HKZaliczeniowe1

Hanna Kranas

25 kwietnia 2017

Litery z polskimi znakami

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

```
## POLSKIE LITERY! ąę!dlugosc obserwacji: 417048[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" "q"
##
## $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 0.02777773 0.02777891 0.02777867 0.02777581 0.02777765
##   drugi    0.02777797 0.02777765 0.02777758 0.02777872 0.02777764
##          symbols
## states              c          k          e          w          z
##   pierwszy 0.02777821 0.02777826 0.02777624 0.02777858 0.02777808
##   drugi    0.02777901 0.02777606 0.02777734 0.02777646 0.02777652
##          symbols
## states              p          n          t          u          s
##   pierwszy 0.02777799 0.02777865 0.02777794 0.02777844 0.02777765
##   drugi    0.02777586 0.02777986 0.02777759 0.02777845 0.02777671
##          symbols
## states              y          l          o          j          ę
##   pierwszy 0.02777642 0.02777862 0.02777617 0.02777808 0.02777759
##   drugi    0.02777772 0.02777792 0.02777888 0.02777786 0.02777533
##          symbols
## states              g          r          ó          ł          ż
##   pierwszy 0.02777738 0.02777775 0.02777940 0.02777806 0.02777693
##   drugi    0.02777663 0.02777901 0.02777922 0.02777888 0.02777805
```

```
##           symbols
## states              ś          ą          ź          b          ć
##   pierwszy 0.02777925 0.02777829 0.02777712 0.02777774 0.02777747
##   drugi    0.02777735 0.02777717 0.02777821 0.02777894 0.02777727
##           symbols
## states              h          f          ń          x          v
##   pierwszy 0.02777682 0.02777844 0.02777573 0.02777933 0.02777717
##   drugi    0.02777960 0.02777882 0.02777853 0.02777707 0.02777761
##           symbols
## states              q
##   pierwszy 0.02777742
##   drugi    0.02777651
```

```
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]))
```

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

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

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

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

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

Litery bez polskich znaków

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

```
## dlugosc obserwacji: 417048[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" "q"
##
## $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 0.03703683 0.03703754 0.03703663 0.03703752 0.03703663
##   drugi    0.03703615 0.03703571 0.03703778 0.03703516 0.03703705
##           symbols
## states              c          k          e          w          z
##   pierwszy 0.03703829 0.03703669 0.03703516 0.03703577 0.03703705
##   drugi    0.03703822 0.03703806 0.03703861 0.03703748 0.03703589
##           symbols
## states              p          n          t          u          s
##   pierwszy 0.03703791 0.03703706 0.03703700 0.03703623 0.03703771
##   drugi    0.03703696 0.03703786 0.03703579 0.03703777 0.03703640
##           symbols
## states              y          l          o          j          g
##   pierwszy 0.03703699 0.03703503 0.03703606 0.03703846 0.03703869
##   drugi    0.03703726 0.03703612 0.03703667 0.03703921 0.03703724
##           symbols
## states              r          b          h          f          x
##   pierwszy 0.03703666 0.03703702 0.03703934 0.03703929 0.03703610
##   drugi    0.03703537 0.03703730 0.03703652 0.03703642 0.03703802
##           symbols
## states              v          q
##   pierwszy 0.03703750 0.03703484
##   drugi    0.03703686 0.03703812
```

```
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]))
```

```
## a d   c w z p t u y o g r b h x
```

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

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

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

```
## m i k e n s l j f v q
```

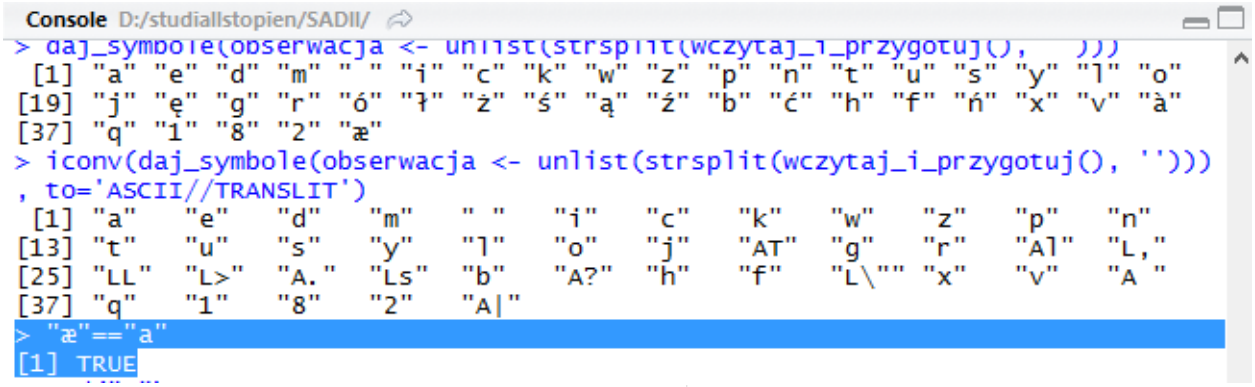
```
#mozna jeszcze plotowac posteriori odjete i patrzec gdzie wiecej samoglosek
```

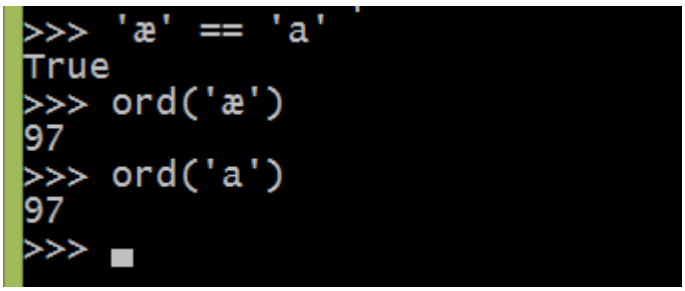
Słowa, bez polskich znaków

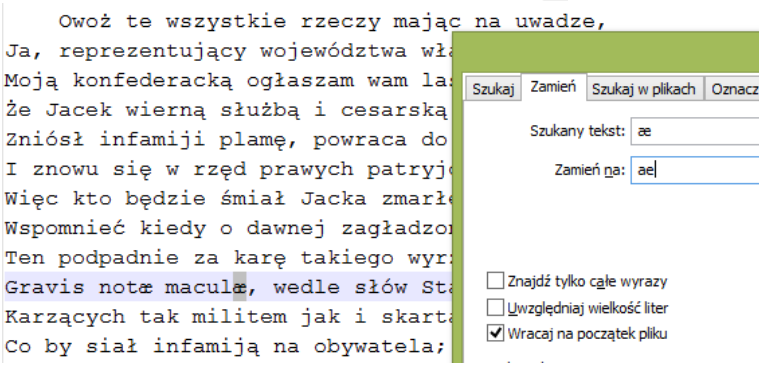
```
slowa_bezpolskie <- testuj(litery_nie_slowa = FALSE,polskie=FALSE)
save(slowa_bezpolskie, file = "slowa.RData")
head(slowa_bezpolskie$hmm$emissionProbs)
```

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)
}
```