|  |  |
| --- | --- |
|  | Project 6 |
|  |  |
|  | Nicolas Eldering  Big Data CSC 643  5/9/18 |

1. **Market Basket Analysis datamining (Apriori) on Databricks using R and/or Spark SQL**

rules<-apriori(df\_trans,

control=list(verbose=F),

parameter=list(supp=0.0001,conf=0.9, target = "rules"))

1. **Generate the top 100 strongest association rules using appropriate support and confidence levels**

Support: inspect(head(sort(rules, by ="support"),100))

Confidence: inspect(head(sort(rules, by ="confidence"),100))

Lift: inspect(head(sort(rules, by ="lift"),100))

1. **Analyze the entire set of association rules by examining the effect of the metrics support, confidence and lift and graph the results**

plot(rules, measure=c("support","confidence"), shading="lift")

1. **Graph the number of association rules generated vs. confidence given a fixed support level**

num\_of\_rules <- vector(mode="integer")

cnt = 1

for (i in seq(from=0.1, to=1, by =0.1)){

rules<-apriori(df\_trans, control = list(verbose=F), parameter= list(supp=0.0001, conf=i))

num\_of\_rules[cnt]= length(rules)

cnt=cnt+1

}

num\_of\_rules

plot(num\_of\_rules, type="o", col="blue", axes =FALSE)

axis(1, at = 1:10, lab=c("0.1","0.2", "0.3","0.4", "0.5","0.6", "0.7","0.8", "0.9","1.0"))

axis(2)

box()

title(main="number of association rules", col.main="blue", font.main=4)

1. **Graph the number of large itemsets found for various support levels**

#large itemsets for support .0001

freq\_itemset <- vector(mode="integer")

cnt = 1

for (i in seq(from=1, to=10, by =1)){

freq\_items <- eclat(df\_trans, parameter = list(supp = .0001, minlen = 1, maxlen = i))

freq\_itemset[cnt]= length(freq\_items)

cnt=cnt+1

}

freq\_itemset

#large itemsets for support .0002

freq\_itemset2 <- vector(mode="integer")

cnt = 1

for (i in seq(from=1, to=10, by =1)){

freq\_items <- eclat(df\_trans, parameter = list(supp = .0002, minlen = 1, maxlen = i))

freq\_itemset2[cnt]= length(freq\_items)

cnt=cnt+1

}

freq\_itemset2

#large itemsets for support .0002

freq\_itemset3 <- vector(mode="integer")

cnt = 1

for (i in seq(from=1, to=10, by =1)){

freq\_items <- eclat(df\_trans, parameter = list(supp = .0003, minlen = 1, maxlen = i))

freq\_itemset3[cnt]= length(freq\_items)

cnt=cnt+1

}

freq\_itemset3

#plot the large itemsets: blue is .0001, red is .0002, yellow is .0003 support, x axis is max length of large itemset

plot(freq\_itemset, type="o", col="blue", axes =FALSE)

axis(1, at = 1:10, lab=c("1","2", "3","4", "5","6", "7","8", "9","10"))

axis(2)

lines(freq\_itemset2, type = "o", col = "red")

lines(freq\_itemset3, type = "o", col = "yellow")

box()

title(main="Large Itemsets", col.main="Red", font.main=4)