeWords,word2)
corp2 <- tm_map(corp2,removeWords,stopwords("german"))
tdm<-TermDocumentMatrix(corp2)

# 말뭉치 문서 matrix 변환
m1<-as.matrix(tdm)

형 변환 시키고 오름차순으로
m2<-sort(rowSums(m1),decreasing=T)

# 빈도수가 5 번이상 나온것들 체크
m3<-m2[m2>=5]
```

wordcloud 도출

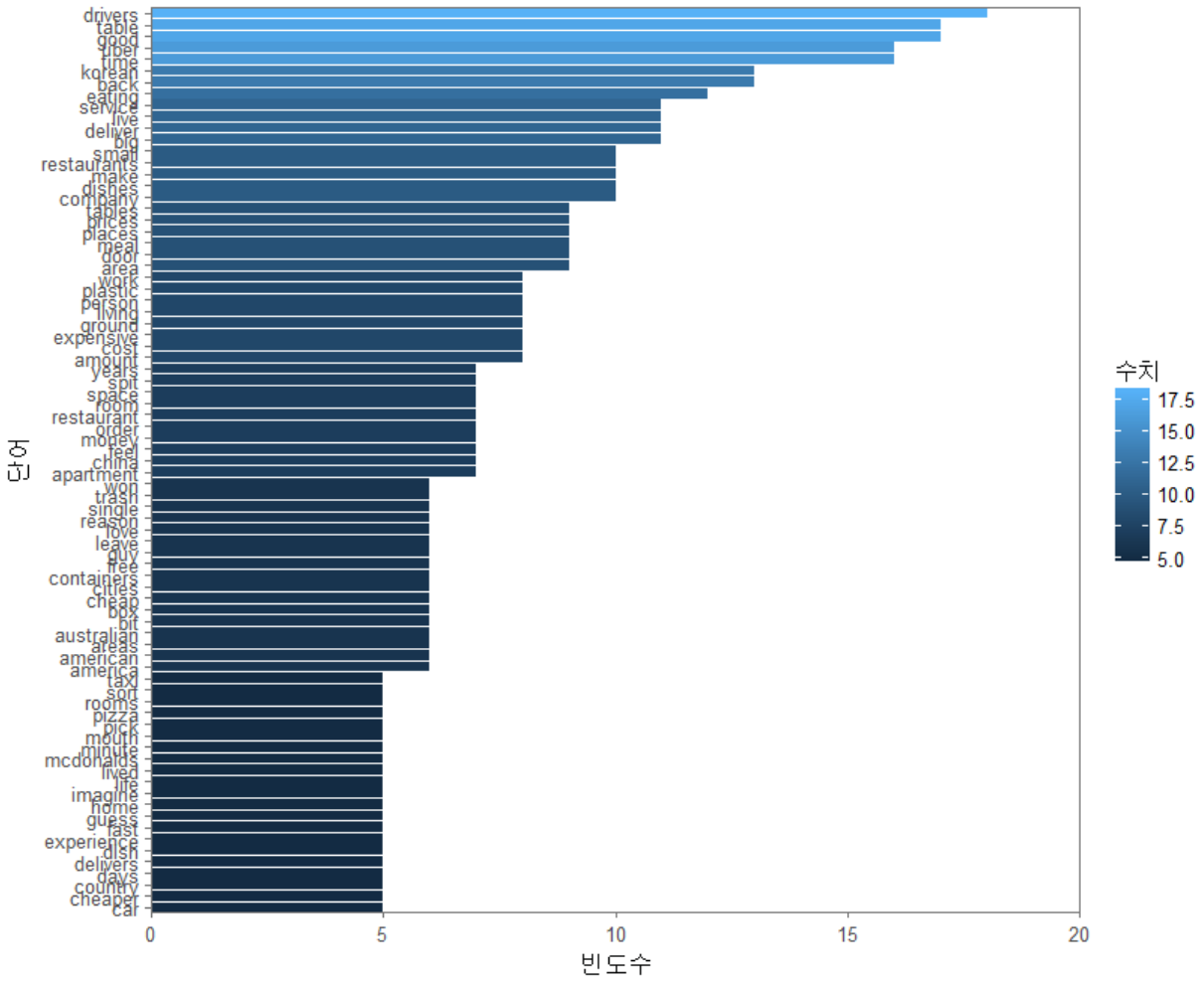
```
wordcloud2(as.table(m3),size=0.5,color = ´random-light´,backgroundColor =
´black´,minRotation = -pi/5)
```


```
# data type 수정
x<-names(m3)
y<-as.numeric(m3)
m4<-data.frame(x,y)


# ggplot으로 단어 빈도수 도식화
library(ggplot2)
library(ggthemes)
```

ggplot 그래프 도출

```
ggplot(m4,aes(x=reorder(x,y),y=y,fill=y))+
      geom_bar(stat=´identity´) +
      labs(x="단어", y="빈도수",fill="수치") + coord_flip(expand = F) +
      theme(axis.text.x=element_text(angle=0,hjust=0.5,vjust=0.5,
```

```
                                                colour="black" , size=8))+
    theme(axis.text.y=element_text(angle=0,hjust=1,vjust=0.4,
                                                colour="black" , size=8))+ylim(0,20) +
theme_few()
```



Comment

uber 에 익숙한 나라답게 한국에서도 uber 를 이용하는지?
어떤 drivers 가 음식를 배달해 주는지? 에 대한
궁금 사항이 높아 보이는것으로 나타낸다. amount, money, cost 등
비용적인 부분에 대한 단어들에 빈도수도 꽤 높게 나온 것을 보면
얼마만큼에 비용지불로 배달음식을 이용할 수 있는지 궁금한 것으로 보인다

```
# 키워드 준비 / 빈도수 11 개 이상 되는것끼리 묶음
keyword<-rownames(as.matrix(m3[(m3)>=11]))
```

```
# 각문장을 문자열로 불러줌
data<-readLines("reddit.txt")

# 명사끼리 묶어줌
library(KoNLP)
data1 <-sapply(data, extractNoun, USE.NAMES=F)

# 2 글자이상 데이터 노출
data2 <- Filter(function(x) {nchar(x)>=2}, data1)

# 비어있는 공간 제거
data2<-data2[1:334]

# for 문으로 데이터의 있는 단어 체크
data3<-c()
for (i in 1:length(data2)){
      index<-intersect(data2[[i]],keyword)
      data3<-rbind(data3,table(index)[keyword])
}

# 이름 & default 값으로 수정
colnames(data3)<- keyword
data3[is.na(data3)]<-0

# 문장의 keyword 가 2 개 미만인 데이터 삭제
for (i in 1:length(data3)){
ifelse(sum(data3[i,])<2,data3[i,]<-NA,data3)
}

data3<-na.omit(data3)


# 위에서 데이터셋 작업중에 오류난 탓에 데이터셋 파일 저장후 불러들임
res<-as.data.frame(data3)
trans<-as.matrix(res,´Transaction´ )


write.csv(trans, file="c:/r/text.txt", row.names=FALSE)
text1<-read.csv("text.txt", header =T )
text2<-as.matrix(text1,´Transaction´ )
```

```
install.packages("arules")
library(arules)
res_rul<-apriori(text2,parameter=list(supp=0.1,conf=0.1,target="rules"))
inspect(sort(res_rul,by='lift'))
```

상관 관계확인

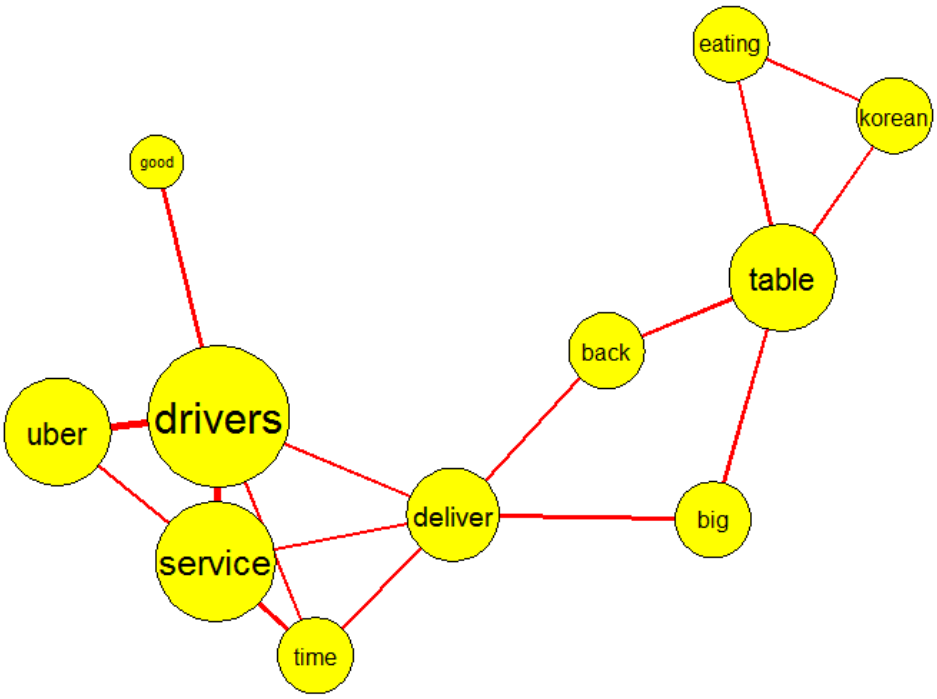
```
     lhs          rhs       support confidence lift     count
[1]  {time}    => {service} 0.1250  1.0000000  3.200000 2
[2]  {service} => {time}    0.1250  0.4000000  3.200000 2
[3]  {uber}    => {drivers} 0.1875  0.7500000  1.714286 3
[4]  {drivers} => {uber}    0.1875  0.4285714  1.714286 3
[5]  {service} => {drivers} 0.1875  0.6000000  1.371429 3
[6]  {drivers} => {service} 0.1875  0.4285714  1.371429 3
[7]  {}        => {eating}  0.1250  0.1250000  1.000000 2
[8]  {}        => {big}     0.1250  0.1250000  1.000000 2
[9]  {}        => {korean}  0.1250  0.1250000  1.000000 2
[10] {}        => {back}    0.1250  0.1250000  1.000000 2
[11] {}        => {time}    0.1250  0.1250000  1.000000 2
[12] {}        => {table}   0.2500  0.2500000  1.000000 4
[13] {}        => {uber}    0.2500  0.2500000  1.000000 4
[14] {}        => {deliver} 0.1875  0.1875000  1.000000 3
[15] {}        => {service} 0.3125  0.3125000  1.000000 5
[16] {}        => {drivers} 0.4375  0.4375000  1.000000 7
```

```
# 단어별로 겹치는 빈도수

b1 <- t(text2)%*%text2

b2<- diag(b1)

b3 <- b1-diag(diag(b1))
```

gplot을 이용해 상관 관계 그래프 도출

```
install.packages("sna")
install.packages("rgl")
```

```
library(sna)
library(rgl)

gplot(b3, displaylabels = T,        #노드레이블을표시
        vertex.cex = sqrt(b2)*1.4,      # 노드 크기 확대 배수
        vertex.col = "yellow",        # 색상
        edge.col ="red",                # 선의 색
        boxed.labels=F,                 # 노드레이블에 박스
        arrowhead.cex = 0,              # 화살표식 크기
        label.pos = 5,              # 노드레이블 위치 0~5
        label.cex = sqrt(b2)*0.7,
        edge.lwd = 1)
```



결론

단어의 상관관계를 분석하여 그래프로 도출해본 결과 빈도수가 가장 많은 drivers 의 관심이 service 라는 단어와 uber 시스템의 관계도가 가장 가까운 것으로 보인다.
배달해주는 사람의 서비스에 대해 관심이 높은것으로 추측하고 그 밖에도
배달 시간이 얼마나 걸리는지 궁금하다는 것을 time 단어에서 추측해 볼 수 있었다.