Module - 3: Machine Learning Techniques using R Part-1

Assignment Solution

Version 2.0

1. Text Clustering

Problem Statement:

Budweiser wants to analyse the response posted by people on Twitter for its Super bowl commercial. It is humongous task for them to go through all the tweets.

Complete the following objectives.

Objective 1:

Group tweets in 10 categories (based on content) using K-Means clustering.

Objective 2:

Using K-Means clustering, find tweets which have reference to words "Clydesdale" and "Budweiser"

Dataset Snapshot:

Bud_Chec	Username	screenname	content	tstamp	Buckets
1	The Athletic Mind	AthleticMindset	The Best Super Bowl Ad Meter Winners E	2013-01-23T14:25:52Z	1
1	niddyclub	niddyclub	Super Bowl Ad Meter countdown No. 7: E	2013-01-23T15:16:23Z	1
1	Sure Sport	suresport	Latest: Super Bowl Ad Meter countdown	2013-01-23T15:17:20Z	1
1	News Collections	MrNgekk	Super Bowl Ad Meter countdown No. 7: E	2013-01-23T15:17:23Z	1
1	dumbwire	dumbwire	vox6 - Super Bowl Ad Meter countdown I	2013-01-23T15:26:36Z	1
1	USA TODAY Sports	USATSportsFeed	Super Bowl Ad Meter countdown No. 7: E	2013-01-23T16:17:34Z	1
1	Phil Kent	Tubaphil	I just voted for Budweiser â€œTeam Clyc	2013-01-23T16:22:24Z	1
1	Wade Spruill	RealTallWolf	Anyone else feel like #BudLight predicte	2013-01-23T19:21:10Z	1
1	Erin McGaughey	ErinLee_CCPR	@target to skip #SuperBowl commercial	2013-01-23T20:05:41Z	1
1	Alexander Matray	A_Matray	RT @ChipsNLaccavole: Can Budweiser m	2013-01-23T20:11:42Z	1
1	Linda	lvh429	I just voted for Budweiser Superfan as Su	2013-01-23T20:14:54Z	2
1	Sprayberry Bottle	Sprayberrybot	Bud Black Crown 6%abv launches at SBS.	2013-01-23T20:38:25Z	2
1	FelicityRussellJon	CherryDare	Coca-cola's social media game for Super I	2013-01-23T21:23:33Z	2
1	Becky Rotter	beckylicious152	I just voted for Budweiser Superfan as Su	2013-01-23T23:33:19Z	2
1	stoops	benutty	Are the rumors that Budweiser cast Blue	2013-01-23T23:35:20Z	2
1	teri debruyne	teridebruyne	I just voted for Budweiser Team Clydesd	2013-01-23T23:38:05Z	2
1	Jonathan Gomez	JonathanGomez	If Bud Light is not preparing another "luc	2013-01-24T01:34:29Z	2
1	Will Fyfe	Will_Fyfe8	That bud light commercial with the creer	2013-01-24T02:46:36Z	2

Solution:

```
library(tm)

setwd("D:\\ ")
tweets<-read.csv("Tweets.csv", header=T)
head(tweets)

tweets1<-Corpus(VectorSource(tweets$content))
tweets1
inspect(tweets1)

tweets2<-tm_map(tweets1, tolower)
tweets2
tweets2[[1]]
tweets2[[173]]

tweets3<-tm_map(tweets2, removeWords, stopwords("english"))
tweets3[[173]]

tweets4<-tm_map(tweets3, removePunctuation)
tweets4[[173]]

tweets5<-tm_map(tweets4, removeNumbers)
tweets5[[1]]
tweets5[[171]]

tweets6<-tm_map(tweets5, stripWhitespace)
tweets6[[1]]
tweets6[[171]]

dtm<-DocumentTermMatrix(tweets6)
dtm
inspect(dtm[1:5,1:10])
findFreqTerms(dtm,10)
findFreqTerms(dtm,5)
dtm_tfxidf <- weightTfIdf(dtm)
inspect(dtm_tfxidf[1:5, 100:103])
m <- as.matrix(dtm_tfxidf)
rownames(m) <- 1:nrow(m)
norm_eucl <- function(m) m/apply(m, MARGIN=1, FUN=function(x) sum(x^2)^.5)
m_norm <- norm_eucl(m)
t1 <- kmeans(m_norm, 10)


cl$cluster
cl$size

comments_out<-cbind(as.character(tweets$content),t1$cluster)
```

```
write.csv(comments_out,"Output_clusters.csv")

##For Objective 2:
#It is actually a search use case. You can use K-Means clustering for search.
#All you need to do is add a new row in the tweets.csv which contains words "Clydesdale" and "Budweiser".
#Now, run K-Means clustering again.
#The tweets which go in same cluster as the new row (created as mentioned above), are the ones which actually have reference to words "Clydesdale" and "Budweiser"
```