Студијски програм: Мастер академске студије информатике

Назив предмета: ПРИПРЕМА ПОДАТАКА

Статус предмета: Изборан на модулу Наука о подацима

Број ЕСПБ: 6

Услов: Уписан одговарајући семестар

Циљ предмета

Оспособљавање студената за обраду изворних, сирових података и припрему поузданих, конзистентних података на бази експлоративне анализе и модерних софтверских окружења за Data Science.

Исход предмета

Савладано градиво оспособиће студента:

- за ефикасан увоз и манипулацију подацима из различитих извора
- за квалитетну експлоративну анализу података
- да овлада вештинама трансформације података од технички чистих до конзистентних података
- да влада модерним софтверским алатима за припрему података.

Садржај предмета

Теоријска настава

Увод. Типови података. Формати. Увоз података у R. Софтверски алати за трансформацију података у R окружењу (dplyr, tidyr, tibble, stringr, magrittr, purr, modelr, lubridate, RODBC…). Wrangling – припрема података за анализу. Увоз локалних података. Увоз CSV, XLSX, XML датотека. XLConnect – манипулација Excel датотекама. Увоз из база података. SQL упити из R-а. Увоз података из статистичких софтверских пакета. Увоз података са web-а. Json. HTML. Експлоративна анализа података. Структуре података. Неуредни и неконзистентни подаци. Чишћење података. Дупликати. Трансформација података. Трансформациона правила. Манипулација редова и колона података. Рашчлањење вектора и датотека. Провера типова. Додавање нових варијабли. Скалирање. Нормализација. Кодирање категоријских података. Пондерисање варијабли. Корекције и импутација. Агрегирање података. Израчунавање збирних статистичких показатеља. Формирање тиблова. Тиблови и структура data.frame. Добијање међузбирова података. Рад са групама података. Радни токови. Скрипте. Од технички чистих података до конзистентних података. Корелациона матрица. Варијансе. Шаблони. Модели. Недостајући подаци (Missing values). Нетипичне тачке (Outliers). Проклетство димензионалности. Редукција података. Редукција димензија. PCA – Principal component Analysis. Релациони подаци. Рад са знаковним низовима. Алати за рад са категоријским варијаблама (пакет: forcats). Рад са датумским подацима и временом. Програмирање.

Практична настава

Примена софтверских алата за припрему података у R окружењу (base пакет, ggplot2, tidyr, dplyr, ggvis, rattle, dplyr пакет, tidyr, tibble, stringr, magrittr, purr, modelr, lubridate, shiny…). Рад на вежбама подразумева примену стеченог знања на решавање конкретних задатака у домену припреме података.

Литература

1. Wickham, Hadley, and Garrett Grolemund, R za statističku obradu podataka, Mikro knjiga, 2017.
2. Boehmke, Bradley C. Data Wrangling with R. Springer, 2016.
3. Buttrey, Samuel E., and Lyn R. Whitaker. A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R. John Wiley & Sons, 2017.

Број часова активне наставе	**Теоријска настава:** 2	**Практична настава:** 1 + 1

Методе извођења наставе

Проблемски-оријентисана настава, практична настава и вежбе уз софтверску подршку, самостални рад студената и консултације.

Оцена знања (максимални број поена 100)

Предиспитне обавезе	**70** поена	**Завршни испит**	**30** поена
активност у току предавања		писмени испит	20
практична настава	20	усмени испит	10
колоквијум-и	20		
семинар-и	30		