HKZaliczeniowe1

Hanna Kranas

25 kwietnia 2017

Litery z polskimi znakami

```
#litery
lit_polskie <- testuj() # z polskimi
```

```
## POLSKIE LITERY! ąę!dlugosc obserwacji: 168319[1] "Macierz emisji gotowa"
## [1] "Hmm zrobiony"
## $States
## [1] "pierwszy" "drugi"
##
## $Symbols
##  [1] "a" "d" "m" " " "i" "c" "k" "e" "w" "z" "p" "n" "t" "u" "s" "y" "l"
## [18] "o" "j" "ę" "g" "r" "ó" "ł" "ż" "ś" "ą" "ź" "b" "ć" "h" "f" "ń" "x"
## [35] "v"
##
## $startProbs
## pierwszy    drugi
##      0.5      0.5
##
## $transProbs
##          to
## from       pierwszy drugi
##   pierwszy     0.75  0.25
##   drugi        0.25  0.75
##
## $emissionProbs
##          symbols
## states              a          d          m                    i
##   pierwszy 0.02856923 0.02857051 0.02857112 0.02857052 0.02857246
##   drugi    0.02857315 0.02857321 0.02857136 0.02857071 0.02857159
##          symbols
## states              c          k          e          w          z
##   pierwszy 0.02857190 0.02857067 0.02857108 0.02857125 0.02857179
##   drugi    0.02857156 0.02857112 0.02857100 0.02857184 0.02857093
##          symbols
## states              p          n          t          u          s
##   pierwszy 0.02857031 0.02857184 0.02857207 0.02857118 0.02857143
##   drugi    0.02857011 0.02857163 0.02857242 0.02857085 0.02857154
##          symbols
## states              y          l          o          j          ę
##   pierwszy 0.02857328 0.02857247 0.02857068 0.02857152 0.02857052
##   drugi    0.02857025 0.02857293 0.02857217 0.02857058 0.02857183
##          symbols
## states              g          r          ó          ł          ż
##   pierwszy 0.02857008 0.02857307 0.02857144 0.02857106 0.02857117
##   drugi    0.02857159 0.02857138 0.02857146 0.02857146 0.02857253
```

```
##           symbols
## states              ś         ą         ź         b         ć
##   pierwszy 0.02857147 0.02857246 0.02857095 0.02857115 0.02857193
##   drugi    0.02857103 0.02857106 0.02857170 0.02857133 0.02856987
##           symbols
## states              h         f         ń         x         v
##   pierwszy 0.02857146 0.02857304 0.02857215 0.02857169 0.02857103
##   drugi    0.02857029 0.02857091 0.02857076 0.02857194 0.02857192
```

```
save(lit_polskie, file = "litery-polskie.RData")
grupy <- lit_polskie$hmm$emissionProbs["pierwszy",]-lit_polskie$hmm$emissionProbs["drugi",]
#jak ujemne to bardziej drugi
print('To jest grupa pierwsza:')
```

```
## [1] "To jest grupa pierwsza:"
```

```
cat(names(grupy[grupy>=0]))
```

```
## m i c e z n l j ę ż ś ą ź ć h ń v
```

```
print('To jest grupa druga:')
```

```
## [1] "To jest grupa druga:"
```

```
cat(names(grupy<0]))
```

```
## a d   k w p t u s y o g r ó ł b f x
```

Litery bez polskich znaków

```
lit_bezpolskie <- testuj(polskie = FALSE) #bez polskich
```

```
## dlugosc obserwacji: 168319[1] "Macierz emisji gotowa"
## [1] "Hmm zrobiony"
## $States
## [1] "pierwszy" "drugi"
##
## $Symbols
##  [1] "a" "d" "m" " " " " "i" "c" "k" "e" "w" "z" "p" "n" "t" "u" "s" "y" "l"
## [18] "o" "j" "g" "r" "b" "h" "f" "x" "v"
##
## $startProbs
## pierwszy    drugi
##      0.5      0.5
##
## $transProbs
##           to
## from       pierwszy drugi
##   pierwszy     0.75  0.25
##   drugi        0.25  0.75
##
## $emissionProbs
##           symbols
## states              a         d         m                   i
##   pierwszy 0.03846036 0.03846163 0.0384634 0.03845919 0.03846097
```

```
##   drugi    0.03846014 0.03846120 0.0384629 0.03846138 0.03846122
##           symbols
## states              c          k          e          w          z
##   pierwszy 0.03846176 0.03846152 0.03846199 0.03846239 0.03845916
##   drugi    0.03846116 0.03846182 0.03846114 0.03846182 0.03846137
##           symbols
## states              p          n          t          u          s
##   pierwszy 0.03846322 0.03846172 0.03846205 0.03846163 0.03846119
##   drugi    0.03846304 0.03846096 0.03846110 0.03846301 0.03846162
##           symbols
## states              y          l          o          j          g
##   pierwszy 0.03846095 0.03846179 0.03845979 0.03846224 0.03846121
##   drugi    0.03846178 0.03846169 0.03846105 0.03846062 0.03846162
##           symbols
## states              r          b          h          f          x
##   pierwszy 0.03846342 0.03846108 0.03846233 0.03846056 0.03846194
##   drugi    0.03846249 0.03846151 0.03846310 0.03846013 0.03845994
##           symbols
## states              v
##   pierwszy 0.03846253
##   drugi    0.03846220
```

```
save(lit_bezpolskie, file = "litery-bezpolskie.RData")
grupy <- lit_bezpolskie$hmm$emissionProbs["pierwszy",]-lit_bezpolskie$hmm$emissionProbs["drugi",]
#jak ujemne to bardziej drugi
print('To jest grupa pierwsza:')
```

```
## [1] "To jest grupa pierwsza:"
```

```
cat(names(grupy[grupy>=0]))
```

```
## m i c e z n j h
```

```
print('To jest grupa druga:')
```

```
## [1] "To jest grupa druga:"
```

```
cat(names(grupy<0]))
```

```
## a d   k w p t u s y l o g r b f x v
```

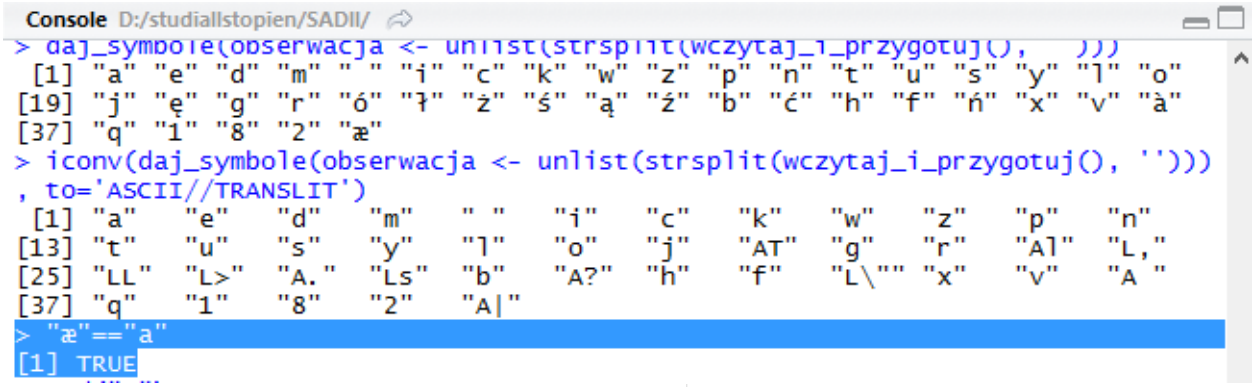
```
#mozna jeszcze plotowac posteriori odjete i patrzec gdzie wiecej samoglosek
```

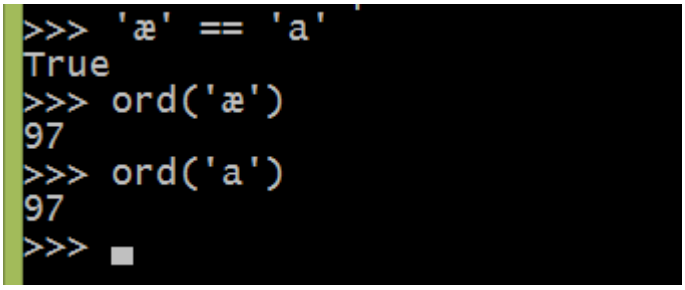
Słowa, bez polskich znaków

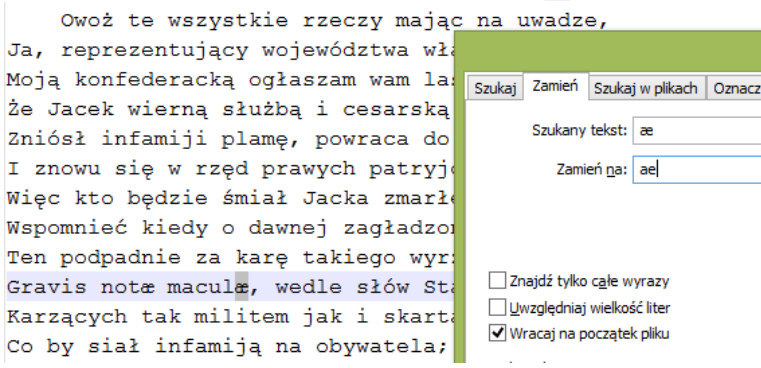
```
slowa_bezpolskie <- testuj(litery_nie_slowa = FALSE,polskie=FALSE)
save(slowa_bezpolskie, file = "slowa.RData")
head(slowa_bezpolskie$hmm$emissionProbs)
```

Przygotowanie danych

Ponieważ ani R ani Python nie radziły sobie z usuwaniem niektórych znaków, usunięte zostały w notepadzie.







Przygotowanie danych wykonane jednak zostało głównie w R:

```
wczytaj_i_przygotuj <- function(nazwa_pliku='pan-tadeusz.txt', polskie = TRUE){
  plik <- readLines(nazwa_pliku) #wczytuje plik
  Encoding(plik) <- 'UTF-8' #zmieniam kodowanie na odpowiednie
  plik <- tryCatch(plik[1:which(plik == '-----')-1],error = function(e) plik, finally= {}) #usuwam wszy
  plik <- gsub("[[:punct:]...-[:digit:]]", "", plik) #usuwam wszystkie znaki przestankowe
  plik <- tolower(plik[plik!=""]) #usuwam  puste
  if(polskie){
    cat('POLSKIE LITERY! ąę!')
  } else {
    plik <- iconv(plik,from="UTF-8",to="ASCII//TRANSLIT") #bez polskich
  }
  return(plik)
}
```