Practical session: 02 (01)

Preliminaries

1. Setting the working directory (the directory where you kept your dataset)

## For example :

## **setwd("C:\\Users\\mullah\\Documents\\mullah-postdoc\\academic\\UT1\\UT1\_lecture\\2020\\lecture-02\_01\\scripts\_data")**

*Change according to your working directory*

2. Loading features

## **features <- read.table("output/topics\_features\_sum\_mean\_max\_std\_var", header=TRUE, row.names=1, sep="\t")**

3. Listing the row names

## rownames(features)

4. List the column names

## colnames(features)

5. Total rows or columns names

## total\_rows <- length(rownames(features))

## total\_column <- length(colnames(features))

## dim(features)

6. Descriptive statistics (AKA, Aggregation functions)

*mean, sd, var, min, max, median, range, and quantile*.

To get a better idea of the distribution of your exploratory variables, you can use the summary() function like this:

## summary(features)

To get a particular statistics, you can apply the following functions:

## mean.features <- apply(features, 2, mean, na.rm=TRUE)

#considering null values in the data (na.rm=TRUE)

summary(mean.features)

7. Selecting some subset of columns

## sum\_variants <- features[,2:5]

## mean\_variants <- features[,6:9]

## ql <- features[,c(“DL”)]

## ql <- features[,1]

ql <- features[,1, drop = FALSE]

**Data Visualization**

############### Scatter plot ########################

Showing scatter plot of the sum variants features

## plot(sum\_variants, col="red")

## plot(sum\_variants [,1], sum\_variants [,2])

############### Bar plot ##################################

Showing bar plot for all the mean variants features

## **barplot(apply(mean\_variants, 2, mean, na.rm=TRUE), las=2)**

Showing bar plot of the Std variants features

## barplot(…..)

############### Box plot #############################

Showing box plot for the mean variants features

## boxplot(mean\_variants, las=2)

############### Heat map #############################

Showing the heat map of the max variants features (convert to matrix first)

## max\_variants <- features[, c(10:13)]

## heatmap(as.matrix(max\_variants))

############### Correlogram plot ############################

correlogram #method="circle, number, color, pie, ellipse, shade

## library(corrplot)

## corrplot(cor(features), tl.cex=0.75, method='circle')

## corrplot(cor(features), type = "lower", tl.cex=0.75, method='circle')

## corrplot(cor(features), type = "upper", tl.cex=0.75, method='circle')

## corrplot(cor(features), type = "upper", tl.cex=0.75, method='pie')

## corrplot(cor(features), type = "lower", tl.cex=0.75, method='pie')

The correlation matrix can be reordered according to the correlation coefficient.

This is important to identify the hidden structure and pattern in the matrix. There are four methods in corrplot (parameter order), named "AOE", "FPC", "hclust", "alphabet"

## corrplot(cor(features), type = "lower", tl.cex=0.75, method='pie', order="AOE")

## corrplot(cor(features), type = "lower", tl.cex=0.75, method='ellipse', order="AOE")

**Exercises:**

1. Plot the Box and Scatter plot for variance variants features
2. Plot Correlogram for different standard deviation and variance variance features
3. Plot Heat map for the variants of “Max” features

**Correlation analysis**

#Pearson correlation coefficient

## cor.test(x, y, method = "pearson", use = "pairwise.complete.obs")

#Spearman’s rank correlation coefficient

## cor.test(x, y, method = "spearman", use = "pairwise.complete.obs")

#kendall tau correlation coefficient

## cor.test(x, y, method = "kendall", use = "pairwise.complete.obs")

Taking two features

## x <- features$STD\_TF

## y <- features$STD\_TFIDF

## cor.test(x, y, method = "pearson", use = "pairwise.complete.obs")

## cor.test(x, y, method = "spearman", use = "pairwise.complete.obs")

## cor.test(x, y, method = "kendall", use = "pairwise.complete.obs")