The R Project for Statistical Computing Overview

Petr Kosina
soso@mail.muni.cz

PA164 Strojové učení a přirozený jazyk

Obsah

Úvod

- Co je R a kde ho získáme
  - `http://www.r-project.org/`
  - Free Software
  - jazyk a prostředí pro statistické výpočty a grafiku
- Co umožňuje
  - jednoduchá práce se soubory
  - množství operací s poli, vektory a maticemi
  - analýza dat
  - grafické výstupy
  - podmínky, cykly, uživatelem definované funkce

- Další výhody
  - podporuje různé platformy
  - možnost zrychlení výpočtů integrací kódu v C, C++ nebo Fortranu
  - Java
  - velké množství balíčků od uživatelů
  - tvorba vlastních balíčků

Začínáme - krátký a nudný manuál

- Ovládání
  - prompt značen >
  - příkazy odděleny enter nebo ;
  - napověda pomocí ?příkaz nebo help.search("hledany_retezec")
  - konec q()
- Proměnné
  - nedeklarované
  - přiřazení nazev_promenne <− hodnota

Typy

- Vektor - příklady vytváření
  - výčet
    - x <- 23
    - x <- c(2,6,4,18)
  - rozsah
    - x <- 1:10
- Matice
  - x <- matrix(1:12,4)
  - funkce cbind() nebo rbind()
  - ...

Typy - další

- Arrays
- Lists
- Data.frame
  - složitější struktury
  - tvořeny z ostatních
  - viz manuál či příklady

Užitečné příkazy

- Vlastní funkce
  - mojefunkce $<-$ function(parametr_1, parametr_2, ...){prikazy}
- Grafika
  - plot()
    - vykresluje
    - popisy os, název, barva vykreslení...
  - par()
    - úprava různých vlastností zobrazení
  - legend()
    - přidá vysvětlivky

tm

- Autor: Ingo Feinerer
- Načtení balíčku
  - library(tm)
  - vyžaduje mimo jiné balíček *slam*
- *tm* umožňuje efektivní práci s textovými daty
  - vytvoření korpusu
  - různé metody předzpracování
  - tvorba term-document matice
  - prohlídku či manipulaci s maticí
  - ...

Struktura Corpus

- Dva typy Corpus
  - VCorpus - volatile corpus
  - PCorpus - permanent corpus
- Parametry
  - Source
  - readerControl

Corpus - parametry

- Vstup různých typů pomocí parametru Source
  - DirSource - adresář
  - VectorSource - vektor znaků
  - DataframeSource - struktury (např. CSV)
  - getSources() - zobrazí dostupné možnosti vstupů
- parametr readerControl - seznam (list) komponent
  - reader - konstrukce textových dokumentů ze vstupu
    - readPlain()
    - readDOC()
    - readPDF()
    - getReaders() - zobrazí dostupné možnosti reader
  - language - nastavení jazyka

Další funkce

- Zápis vytvořeného pomocí writeCorpus()
- Výpis korpusu
  - print() - jen základní informace
  - summary() - více detailů o meta-datech
  - inspect() - celý obsah

Transformace

- Funkce tm_map()
  - PlainTextDocument - plain text
  - stripWhitespace - odstranění bílých znaků
  - tolower - převod na malá písmena
  - odstranění stop slov př. reuters $<-$ tm_map(reuters, removeWords, stopwords("english"))
  - stemDocument - převod na kořen slova

Term-document matice

- Objekt vhodný pro dolování
  - TermDocumentMatrix - termy řádky, dokumenty sloupce
  - DocumentTermMatrix - opak
- Operace
  - findFreqTerms() - časté vzory
  - findAssoc() - nalezení asociací ke zvolenému termu (se zadanou korelací)

Závěr

- Příklad na rozhodovací stromy
- Více zájemci naleznou v dizertaci Ingo Feinerera