Lab 10. Dimensionality reduction. PCA. t-SNE

```
library(tidyverse)
library(ggfortify)
#Sys.setlocale(locale = "ru_RU.UTF-8")
```

Principal component analysis (PCA)

1 Gospels' frequency word lists

The gospels of Matthew, Mark, and Luke are referred to as the Synoptic Gospels and stand in contrast to John, whose content is comparatively distinct. This dataset (https://tinyurl.com/y8tcf3uw (https://tinyurl.com/y8tcf3uw)) contains frequency of selected words (without stopwords, without pronouns and without frequent word "Jesus") as attested in four gospels of the New Testament.

For some visualisations you will need assign row names to the dataframe:

```
gospels <- read.csv("https://tinyurl.com/y8tcf3uw")
row.names(gospels) <- gospels$word
```

1.1 Apply PCA to four continuous variables. What is the cumulative proportion of explained variance for the first and second component?

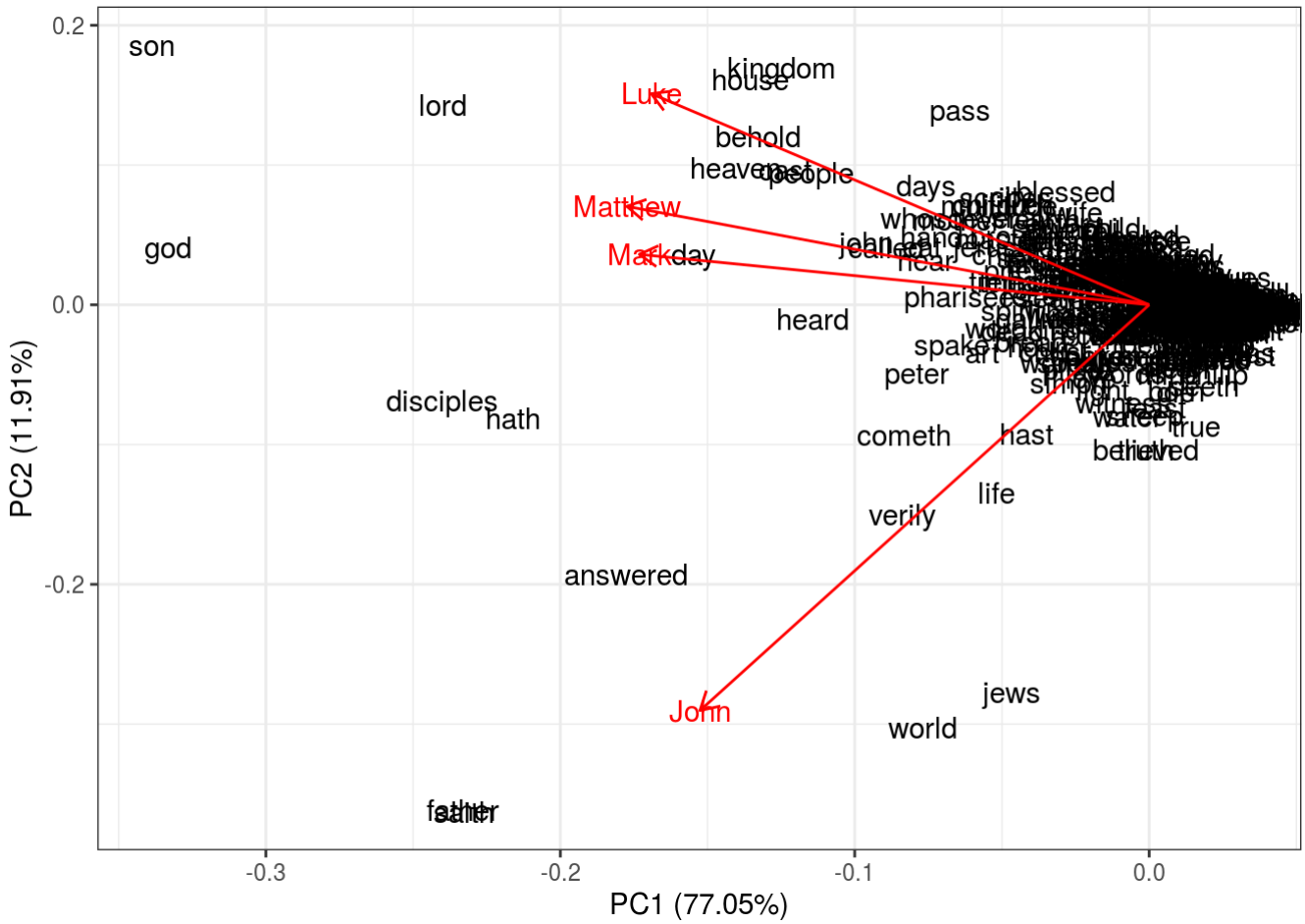
```
PCA <- prcomp(gospels[,2:5], center = TRUE, scale. = TRUE)
summary(PCA)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7556 0.6903 0.50983 0.42619
## Proportion of Variance 0.7705 0.1191 0.06498 0.04541
## Cumulative Proportion  0.7705 0.8896 0.95459 1.00000
```

1.2 Use the `autoplot()` function of the library ggfortify for creating plot like this.

See more examples here: https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html (https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html)

```
autoplot(PCA,
         shape = FALSE,
         loadings = TRUE,
         label = TRUE,
         loadings.label = TRUE)+
  theme_bw()
```



1.3 Predict the coordinates for the word “Jesus”, which have the following frequencies: John = 0.05, Luke = 0.01, Mark = 0.02, Matthew = 0.02.

```
predict(PCA, data.frame(John = 0.05, Luke = 0.01, Mark = 0.02, Matthew = 0.02))
```

```
##             PC1       PC2       PC3       PC4
## [1,] -22.60497 -9.599171 2.367918 2.104944
```