LAST DM CLASS

Nothing is Impossible. The word itself says “I’m possible”. – Audrey Hepburn

Agenda

Discuss cluster homework

Discuss RMF analysis

Discuss Final Exam

Data set

"gradyear"	"gender"	"age"	"friends"	"basketball"
"football"	"soccer"	"softball"	"bible"	"cheerleading"
"volleyball"	"swimming"	"baseball"	"tennis"	"sports"
"sexy"	"hot"	"kissed"	"dance"	"band"
"marching"	"music"	"rock"	"hair"	"dress"
"god"	"church"	"jesus"	"blonde"	"mall"
"shopping"	"clothes"	"hollister"	"die"	"abercrombie"
"death"	"drunk"	"drugs"	"cute"	"sex"

Preprocessing and results

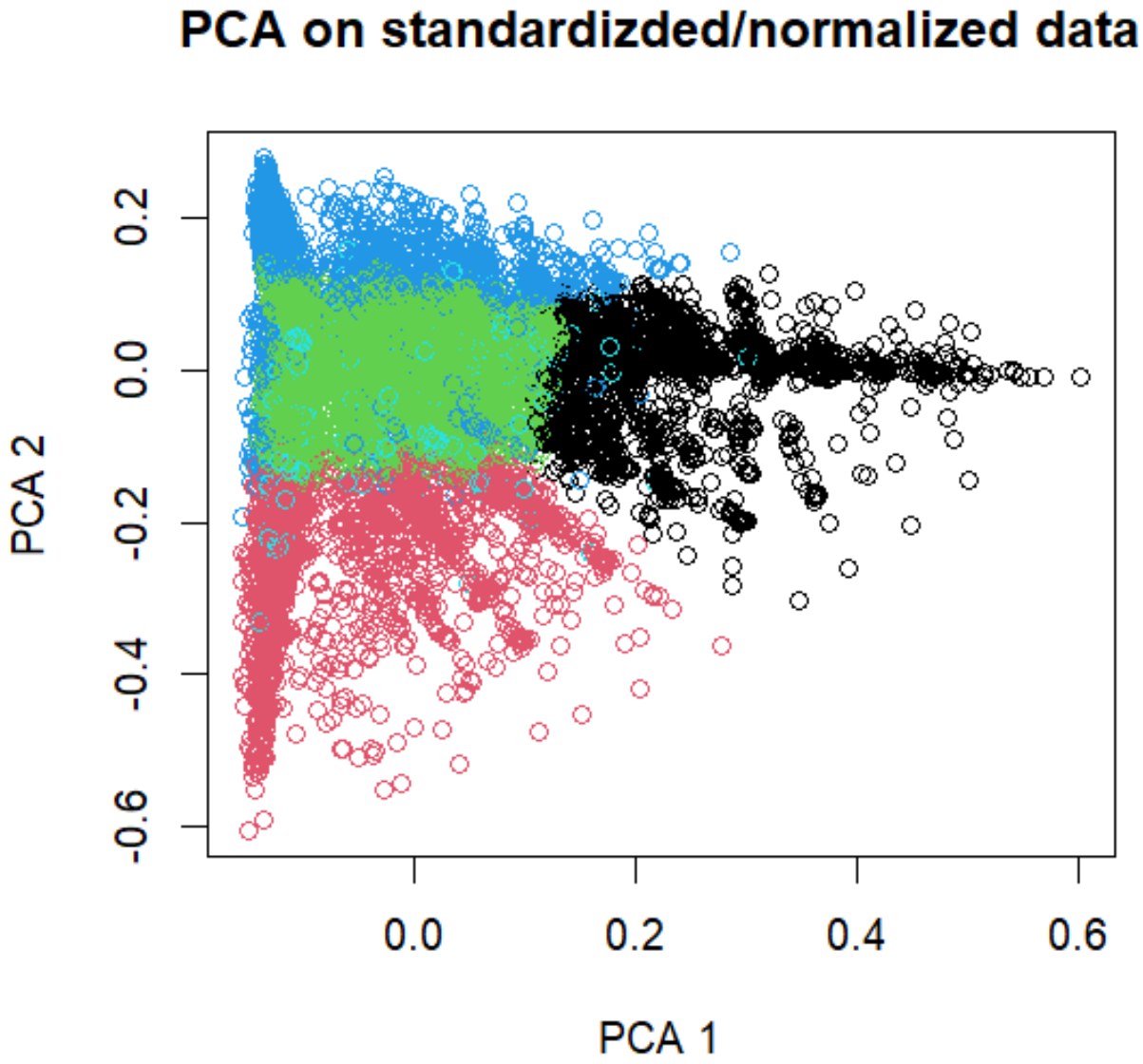
Standardized data by row…each value was a percent of the total words for that individual

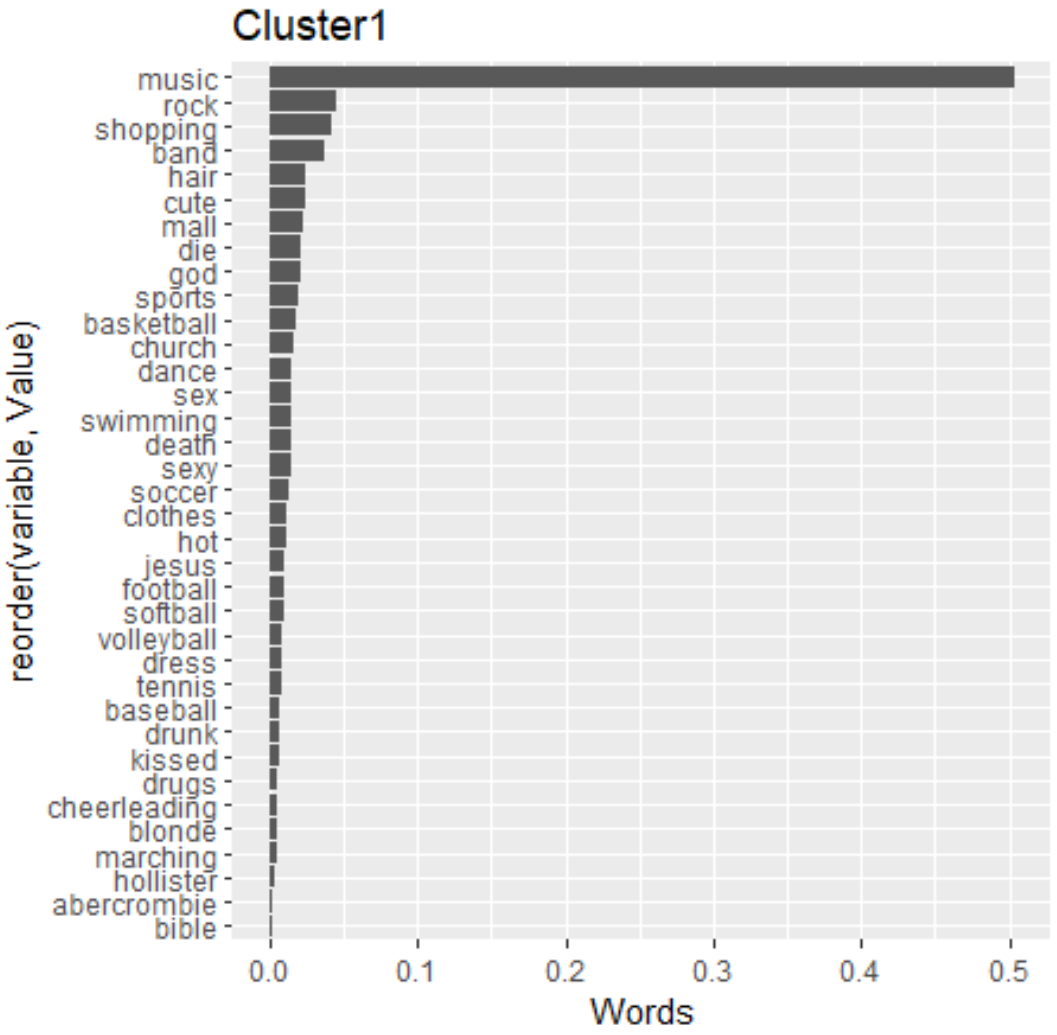
Removed rows that had all 0's

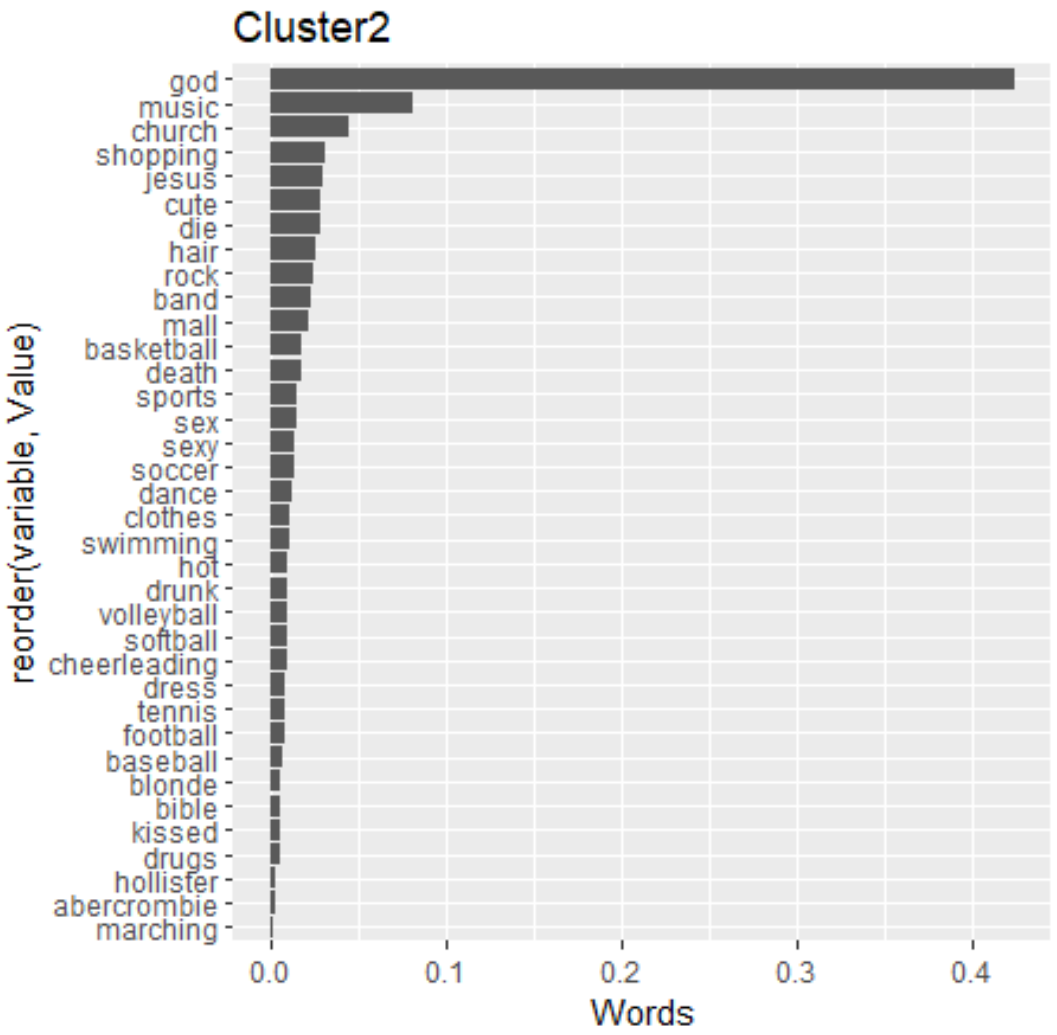
Normalized data by using log(x+1)

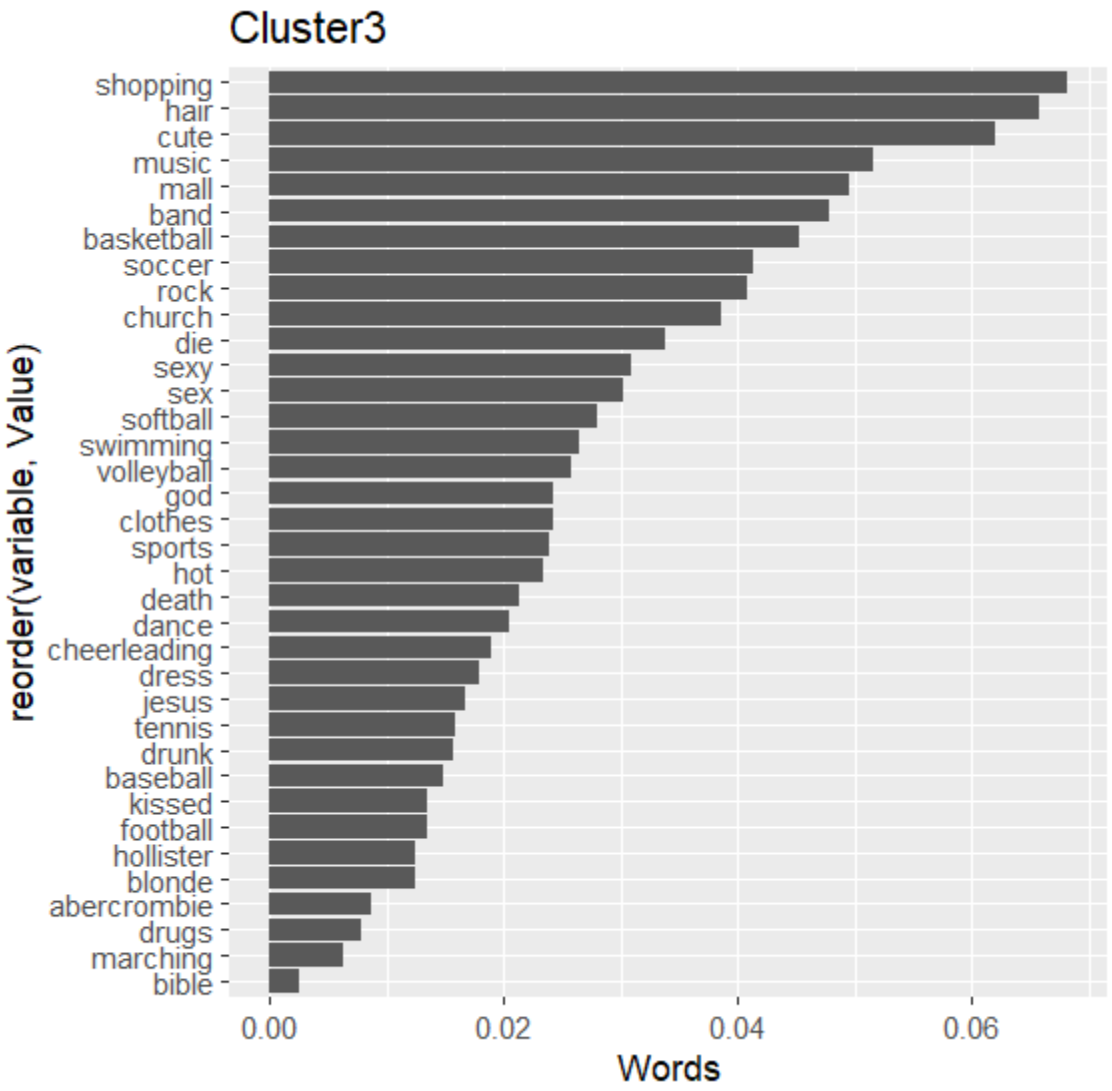
Final solution: k-means with 5 clusters

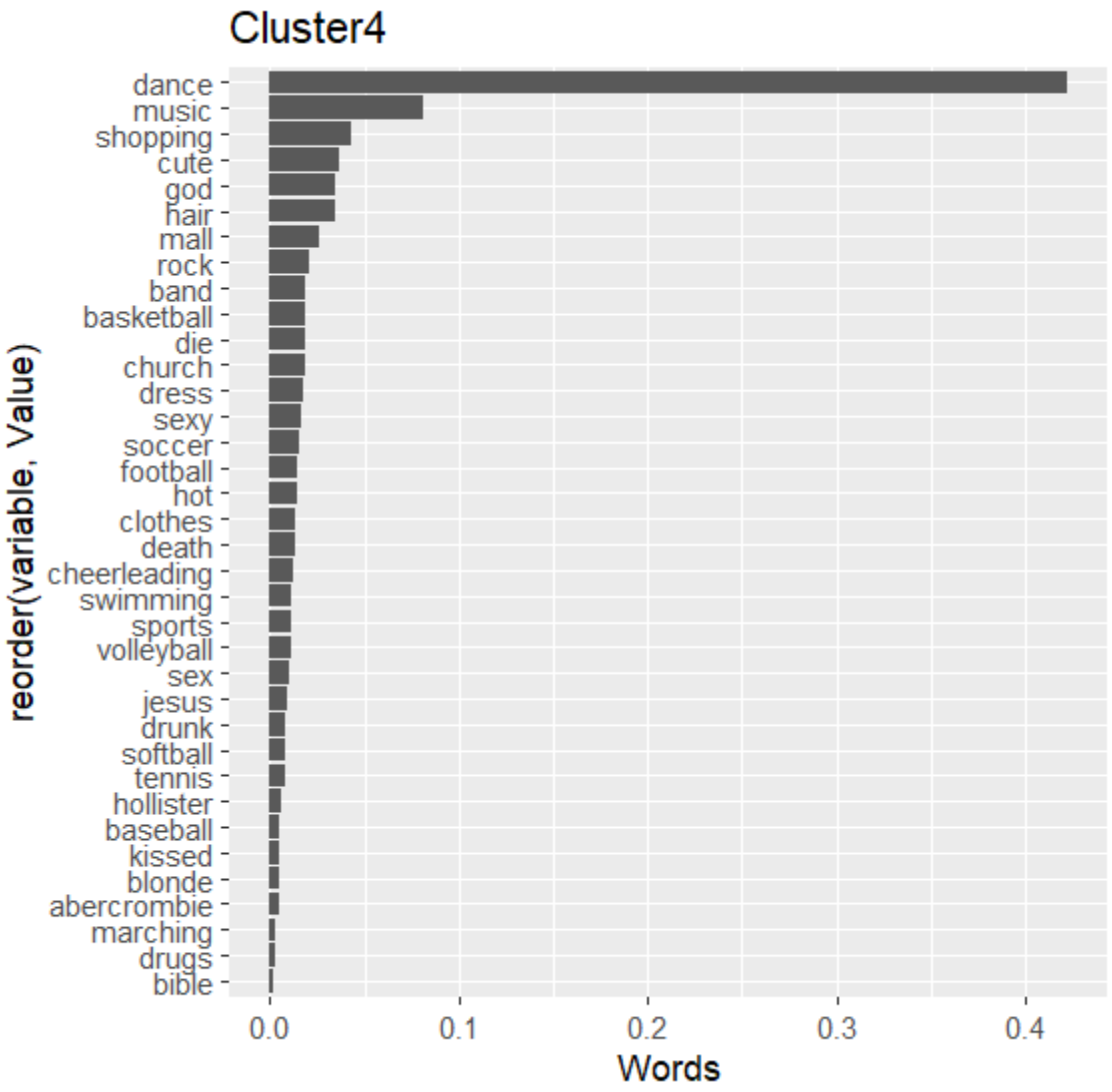
- Cluster 1: Rock-n-Roll Cluster…interested in music (especially rock)
- Cluster 2: Religious cluster….interested in religion
- Cluster 3: Shoppers….although, their interest are all over the place (also has the highest number of "friends")
- Cluster 4: Dancers….interested in dancing
- Cluster 5: Athletes…interested in sports (especially football)

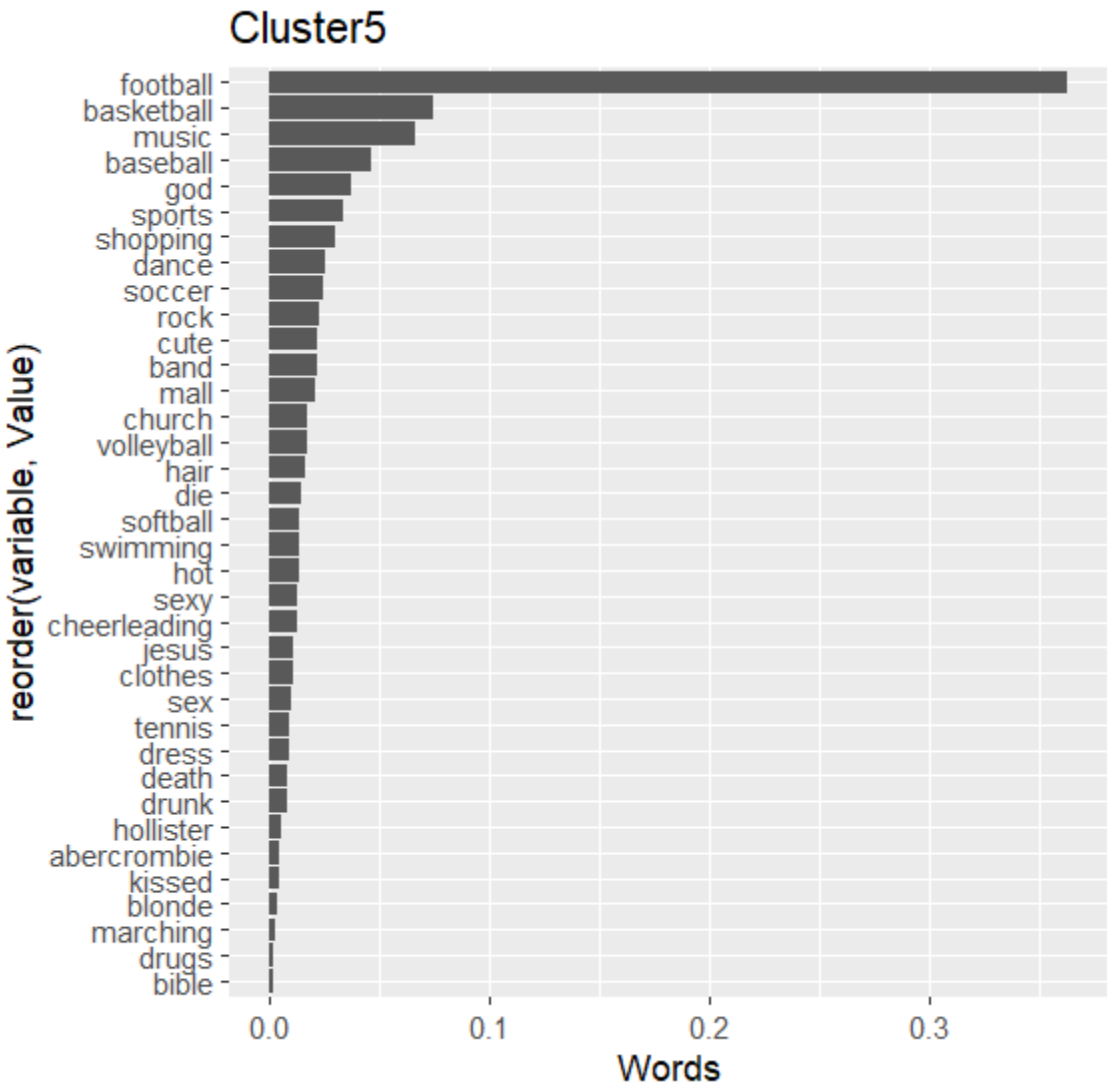

PCA on standardizded/normalized data
PCA 2
0.2
0.0
-0.2
-0.4
-0.6
0.0
0.2
0.4
0.6
PCA 1


Cluster1
reorder(variable, Value)
music
rock
shopping
band
hair
cute
mall
die
god
sports
basketball
church
dance
sex
swimming
death
sexy
soccer
clothes
hot
jesus
football
softball
volleyball
dress
tennis
baseball
drunk
kissed
drugs
cheerleading
blonde
marching
hollister
abercrombie
bible
0.0
0.1
0.2
0.3
0.4
0.5
Words


Cluster2
reorder(variable, Value)
god
music
church
shopping
jesus
cute
die
hair
rock
band
mall
basketball
death
sports
sex
sexy
soccer
dance
clothes
swimming
hot
drunk
volleyball
softball
cheerleading
dress
tennis
football
baseball
blonde
bible
kissed
drugs
hollister
abercrombie
marching
0.0
0.1
0.2
0.3
0.4
Words


Cluster3
shopping
hair
cute
music
mall
band
basketball
soccer
rock
church
die
sexy
sex
softball
swimming
volleyball
god
clothes
sports
hot
death
dance
cheerleading
dress
jesus
tennis
drunk
baseball
kissed
football
hollister
blonde
abercrombie
drugs
marching
bible
reorder(variable, Value)
0.00
0.02
0.04
0.06
Words


Cluster4
dance
music
shopping
cute
god
hair
mall
rock
band
basketball
die
church
dress
sexy
soccer
football
hot
clothes
death
cheerleading
swimming
sports
volleyball
sex
jesus
drunk
softball
tennis
hollister
baseball
kissed
blonde
abercrombie
marching
drugs
bible
reorder(variable, Value)
0.0
0.1
0.2
0.3
0.4
Words


Cluster5
football
basketball
music
baseball
god
sports
shopping
dance
soccer
rock
cute
band
mall
church
volleyball
hair
die
softball
swimming
hot
sexy
cheerleading
jesus
clothes
sex
tennis
dress
death
drunk
hollister
abercrombie
kissed
blonde
marching
drugs
bible
reorder(variable, Value)
0.0
0.1
0.2
0.3
Words

RFM Analysis

Recency, Frequency, Monetary Value analysis

RFM can also be used to segment customers

- Can create features on these ideas and cluster customers on it
- Can give each customer a “score” on each measure and use that to identify “clusters”
  - For example, if scores go from 1-5, a customer that has a 555 is one that made recent purchases, has bought several items and has spent a good amount of money with their purchases
  - Someone who is 551 is an individual that has made recent purchases, has bought several items, but has not spent that much money
  - Need to really think what defines each “level” of these values

Review for exam

Bootstrapping (what it is)

Number of observations per variable

Dealing with transactional data

Dealing with Missing values

Bonferroni correction

FWER and FDR

Calculate support, confidence and lift for Association Analysis (interpret each of these values)

Antecedent, Consequence

Review continued

Decision trees

- Terminology
- Advantages and Disadvantages

Gini, information, SSE criteria

Predicted probabilities, predicted classes, predicted values

How to split categorical variables in a decision tree

Dealing with missing values in a decision tree

What is meant by purity of a node

Pruning and prepruning a tree

Difference between CART and conditional trees

Review continued

Clustering: Hard versus fuzzy

Clustering: Hierarchical versus flat

How does k-means work/different clusters (random seeds)

- Kmeans—converges in small number of iterations
- Advantages/disadvantages of kmeans

Distance measures (Euclidean)

How does hierarchical clustering work

Linkage

Advantages/disadvantages for hierarchical clustering

DBSCAN – what it is and what is it good for

What is variable clustering and what is it used for

Review continued

k-NN

advantages/disadvantages for k-NN

MDS versus PCA

GOF/stress for MDS