Statistika za jezikoslovno istraživanje

Damir Ćavar
Sveučilište u Zadru
7. travnja 2010.

Podaci, izvori i kodiranje

Jezični podatci za statističku analizu

- Oblici:
  - tekst: knjige, novine, časopisi itd.
  - audio snimci
  - video snimci
  - slike, grafike itd.

Jezični podatci

- Klasični i prirodni oblici (jezičnih) podataka
  - tiskano = analogno i u obliku piktograma
  - snimljeno = analogno u obliku medija za snimanje

Jezični podatci

- Analogni podatci
  - Gubitak kvalitete (faktori npr. vrijeme, materijal)
  - Šum (može biti i povezan npr. s ograničenim tehnologijama, gubitkom kvalitete)

Jezični podatci

- Suvremeni oblici:
  - digitalni podatci na računalu
- Što to znači digitalni podatci?
  - tekst, audio, video, slike itd.
- Digitalni kod je diskretan

Jezični podatci

- Digitalno kodiranje:
  - Informacija → Byte → Bit → Napon ili struja u elektroničkim sklopovima
  - Kodiranje u digitalni oblik za obradu na računalu

Jezični podatci

- Digitalno kodiranje
  - gubi se informacija iz analognih izvora
    - tolerira se gubitak informacije o taktilnim i vizualnim osobinama, stanju, izboru boja i znakova itd. ali:
  - dobije se mogućnost jednostavnije, brže i bolje obrade i arhiviranja
    - velikih količina podataka u kratkom vremenu sa specifičnim alatima itd.

Jezični podatci

- Tipovi podataka
  - Vremensko neovisni (statični ili diskretni podatci)
    - tekst, slika itd.
  - Vremensko ovisni (dinamični, vrijeme se mora kodirati, pojedinačne informacije nebitne)
    - video, audio snimci, itd.
- Za svaki takav tip podataka postoji poseban način kodiranja.

Primjeri kodiranja

- Tekst (pisani):
  - niz znakova: {a, b, c, d, e, ...}
  - složeno u: riječi, rečenice, paragrafe, poglavlja itd.

Primjeri kodiranja

- Analogni u digitalni tekst
  - Brailleovo kodiranje
  - Morseovo kodiranje
  - Baudotovo kodiranje
  - ASCII kod u računalima
  - Unicode

Kodiranje i oznake

- Kod i oznake za tekstualni zadržaj
- Kod za tekstualni oblik, strukturu i semantiku dokumenta
- Meta-informacija
  - autor, naslov, datum izdanja, jezik, anotator itd.

R 2

R pomoć

```
help(KOMANDA)
```

- ?KOMANDA ili ?FUNKCIJA

```
?sqrt
```

```
help.start()
```

- Pretraga dokumentacije po ključnim riječima:

```
??deviation
```

Podaci

- Vektori i računanje s vektorima:
  - funkcija `c()`
    - uzima kao parametar niz podataka
    - vraća vector s tim podacima složenim u redoslijedu parametara

Operacije s vektorima

- Primjer:

```
x <- c( 0.3, 0.2, 0.24, 0.29 )

x = c( 0.3, 0.2, 0.24, 0.29 )

assign(“x”, c( 0.3, 0.2, 0.24, 0.29 ))

c( 0.3, 0.2, 0.24, 0.29 ) -> x
```

- `x` je varijabla koja pokazuje na memoriju koja sadrži podatke vektora u binarnom obliku.

Funkcije s vektorima

- Primjeri:

```
x
2 * x
x - mean(x)
length(x)
sum(x)
```

Prijevod u R

- Varijanca: $$s^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

  - zbrajanje svih vrijednosti u vektoru x:

    `sum(x)` $\sum_{i=1}^{N} x_i$

  - isto to zbrajanje i dodatno odbijanje aritmetičke sredine od svih vrijednosti u vektoru x:

    `sum(x-mean(x))` $\sum_{i=1}^{N}(x_i - \bar{x})$

Prijevod u R

- Varijanca: $$s^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

  - zbrajanje kvadratnih vrijednosti u vektoru `x` umanjene za aritmetičku sredinu:

    `sum((x-mean(x))^2)` $\sum_{i=1}^{N}(x_i - \bar{x})^2$

  - `N` je broj vrijednosti u vektoru `x`, što znači da je `N` dužina vektora `x`:

    `length(x)`

Prijevod u R

- Varijanca: $$s^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- sve u jednoj funkciji:

```
sum((x-mean(x))^2)/(length(x)-1)
```

- ili jednostavno:

```
var(x)
```

Prijevod u R

- Standardna devijacija: $s_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- ????

Prijevod u R

- Standardna devijacija

$$s_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- sve u jednoj funkciji:

```
sqrt(sum((x-mean(x))^2)/(length(x)-1))
```

- ili:

```
sqrt(var(x))
```

- ili jednostavno:

```
sd(x)
```

R osnove

- Procesiranje vektora:

```
x > 5

y <- x < 5

y <- 1:10

y <- seq(1, 10, by=2)

y <- rep(x, times=3)

y <- rep(x, each=3)
```

R osnove

- Konstrukcija podataka:
  - Ako nema specifičnih podataka (u statističkom smislu): NA znači “*not available*”

```
x <- c(NA, 3, 2, 4, 3, NA, 3, 2)

is.na(x)
```

R osnove

- Nešto nije broj: “Not a Number: NaN”

```
0/0

log(-2)

sqrt(-1)
```

R osnove

- Selekcija dijela vektora:

```
x[2]

x[2:5]

x[-1]

x[-(6:8)]

x[x>3]

x[!(is.na(x)) & x>2]
```

R osnove

- Selekcija dijela vektora:

```
x[!(is.na(x)) & x>2]

(x+1)[!(is.na(x)) & x>2]
```

R osnove

- Imenovanje razreda:

```
boje <- c(3, 4, 2, 5)

names(boje) <- c("crveno", "sivo", "bijelo", "plavo")

izbor <- boje[c("crveno", "bijelo")]
```

R osnove

- Manipulacija vektora:

```
x <- c(NA, 3, 2, 4, 3, NA, 3, 2)

x[is.na(x)] <- 0

x
```

R osnove

- Manipulacije vektora:

```
x <- c(-1, 2, 4, -3, 4)
```

```
x[x<0] <- -x[x<0]
```

ili

```
x <- abs(x)
```

Domaći

- Prevedite sljedeću formulu u R
- pretpostavljamo da imamo vektor `x` s vjerojatnostima pojavnica npr. riječi u korpusu, i p(x) su upravo vjerojatnosti:

```
x <- c(0.001, 0.0018, 0.000032, 0.002)
```

$$-\sum_{i=1}^{n} p(x_i) log_b p(x_i)$$