Discovery_dataset

Ariel-ac4391

11/22/2018

Here, I recapitulate the main step related in the research paper with the gaphes associated

The first step is data cleansing :

```
training_data=read.csv("data/Data_User_Modeling_training_Dataset.csv")
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

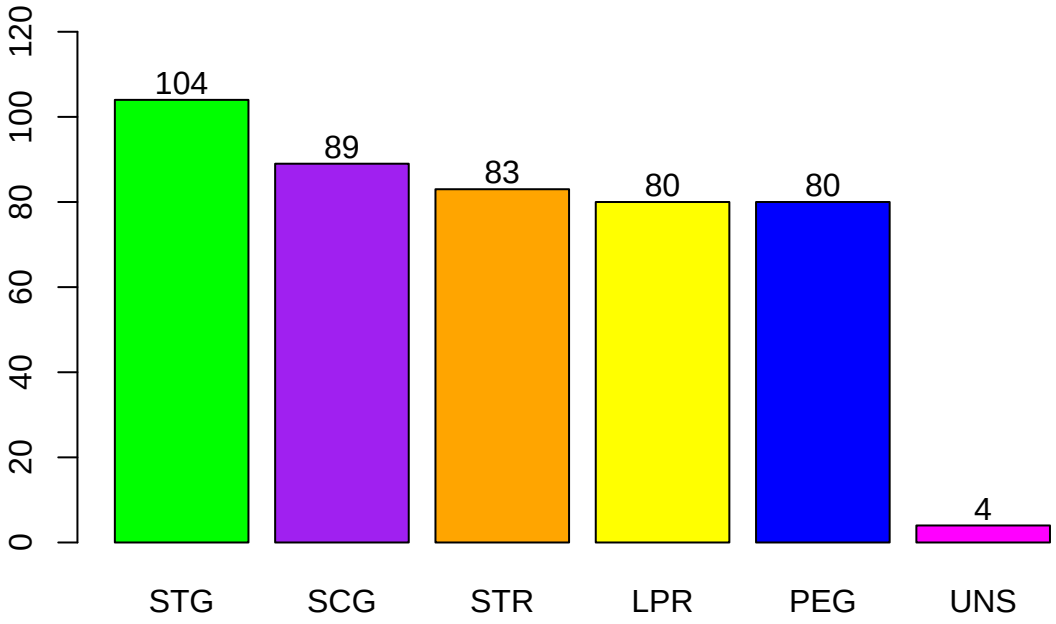
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
summary(training_data)
```

```
##       STG              SCG              STR              LPR
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2407   1st Qu.:0.2100   1st Qu.:0.2913   1st Qu.:0.2500
##  Median :0.3270   Median :0.3025   Median :0.4900   Median :0.3300
##  Mean   :0.3711   Mean   :0.3557   Mean   :0.4680   Mean   :0.4327
##  3rd Qu.:0.4950   3rd Qu.:0.4975   3rd Qu.:0.6900   3rd Qu.:0.6475
##  Max.   :0.9900   Max.   :0.9000   Max.   :0.9500   Max.   :0.9900
##       PEG              UNS
##  Min.   :0.0000   High    :63
##  1st Qu.:0.2500   Low     :83
##  Median :0.5000   Middle  :88
##  Mean   :0.4585   very_low:24
##  3rd Qu.:0.6600
##  Max.   :0.9300
```

```
attach(training_data)

# Number of distinct values in each feture
a = n_distinct(STG)
b = n_distinct(SCG)
c = n_distinct(STR)
d = n_distinct(LPR)
e = n_distinct(PEG)
f = n_distinct(UNS)
num_distinct = c(a,b,c,d,e,f)
plot = barplot(num_distinct, names = c("STG", "SCG", "STR", "LPR", "PEG", "UNS"), ylim=c(0,120), xlab="A
text(plot,num_distinct + 4,labels=as.character(num_distinct))
```



All Features

```
# stat summary of the data
summary(STG,SCG,STR,LPR,PEG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2407  0.3270  0.3711  0.4950  0.9900
```