Lab 12. Claster analysis

```
library(tidyverse)
```

1. Baltic Swadesh lists

The dataset contains results of the comparison of the Swadesh lists from three Baltic languages (Old Prussian, Lithuanian and Latvian) and outlier. 1 — means the same cognates. Calculate a distance matrix.

1.1 What is the distance between Lithuanian and Latvian? Use dist() function.

1.2 Between which two languages the distance is minimum?

1. 3 What does it mean? Why there is a minimum distance between that languages?

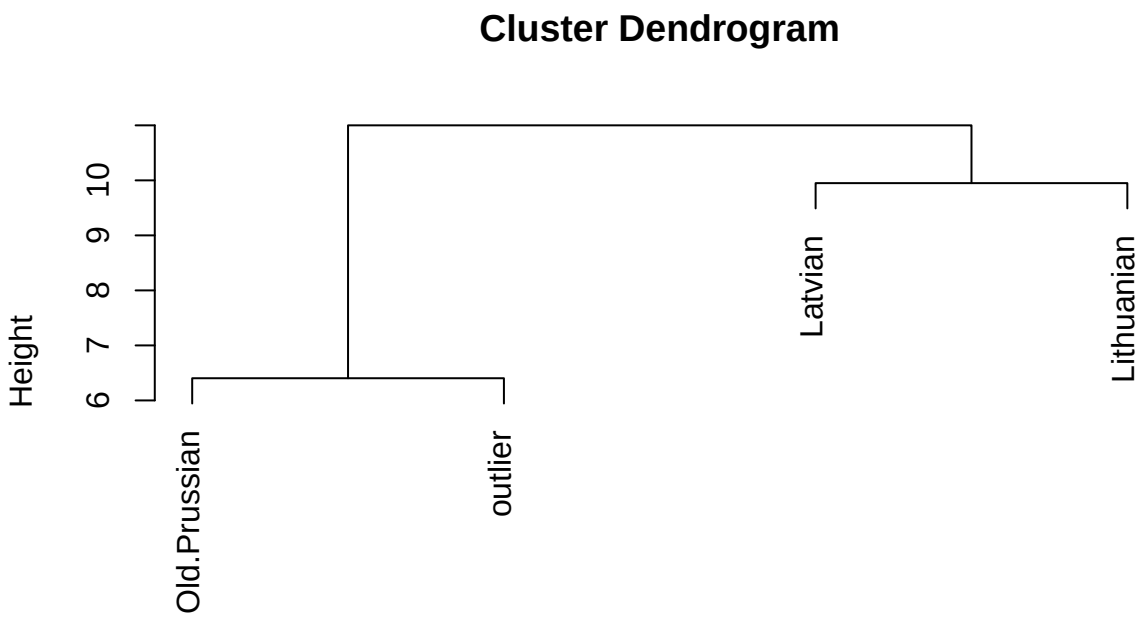
1.4 Will the distances changed, if you remove an outlier column?

1.5 Make a k-means clustering with k = 3. What languages are in the same cluster? Use kmeans() function and set.seed(42).

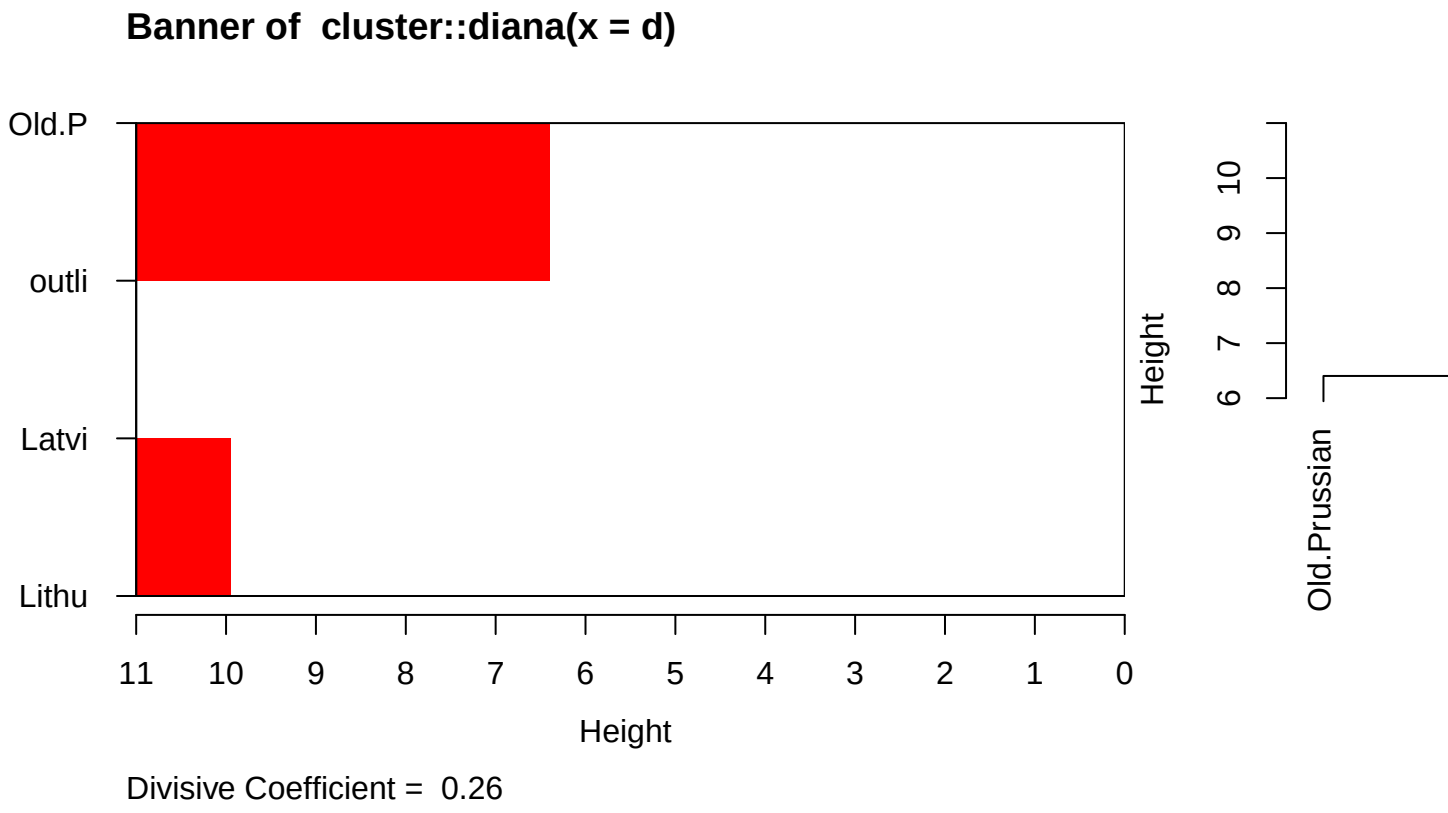
1.6 Does clusterization made with hierarchical algorithm differ from the results obtained by k-means? Use hclust() function.

1.7 Plot the dendrogram which use the plot() function. 1.8 Compute a divisive hierarchical clustering using diana function from cluster package for our dataset. Are there any differences with results obtained by agglomerative clustering algorithm?

```
df <- read.csv("https://raw.githubusercontent.com/agricolamz/2018-MAG_R_course/master/data/baltic.csv")
d <- dist(t(df[,3:6]))
set.seed(42)
k <- kmeans(d, 3)
hc <- hclust(d)
diana <- cluster::diana(d)
plot(hc)
```


Cluster Dendrogram
Height
6
7
8
9
10
Old.Prussian
outlier
Latvian
Lithuanian
d
hclust (*, "complete")

```
plot(diana)
```


Banner of cluster::diana(x = d)
Old.P
outli
Latvi
Lithu
11
10
9
8
7
6
5
4
3
2
1
0
Height
Divisive Coefficient = 0.26
Height
6
7
8
9
10
Old.Prussian

2. Orientation set

This set is based on (Chi-kuk 2007). Experiment consisted of a perception and judgment test aimed at measuring the correlation between acoustic cues and perceived sexual orientation. Naïve Cantonese speakers were asked to listen to the Cantonese speech samples collected in Experiment and judge whether the speakers were gay or heterosexual. There are 14 speakers and following parameters: * [s] duration (s.duration.ms) * vowel duration (vowel.duration.ms) * fundamental frequencies mean (F0) (average.f0.Hz) * fundamental frequencies range (f0.range.Hz) * percentage of homosexual impression (perceived.as.homo) * percentage of heterosexal impression (perceived.as.hetero) * speakers' orientation (orientation) * speakers' age (age)

Make a distance matrix bassed on age, s.duration.ms, vowel.duration.ms, average.f0.Hz and f0.range.Hz.

2.1 Between which two speakers the distance is minimum?

2.2 Make k-means clusterisation with k = 2. Are selected parameters good for distinguishing informants of the different orientation? Use set.seed(42).

2.3 Fast cluster visualisation: plot(my_df, col = km$cluster)

2.4 What is the mean and standard deviation of the variable perceived.as.homo.percent within cluster 1.

2.5 Make hierarchical clustering . Which informant is in the same cluster with informant L? Use plot() function.

2.5 Make hierarchical clustering . Which informant is in the same cluster with informant L? Use plot() function.

2.6 Compute a divisive hierarchical clustering using diana function from cluster package for our dataset. Are there any differences with results obtained by agglomerative clustering algorithm?

`pvclust()` provides two types of p-values: AU (Approximately Unbiased) p-value and BP (Bootstrap Probability) value. AU p-value, which is computed by multiscale bootstrap resampling, is a better approximation to unbiased p-value than BP value computed by normal bootstrap resampling. Note that the interpretation of the p-value is different here from our previous case studies: the closer the p-value to 1, the more empirical support the cluster has. The AU value is considered to be the more precise measure.

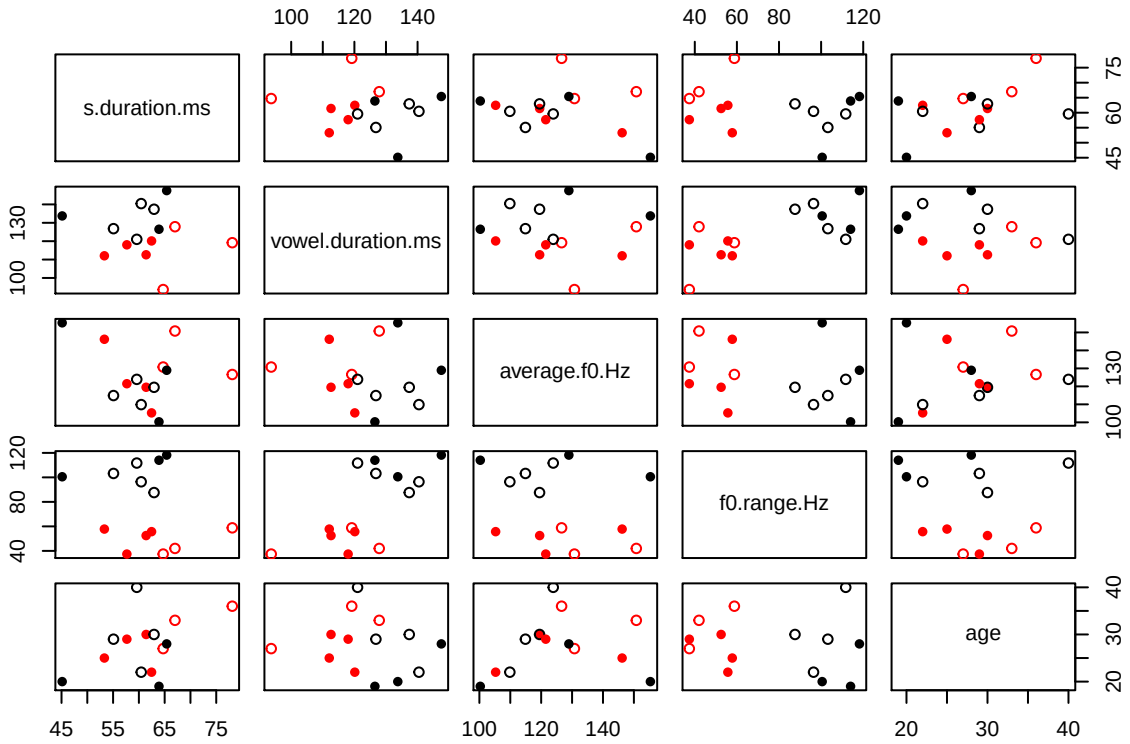
```
df <- read.csv("https://raw.githubusercontent.com/LingData2019/LingData2020/master/data/orientation.csv
row.names(df) <- df$speaker
d <- dist((df[,-c(6:9, 1)]))
min(d)
```

```
## [1] 15.79903
```

```
set.seed(42)
km <- kmeans(d, 2)
cbind.data.frame(cluster = km$cluster, orientation = df$orientation) %>%
  count(cluster, orientation)
```

```
## # A tibble: 4 x 3
##   cluster orientation     n
##     <int> <fct>       <int>
## 1       1 hetero          3
## 2       1 homo            4
## 3       2 hetero          4
## 4       2 homo            3
```

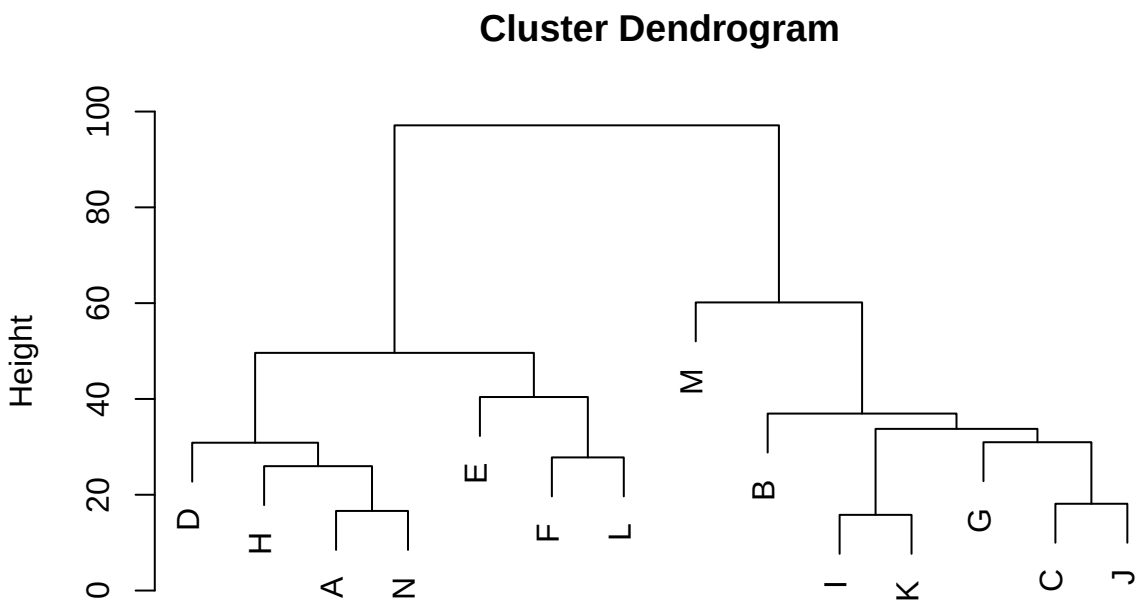
```
plot(df[,-c(6:9, 1)], col = km$cluster, pch = c(16, 1)[df$orientation])
```



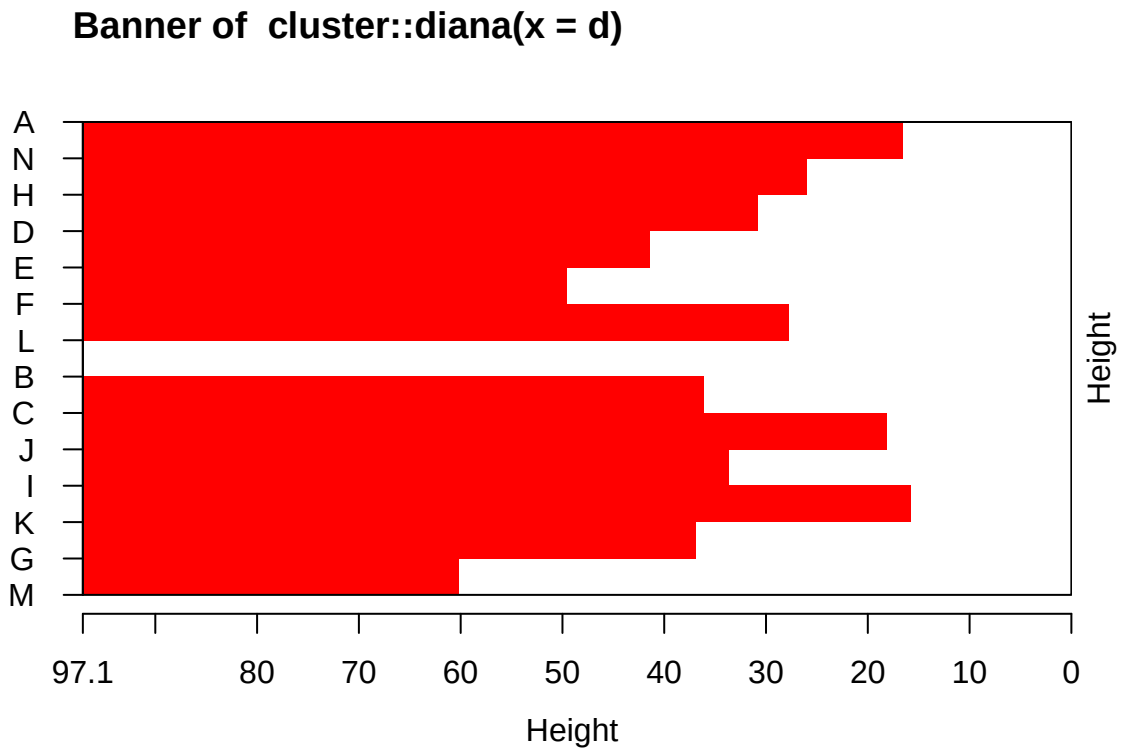

```
df %>%
  mutate(cluster = km$cluster) %>%
  group_by(cluster) %>%
  summarise(mean = mean(perceived.as.homo.percent),
            sd = sd(perceived.as.homo.percent))
```

```
## # A tibble: 2 x 3
##   cluster  mean    sd
##     <int> <dbl> <dbl>
## 1       1 0.611 0.248
## 2       2 0.469 0.244
```

```
hc <- hclust(d)
plot(hc)
```




```
diana <- cluster::diana(d)
plot(diana)
```



Divisive Coefficient = 0.71

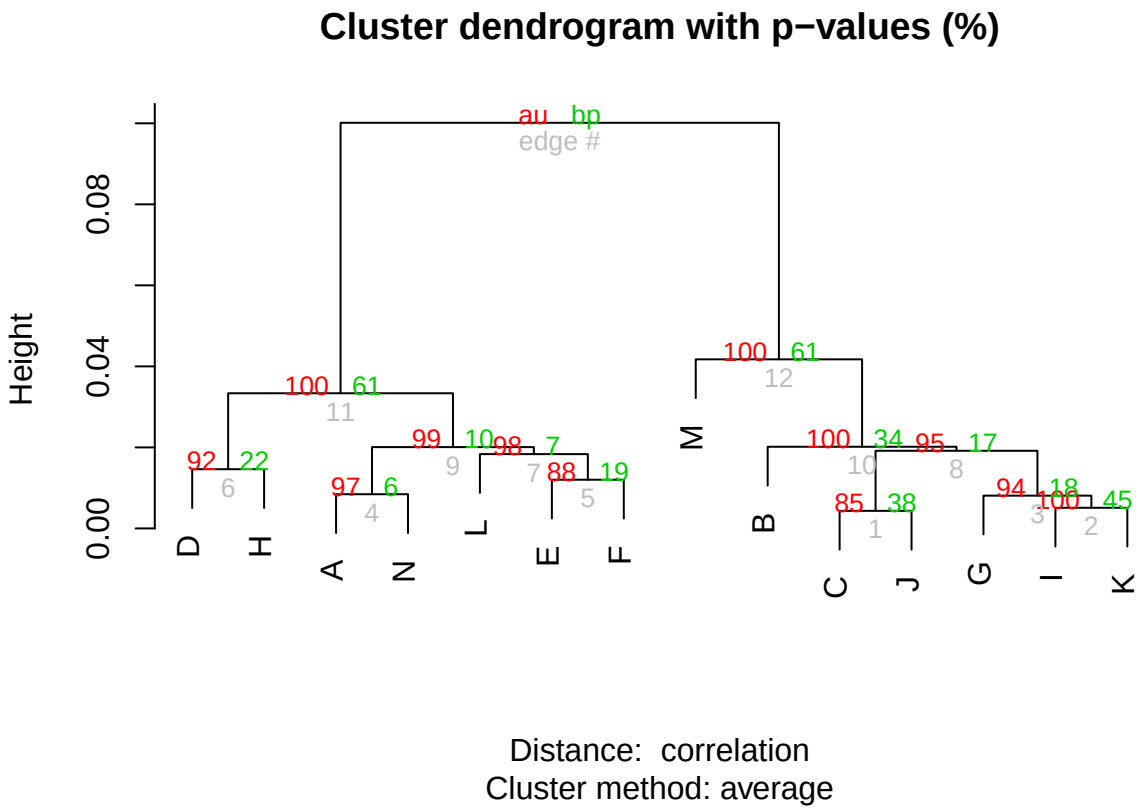
```
p <- pvclust::pvclust(t(df[,-c(1, 9, 10)]),
                      method.dist="cor",
                      method.hclust="average",
                      nboot=100)
```

```
## Bootstrap (r = 0.43)... Done.
```

```
## Warning: inappropriate distance matrices are omitted in computation: r =
## 0.428571428571429
```

```
## Bootstrap (r = 0.57)... Done.
## Bootstrap (r = 0.71)... Done.
## Bootstrap (r = 0.86)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.14)... Done.
## Bootstrap (r = 1.29)... Done.
```

```
plot(p)
```



3. Lexical contact in Daghestan

Data by Misha Daniel, Iliya Chechuro, Samira Verhees.

There 225 nouns coming from mid-range of the The World Loanword Database list (http://wold.clld.org/) collected from Archi, Bezhta and Avar language. Since it is possible to have multiple words with selected meaning and some words could appear to be in different languages they were grouped into cognate sets.

Create row names by following command

```
row.names(data) <- paste(1:294, df$english)
```

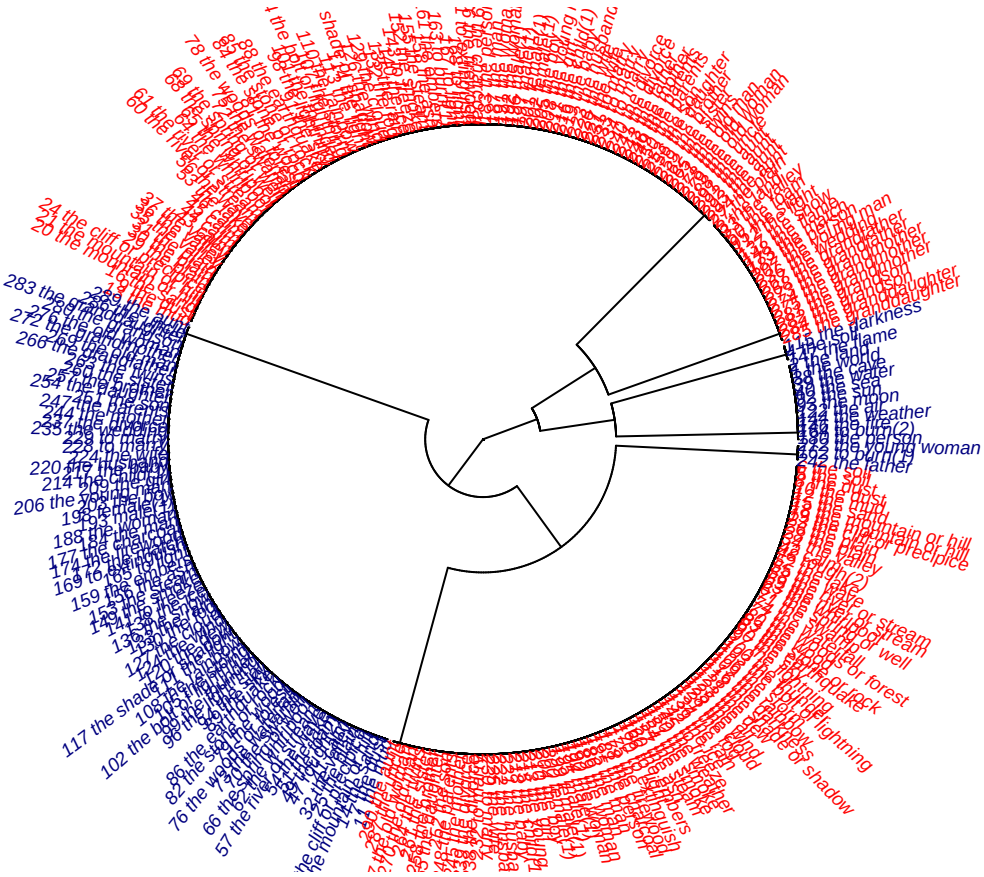
3.1 Make hierarchical clustering for this data and visualise it.

3.2 Play around with different types: “phylogram”, “cladogram”, “unrooted”, “radial”

3.3 Play around with different Avar and Bezhta variables.

3.4 Think about clusters that appeared from these clusterisation. What could we say about cognate sets?

```
library(ape)
df <- read.csv("https://goo.gl/4sJqv1")
data <- df[,-c(1:3)]
row.names(data) <- paste(1:294, df$english)
data %>%
  dist() %>%
  hclust() ->
  hc
plot(as.phylo(hc),
     type = "fan",
     cex = 0.6,
     no.margin = TRUE,
     tip.color = c("red",
                   "navy")[as.factor(df$archi)])
```



- blue are the Bezhta unique words
- red are the Archi unique words