

# HKZaliczeniowe1

Hanna Kranas

25 kwietnia 2017

## Litery z polskimi znakami

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

## POLSKIE LITERY! aę!dlugosc obserwacji: 41941[1] "Macierz emisji gotowa"
## [1] "Hmm zrobiony"

save(wyniki, file = "litery-polskie.RData")
grupy <- wyniki$hmm$emissionProbs["pierwszy",]-wyniki$hmm$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 ć h ń
print('To jest grupa druga:')

## [1] "To jest grupa druga:"
cat(names(grupy[grupy<0]))

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

## Litery bez polskich znaków

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

## dlugosc obserwacji: 41941[1] "Macierz emisji gotowa"
## [1] "Hmm zrobiony"

save(wyniki, file = "litery-bezpolskie.RData")
grupy <- wyniki$hmm$emissionProbs["pierwszy",]-wyniki$hmm$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 s j h
print('To jest grupa druga:')

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

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

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

```
#można jeszcze plotować posteriori odjęte i patrzeć gdzie więcej samogłosek  
#podobnie z sufiksami
```

## Słowa, bez polskich znaków

```
symbole <- slowa()
```

```
## POLSKIE LITERY! æ!dlugosc obserwacji: 68184
```

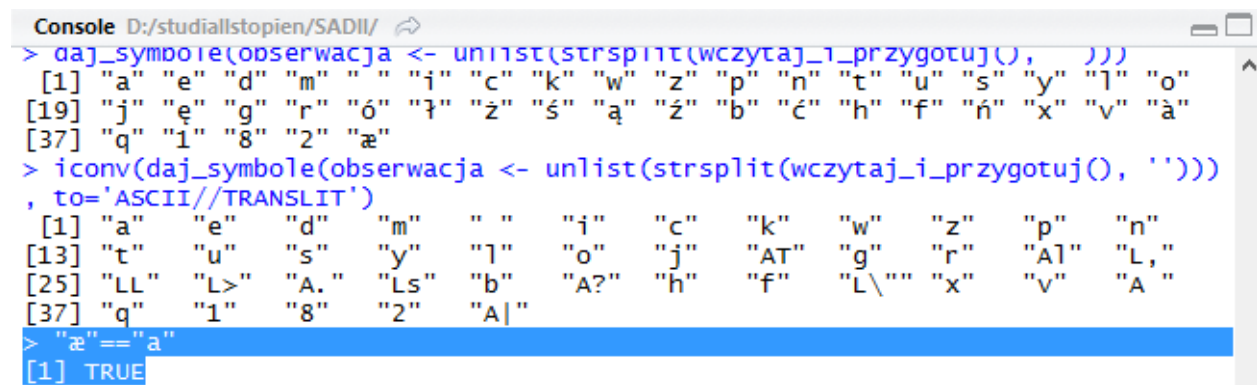
```
print(length(symbole))
```

```
## [1] 2444
```

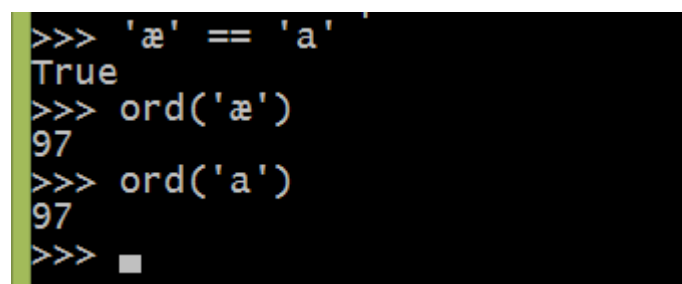
```
#wyniki <- testuj(litery_nie_slowa = FALSE,polskie=FALSE)  
#head(wyniki$hmm$emissionProbs)  
# Error in if (d < delta) { : missing value where TRUE/FALSE needed
```

## Pytania:

te dziwne znaki usunięte w notepadzie, bo zarówno R jak i python nie umiały ich rozróżnić - czy to ok?



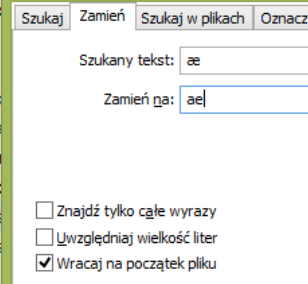
```
Console D:/studiallstopien/SADII/
> daj_symbole(obserwacja <- unlist(strsplit(wczytaj_i_przygotuj(), '')))
[1] "a" "e" "d" "m" " " "i" "c" "k" "w" "z" "p" "n" "t" "u" "s" "y" "l" "o"
[19] "j" "ę" "g" "r" "ó" "ł" "ż" "ś" "ą" "ź" "b" "ć" "h" "f" "ń" "x" "v" "à"
[37] "q" "1" "8" "2" "æ"
> iconv(daj_symbole(obserwacja <- unlist(strsplit(wczytaj_i_przygotuj(), ''))),
, to='ASCII//TRANSLIT')
[1] "a" "e" "d" "m" " " "i" "c" "k" "w" "z" "p" "n"
[13] "t" "u" "s" "y" "l" "o" "j" "AT" "g" "r" "Al" "L,"
[25] "LL" "L>" "A." "Ls" "b" "A?" "h" "f" "L\" "x" "v" "A"
[37] "q" "1" "8" "2" "A|"
> "æ"=="a"
[1] TRUE
```



```
>>> 'æ' == 'a'
True
>>> ord('æ')
97
>>> ord('a')
97
>>>
```

Owoż te wszystkie rzeczy mając na uwadze,

Ja, reprezentujący województwa wł  
Moją konfederacką ogłaszam wam la  
że Jacek wierną służbą i cesarską  
Zniósł infamiji płamę, powraca do  
I znowu się w rząd prawych patryjc  
Więc kto będzie śmiał Jacka zmarł  
Wspomnieć kiedy o dawnej zagładzo  
Ten podpadnie za karę takiego wyr  
Gravis notæ maculæ, wedle słów St  
Karzących tak militem jak i skarta  
Co by siał infamiją na obywatela;



Szukaj Zamień Szukaj w plikach Oznacz

Szukany tekst: æ

Zamień na: a

☐ Znajdź tylko całe wyrazy

☐ Uwzględnij wielkość liter

☒ Wróć na początek pliku

jak z tym lekkim zaburzeniem? czy jest na to jakiś sprytniejszy sposób? (i czy ten w ogóle jest słuszną drogą...)

```
macierz_emisji <- numeric(0)
for(i in 1:length(stany)){
  x <- rep(1/length(symbole),length(symbole))
  x[i] <- x[i] - 1/100000000 #dajemy małe zaburzenie
  x[i+1] <- x[i+1] + 1/100000000 #tu też, żeby się nadal sumowały w wierszu do 1
  macierz_emisji <- c(macierz_emisji,x)
}
```

czy jak nie dzieli w żadnym przypadku na samogłoski/spółgłoski to należy się martwić czy to była podpucha ze tak wyjdzie? (patrz: wyniki wyżej)

przy słowach błąd (patrz: wyżej)

Już coś szukałam na stackoverflow 1

Już coś szukałam na stackoverflow 2

zrobic dla 3-4tys znakow (czyli w testuj uciete [(1:4000)])

jesli chodzi o slowa, to mozliwe ze za male sa te prawdopodobienstwa ale jeszcze zobaczymy co wymyslimy jutro - moze jakies klastrowanie? moze jakies podmienianie?

albo zaraportować i napisać że sufiksy 2literowe lub zrobic klastrowanie sufiksów