Question 4

Bejan Sadeghian

August 7, 2015

Question 4 of Homework 1

Objective

To evaluate the users on Twitter that follow the company NutrientH2O and their last 1000 tweets to develop a suggestion to the marketing team on what their users enjoy most and what they should incorporate in their advertisements to draw attention from their interest groups.

Conclusion

My message to the marketing team would be the follows. Cluster 10, 14, and 15 are their largest groups of consumers to target so they should focus their efforts there initially.

- Cluster 10 individuals are the type of person who enjoys keeping up with current events, dating, and home and garden. They dont enjoy religion, researching about schooling, food, sports, or photo sharing necessarily. Because they dont have an interest in schooling research or religion and they are interested in dating and current events I would make the assumption that these people are younger people probablly in their mid 20s. Not necessarily the “hipster genre” but something of a more well rounded person who is starting their life.
- Cluster 14 individuals are the type of person who has interests in business, eco-friendly ideas, current events, talking, shopping, and photo sharing. What they dont enjoy are things like religion, outdoors, beauty, news, cooking, or food. I would characterize this user group as older, maybe in the 30s, who have careers and read about business/economics, they shop and talk alot on online which indicates they are online to discuss ideas as opposed to blurting information. With beauty being a lower interest item, this tells me they are older as well because they may already have a significant other and while they dont completely ignore beauty, they probably already have established methods and dont need to have that on their mind.
- Cluster 15 individuals enjoy outdoors, eco-friendly ideas, food, cooking, personal fitness, and health nutrition. What they dont necessarily enjoy discussing is politics, automotives, sports fanatacism, school, beauty, or university. I would assess this person as being a younger individual probablly out of college. They are definetly an outdoors person given they include outdoors and ‘eco’. They are probablly a bit of a health nut (hence the fitness, nutrition, and cooking).
- The three top clusters of individuals all seem to be in the millennial category. Thats very telling about NutrientH2O’s key demograph.
    - If NutrientH2O wanted to branch out further they could take a look at the fourth largest cluster (cluster 5) of individuals.

- These individuals are very different from the top three in that they are highly interested in school, family, parenting, and religion. They are interested in food and sports as well. What they are least interested in are cooking, health nutrition, personal fitness, news, photo sharing, and politics. This type of individual is definetly a parent or soon to be parent. Likely already an established family. If NutrientH2O wanted to expand their customer base they could focus on the family aspect when marketing to include this group. If we had just looked at the top three groups we would not have thought family would be of interest but for this group they are.

Method

Step 1) I imported the ggplot2, reshape, and foreach libraries and the social_marketing.csv file

```
library(ggplot2)
library(reshape)
library(foreach)

rawdata = read.csv('social_marketing.csv')
```

Step 2) I then removed the first column (userIDs) and scaled the data in preparation for K-Means clustering

```
moddata = rawdata[,-1]

datascaled = scale(moddata, center=TRUE, scale=TRUE)
```

Step 3) For K-Means I ran an optimization of the K value with nstart=10 to find the best number of centroids to run Kmeans with. You can see that information below. ‘’’ K = 40 #Try K-means ErrorArray = rep(0,K) for(i in 1:K){ dataCluster = kmeans(datascaled,centers = i, nstart = 10)# method = ‘kmeans’, dist=‘euclidean’, save.data=TRUE) ErrorArray[i] = dataCluster$tot.withinss } plot(ErrorArray) ‘’’

Step 4) After finding the optimized K value (15 in this case), I performed the K-Means by the R function kmeans()

```
#Perform K-Means with the optimized K value
K=15
dataCluster = kmeans(datascaled,centers = K, nstart = 10)
```

Step 5) To find what characterizes each cluster of user, I wanted to see the centers for each cluster for each variable, I had to do some cleaning of the data and renaming the variables to be able to utilize it in plotting, you can see that below.

```
Centers = dataCluster$centers
transposeCenters = t(Centers)
transposeCenters = as.data.frame(transposeCenters)
transposeCenters = cbind(transposeCenters,rownames(transposeCenters))
colnames(transposeCenters) = c('one','two','three','four','five','six','seve
n','eight','nine','ten','eleven','twelve','thirteen','fourteen','fifteen','ite
m')
meltCenters = melt(transposeCenters,id=c('item'))
```

Step 6) I then plot this information in a bar chart. With 15 clusters you can see that doing any analysis will be very difficult just looking at the graph so I need to take it a step further.

```
ggplot(meltCenters, aes(x=item, fill=variable, y=value)) + geom_bar(position="d
odge", stat="identity") + theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

Step 7) I then took the top interest and lowest interest (based on cluster centers). I also added the count of each cluster so I could make an informed decision on what I was looking at. My evaluation and marketing recommendation for nutrientH2O is in the conclusions above.

```
LowInterest = foreach(i = 1:(length(transposeCenters)-1),.combine=cbind, .inord
er=TRUE) %do% {
  head(rownames((transposeCenters[with(transposeCenters, order(transposeCenters
[,i])), ])))
}
LowInterest = rbind(LowInterest,t(as.matrix(dataCluster$size)))

HighInterest = foreach(i = 1:(length(transposeCenters)-1),.combine=cbind, .inor
der=TRUE) %do% {
  tail(rownames((transposeCenters[with(transposeCenters, order(transposeCenters
[,i])), ])))
}
HighInterest = rbind(HighInterest,t(as.matrix(dataCluster$size)))

print('Higher Interest Items for each Cluster')
print(HighInterest)
print('Lower Interest Items for each Cluster')
print(LowInterest)
```

```
library(ggplot2)
library(reshape)
library(foreach)

set.seed(1)
rawdata = read.csv('social_marketing.csv')

moddata = rawdata[,-1]

datascaled = scale(moddata, center=TRUE, scale=TRUE)

K = 40
#Try K-means
ErrorArray = rep(0,K)
for(i in 1:K){
  dataCluster = kmeans(datascaled,centers = i, nstart = 10)# method = 'kmean
s', dist='euclidean', save.data=TRUE)
  ErrorArray[i] = dataCluster$tot.withinss
}
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```

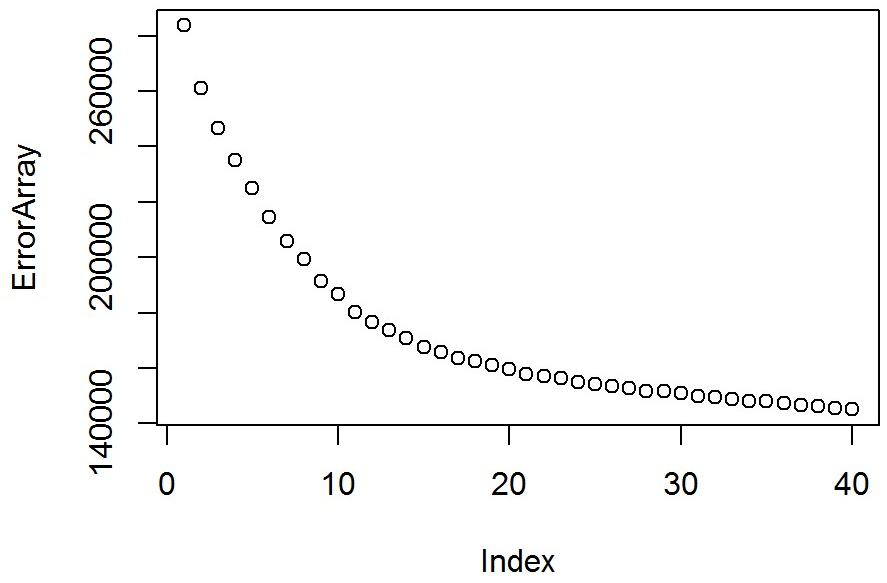
```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```

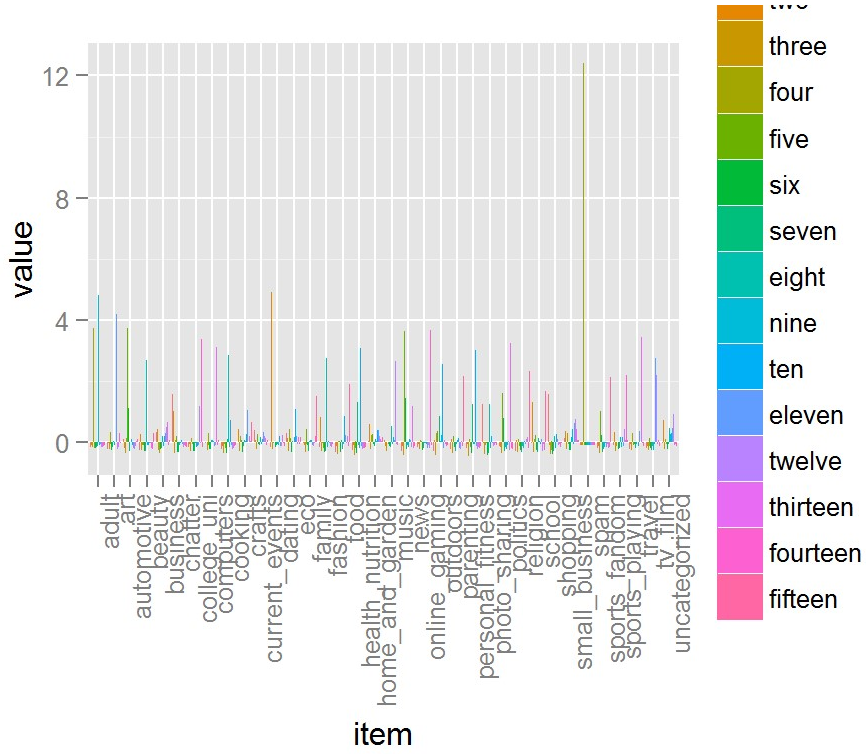
```
plot(ErrorArray)
```




```
#Perform K-Means with the optimized K value
K=15
dataCluster = kmeans(datascaled,centers = K, nstart = 10)

Centers = dataCluster$centers
transposeCenters = t(Centers)
transposeCenters = as.data.frame(transposeCenters)
transposeCenters = cbind(transposeCenters,rownames(transposeCenters))
colnames(transposeCenters) = c('one','two','three','four','five','six','seve
n','eight','nine','ten','eleven','twelve','thirteen','fourteen','fifteen','ite
m')
meltCenters = melt(transposeCenters,id=c('item'))

ggplot(meltCenters, aes(x=item, fill=variable, y=value)) + geom_bar(position="d
odge", stat="identity") + theme(axis.text.x = element_text(angle = 90, hjust =
1))
```




```
LowInterest = foreach(i = 1:(length(transposeCenters)-1),.combine=cbind, .inord
er=TRUE) %do% {
  head(rownames((transposeCenters[with(transposeCenters, order(transposeCenters
[,i])), ])))
}
LowInterest = rbind(LowInterest,t(as.matrix(dataCluster$size)))

HighInterest = foreach(i = 1:(length(transposeCenters)-1),.combine=cbind, .inor
der=TRUE) %do% {
  tail(rownames((transposeCenters[with(transposeCenters, order(transposeCenters
[,i])), ])))
}
HighInterest = rbind(HighInterest,t(as.matrix(dataCluster$size)))

print('Higher Interest Items for each Cluster')
```

```
## [1] "Higher Interest Items for each Cluster"
```

```
print(HighInterest)
```

```
##      result.1         result.2          result.3          result.4
## [1,] "eco"            "home_and_garden" "dating"          "outdoors"
## [2,] "business"       "uncategorized"   "small_business"  "small_business"
## [3,] "current_events" "fashion"         "home_and_garden" "art"
## [4,] "photo_sharing"  "chatter"         "uncategorized"   "eco"
## [5,] "chatter"        "school"          "adult"           "adult"
## [6,] "shopping"       "dating"          "spam"            "spam"
## [7,] "857"            "187"             "2706"            "49"
##      result.5        result.6          result.7          result.8
## [1,] "outdoors"      "home_and_garden" "eco"             "uncategorized"
## [2,] "family"        "outdoors"        "food"            "music"
## [3,] "sports_fandom" "sports_fandom"   "cooking"         "photo_sharing"
## [4,] "politics"      "politics"        "outdoors"        "beauty"
## [5,] "news"          "automotive"      "personal_fitness" "fashion"
## [6,] "automotive"    "news"            "health_nutrition" "cooking"
## [7,] "190"           "442"             "709"             "451"
##      result.9         result.10          result.11
## [1,] "automotive"     "cooking"          "home_and_garden"
## [2,] "eco"            "food"             "uncategorized"
## [3,] "outdoors"       "eco"              "small_business"
## [4,] "uncategorized"  "outdoors"         "crafts"
## [5,] "small_business" "personal_fitness" "tv_film"
## [6,] "adult"          "health_nutrition" "art"
## [7,] "190"            "313"              "251"
##      result.12        result.13        result.14        result.15
## [1,] "business"       "small_business" "small_business" "family"
## [2,] "small_business" "business"       "family"         "school"
## [3,] "uncategorized"  "news"           "art"            "food"
## [4,] "college_uni"    "computers"      "sports_playing" "sports_fandom"
## [5,] "tv_film"        "politics"       "college_uni"    "parenting"
## [6,] "music"          "travel"         "online_gaming"  "religion"
## [7,] "255"            "310"            "334"            "638"
```

```
print('Lower Interest Items for each Cluster')
```

```
## [1] "Lower Interest Items for each Cluster"
```

```
print(LowInterest)
```

```
##      result.1           result.2        result.3
## [1,] "food"             "automotive"    "personal_fitness"
## [2,] "outdoors"         "politics"      "outdoors"
## [3,] "news"             "cooking"       "health_nutrition"
## [4,] "religion"         "food"          "news"
## [5,] "beauty"           "sports_fandom" "food"
## [6,] "health_nutrition" "news"          "photo_sharing"
## [7,] "857"              "187"           "2706"
##      result.4         result.5           result.6
## [1,] "business"       "health_nutrition" "photo_sharing"
## [2,] "shopping"       "cooking"          "shopping"
## [3,] "tv_film"        "fashion"          "health_nutrition"
## [4,] "sports_playing" "adult"            "personal_fitness"
## [5,] "beauty"         "personal_fitness" "cooking"
## [6,] "photo_sharing"  "religion"         "crafts"
## [7,] "49"             "190"              "442"
##      result.7        result.8        result.9           result.10
## [1,] "photo_sharing" "sports_fandom" "tv_film"          "college_uni"
## [2,] "chatter"       "food"          "politics"         "travel"
## [3,] "sports_fandom" "politics"      "photo_sharing"    "politics"
## [4,] "school"        "tv_film"       "shopping"         "automotive"
## [5,] "religion"      "religion"      "health_nutrition" "beauty"
## [6,] "automotive"    "adult"         "religion"         "spam"
## [7,] "709"           "451"           "190"              "313"
##      result.11       result.12          result.13
## [1,] "automotive"    "school"           "sports_fandom"
## [2,] "online_gaming" "politics"         "beauty"
## [3,] "outdoors"      "cooking"          "online_gaming"
## [4,] "parenting"     "health_nutrition" "cooking"
## [5,] "sports_fandom" "automotive"       "health_nutrition"
## [6,] "computers"     "personal_fitness" "fashion"
## [7,] "251"           "255"              "310"
##      result.14          result.15
## [1,] "school"           "politics"
## [2,] "beauty"           "health_nutrition"
## [3,] "religion"         "chatter"
## [4,] "news"             "college_uni"
## [5,] "health_nutrition" "uncategorized"
## [6,] "personal_fitness" "personal_fitness"
## [7,] "334"              "638"
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.