EDA_sentiment_stanford

Sandar Felicity Lim

5/20/2018

Note

20/5/2018: With super small amount of stories (n=100). Will run the analysis with larger dataset later

data

```
dat<-read.table("/Users/sandar/nlu-project2/data/Stanford_highscores.txt", sep=",", col.names = c("titl
head(dat)
```

```
##   title s1 s2 s3 s4 Class
## 1     2  3  3  2  3     3
## 2     2  1  2  1  3     1
## 3     2  3  1  2  1     1
## 4     2  2  2  2  2     2
## 5     1  1  2  3  0     3
## 6     2  2  2  2  1     1
```

legend

Very sad = 0 sad = 1 neutral = 2 happy = 3 very happy = 4

stratified sampling

```
set.seed(123)
train.index <- createDataPartition(dat$Class, p = .7, list = FALSE)
train <- dat[ train.index,]
test  <- dat[-train.index,]
test2 <- test[,-6]
```

Check 1: Linear Discriminant Analysis

```
m1 <- lda(Class~., data = train)
m1
```

```
## Call:
## lda(Class ~ ., data = train)
##
## Prior probabilities of groups:
##          1          2          3          4
## 0.35211268 0.29577465 0.33802817 0.01408451
##
## Group means:
```

```
##      title       s1       s2       s3       s4
## 1 2.000000 1.920000 2.000000 1.560000 1.880000
## 2 2.047619 1.904762 1.952381 1.952381 1.952381
## 3 2.000000 1.916667 2.208333 2.125000 1.875000
## 4 2.000000 2.000000 1.000000 2.000000 3.000000
##
## Coefficients of linear discriminants:
##               LD1        LD2         LD3
## title  0.03198289 -0.6078592 -2.73073888
## s1     0.10221892 -0.1789778  0.35828790
## s2    -0.39534743  1.0210803  0.31265835
## s3    -1.30602126 -0.5805287 -0.04682902
## s4     0.41179308 -0.7445015  0.52618299
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.6043 0.3662 0.0295
```

```
test$predict <- predict(m1,newdata=test2)$class

classError(test$predict,test$Class)
```

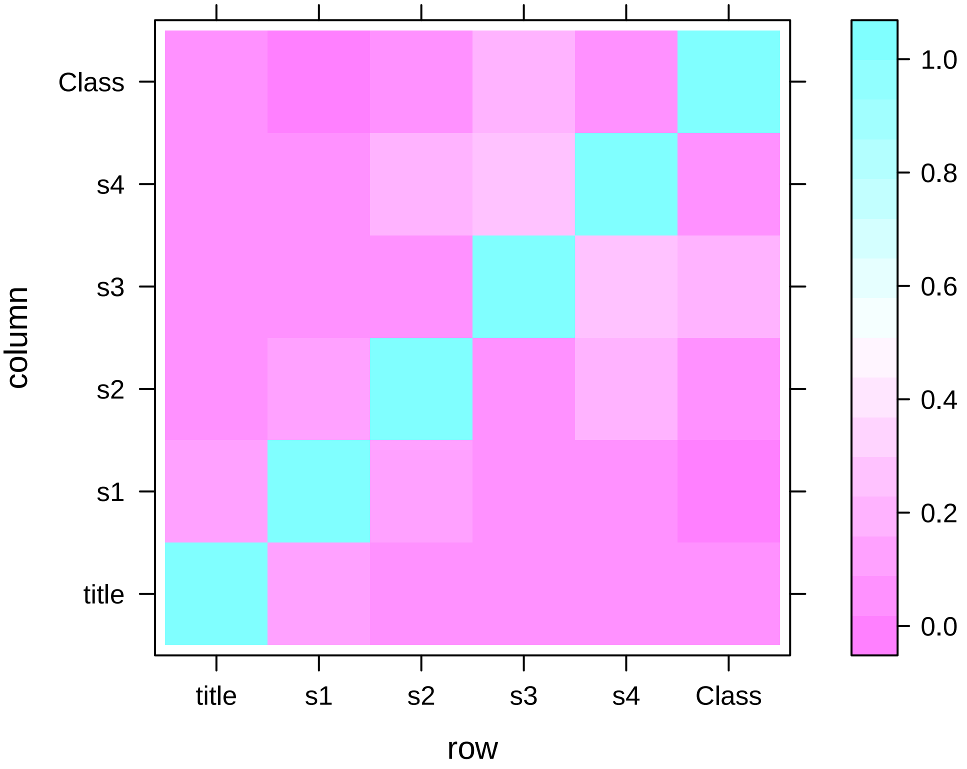
```
## $misclassified
##  [1]  2  4  5  7 11 12 14 15 17 18 20 24 25 27 28
##
## $errorRate
## [1] 0.5172414
```

Check 2: Correlations

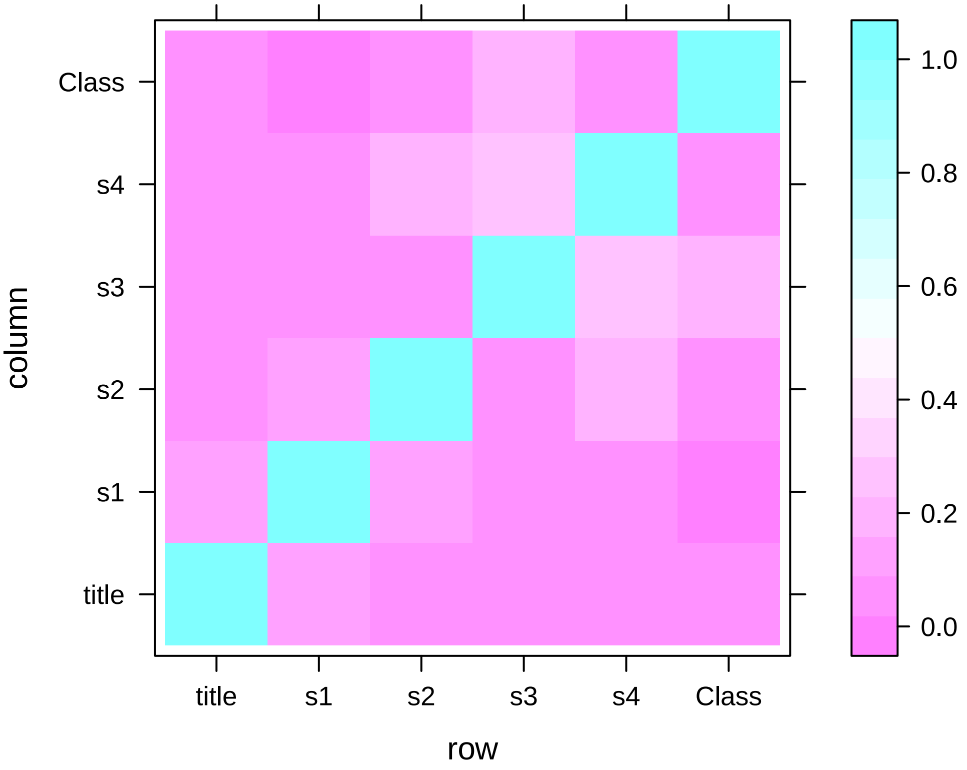
```
cor1<-as.matrix(cor(dat))
cor2<-cor(filter(dat,Class==3|4))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

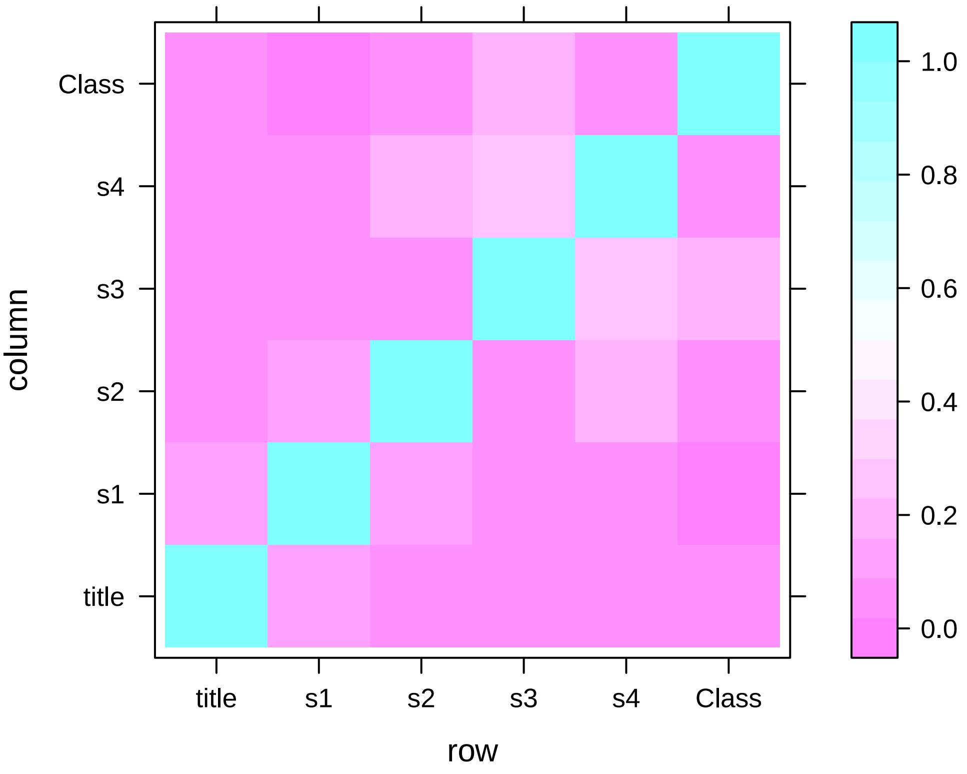
```
cor3<-cor(filter(dat,Class==0|1))

levelplot(cor1)
```



```
#happy ending:
levelplot(cor2)
```



```
#sad ending:
levelplot(cor3)
```



EDA 1: Longitudinal clustering

```
comp.dat <- data.frame(ID=seq(1,100,1), dat )
comp.time <- data.frame(ID=seq(1,100,1), title =0, s1= 1, s2=2, s3=3,s4=4, Class=5)
s1 = step1measures(comp.dat, comp.time, ID = TRUE)
```

```
## [1] "Correlation of m5 and m6 : 1"
## [1] "Correlation of m12 and m13 : 1"
```

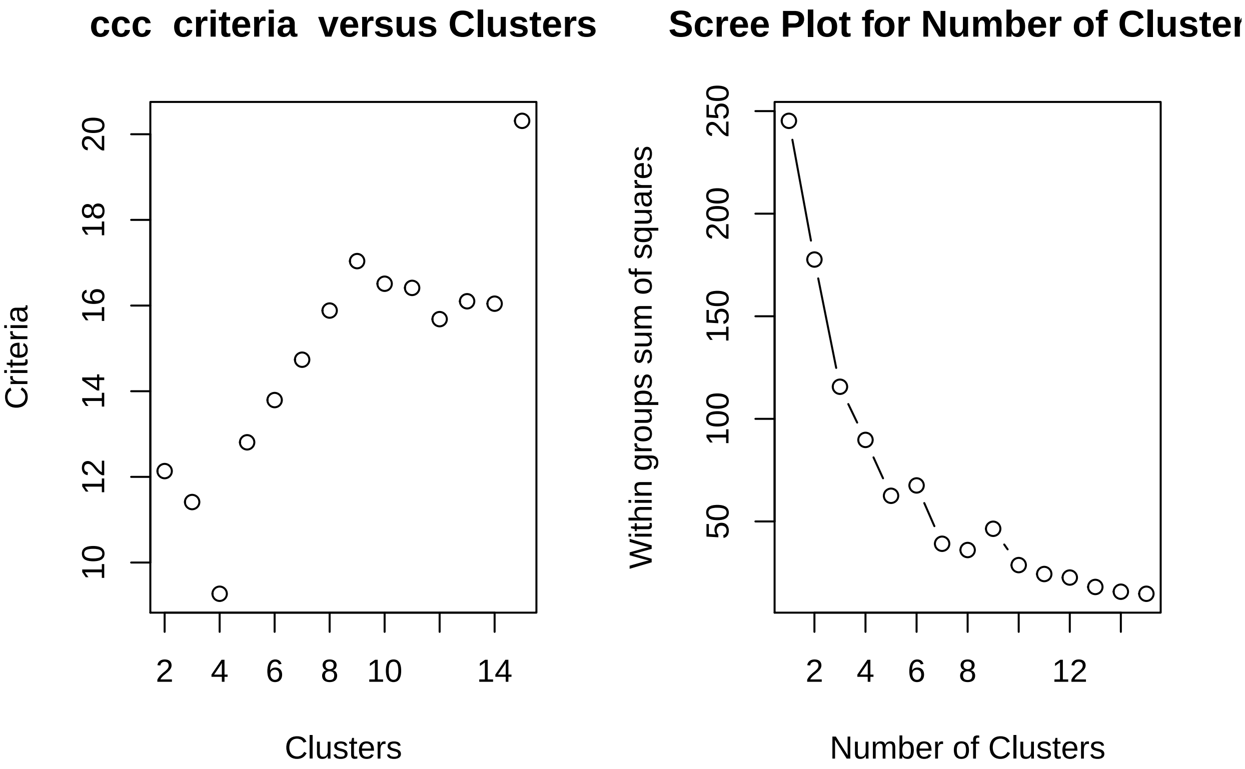
```
s2 = step2factors(s1)
```

```
## [1] "m6 is removed because it is perfectly correlated with m5"
## [2] "m13 is removed because it is perfectly correlated with m12"
## [1] "Computing reduced correlation e-values..."
```

```
s2 = step2factors(s1)
```

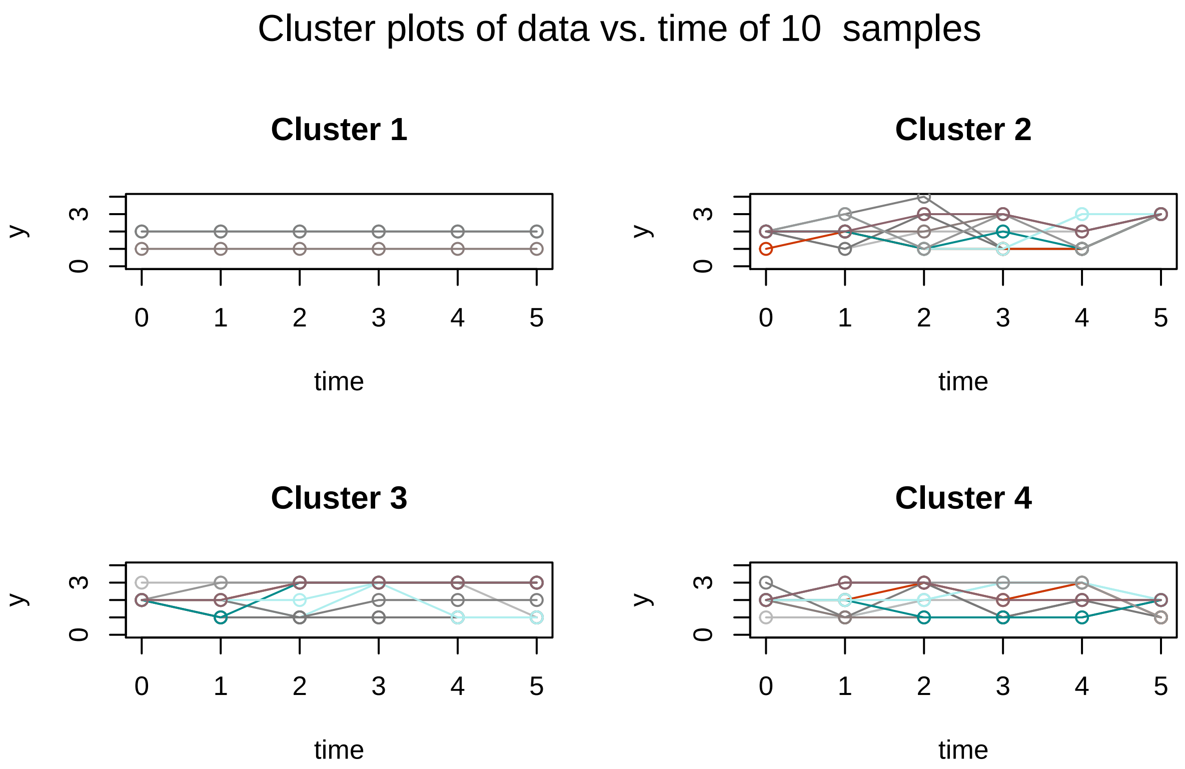
```
## [1] "m6 is removed because it is perfectly correlated with m5"
## [2] "m13 is removed because it is perfectly correlated with m12"
## [1] "Computing reduced correlation e-values..."
```

```
step3clusters(s2,  nclusters = NULL,nstart=50,
              criteria = "ccc", forced.factors = NULL)
```



```
## Number of observations:  100
##
## Cluster distribution:
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
##  3  9  8 14  8  8  6  1  9  4  1  7 11  7  4
##
## Measures with max.loading in factors:  m2 m5 m12 m23
##
## If you report these results, please cite:
## Sylvestre MP, et al. (2006). Classification of patterns of delirium severity scores over time in an
## International Psychogeriatrics,18(4), 667-680. doi:10.1017/S1041610206003334.
```

```
#4 clusters look good
s3 = step3clusters(s2,  nclusters = 4)
plot(s3)
```



EDA 2: K-Means for Longitudinal Data

```
colnames(comp.dat)<- c("ID","s0", "s1","s2","s3","s4","s5")
dat.long<- melt(comp.dat, idvar="ID", measure.vars =c("s0", "s1","s2","s3","s4","s5") )
cld1 <-clusterLongData(traj=dat,idAll = seq(1,100,1), time=c(1,2,3,4,5,6))
kml(cld1,3,2)
```

```
##  ~ Fast KmL ~
## **S
```