OSNOVE PROGRAMSKOG JEZIKA R

Projektni zadatak

22.12.2016

Opis zadatka

U sklopu vještine "Osnove programskog jezika R" predviđena je izrada projektnog zadatka. Zadatak se sastoji u sljedećem:

Studenti moraju provesti eksploratornu analizu (minimalno) dva podatkovna skupa iz zadane kolekcije podatkovnih skupova. Rezultat analize mora biti PDF izvještaj koji će:

- **opisati proces učitavanja i prilagodbe podataka**
- **prikazati niz interesantnih vizualizacija dobivenih iz odabranog podatkovnog skupa**
- **iznijeti zaključke o podatkovnom skupu utemeljene na rezultatima analize**

Predviđen je rad u korisničkom sučelju *RStudio*. Projektni zadatak radi se u parovima, iako je moguće da student po želji radi i sam. Na predavanju 4.1.2016. kod popisivanja morati ćete navesti s kim u paru radite projekt.

ROK ZA PREDAJU PROJEKTNOG ZADATKA JE 1.2.2017.

Projektni zadatak predajete preko *Moodle* sustava, slično predaji laboratorijskih vježbi i domaćih zadaća. Za izradu projektnog zadatka poslužite se saznanjima iz do sada obrađenih lekcija te danim podsjetnicima. Analiza ne mora sadržavati prediktivne modele budući da to još nije obrađeno, no po želji studenta koji imaju dovoljno znanja za provedbu ovakvog tipa analize mogu također isto uključiti u izvještaj.

Ponuđeni podatkovni skupovi su sljedeći:

GoT_battles	bitke iz *Game of Thrones* serijala knjiga
GoT_character-deaths	likovi (i njihov životni status) iz *Game of Thrones* serijala
HighestMountains	najviše planine svijeta
HR_StanPremaSpoluStarosti	popis stanovništva gradova RH prema spolu i starosti
HR_ZG_RI_OS_ST_Stan	popis stanovništva naselja većih gradova RH prema spolu i starosti
HR_zaposleni_neto_placa	neto plaće u pojedinom poslovnom sektoru u RH
HR_ZupanijeTurizam	turističke informacije
IGN_game_reviews	informacije o recenzijama računalnih igara portala IGN
IMDB_movie_dataset	informacije o filmovima sa portala IMDB
SettlersOfCatan	statistike o partijama društene igre *Settlers of Catan*
Simpsons_episodes	podaci o emitiranjima i gledanosti tv serije *The Simpsons*
speed-dating-experiment	podaci o rezultatima *speed dating* događaja
SuperMarioMakerDatabase	podaci o nivoima za igru *Super Mario Maker* na konzoli *Nintendo Wii U*

Glavni izvori podatakovnih skupova su **Kaggle** (www.kaggle.com) i **Državni zavod za statistiku** (www.dsz.hr).

U nastavku slijedi kratki naputak kako organizirati proces analize podataka kojeg se poželjno pridržavati kako bi se zadatak što učinkovitije izvršio.

Organizacija procesa analize podataka

Budući da je analiza podataka često zahtjevan i složen proces, preporučuje se unaprijed pripremiti određenu organizacijsku infrastrukturu kako bi se lakše upravljalo procesom, omogućilo lako uočavanje i ispravljanje grešaka te stvorila podrška za jednostavno ponavljanje i prilagodba već gotovih koraka procesa.

Postoji puno preporuka o tome kako najbolje organizirati ovakvu analizu, a također i adekvatna softverska podrška u obliku dodatnih paketa, sučelja i praćenih tehnologija. Što je analiza složenijeg tipa, to su i potrebe za višom razinom organizacije utoliko i važnije, a pogotovo ako se analiza provodi u višekorisničkom okruženju gdje je potrebno ne samo pratiti svoj vlastiti posao, nego se koordinirati sa ostalim članovima projektnog tima.

Za potrebe projektnog zadatka preporučuje se jednostavnija inačica organizacije postupka analize koja je prilagođena manje zahtjevnim analitičkim okruženjima te ne zahtjeva nikakvu dodatnu softversku podršku pored stvaranja određene hijerarhije podatkovnih mapa te držanje određenih konvencija tijekom procesa analize. Postupak je sljedeći:

1. Za potrebe analize određenog podatkovnog skupa potrebno je otvoriti zasebnu mapu koja će se koristiti strogo za tu analizu. Ovo se može izvesti kroz razvojno sučelje stvaranjem novog "projekta", ili jednostavno stvaranjem odgovarajuće mape na disku i pozicioniranjem u istu prije početka analize.
2. U osnovnu mapu projekta preporučuje se stvoriti podmape sljedećih naziva:

a) **R** - u ovu mapu stavljaju se vlastite R skripte koje će se koristiti u projektu; najčešće se radi o različitim funkcijama (npr. za čišćenje i obradu podataka) za koje se očekuje da će se pozivati više puta i kojima nije mjesto u izvještaju. Datoteke u ovoj mapi nazivamo *ime_datoteke.R* a učitavamo ih uz pomoć funkcije `source`.

b) **data** - u ovoj mapi nalaze se podaci - bilo da se radi o ulaznim podacima, ili podacima koji su prošli određene korake čišćenja i prilagodbe. Podatke je preporučeno spremati u obliku *CSV* datoteke, sa jasno danim nazivima (ako imamo više iteracija međurezultata koristimo se brojevima koji jasno naznačuju slijed prilagodbe - npr. `podaci_01.csv`, `podaci_02.csv`...).

c) **figures** - u ovu mapu stavljamo sve interesantne vizualizacije koje smo stvorili tijekom prediktivne analize. Preporučuje se pohranjivati ih u *PDF* ili *PNG* obliku. Vizualizacija mora imati dovoljno podataka kako bi se kasnije mogla interpretirati (i rekonstruirati), a za vizualizacije koje ćemo kasnije ugrađivati u izvještaj preporuka je sačuvati i programski kod koji ih je stvorio. Ukoliko želimo, u ovoj mapi možemo napraviti i podmape **expl** i **report** kako bi razdvojili vizualizacije kojima samo istražujemo podatke od onih koje ćemo uključiti u izvještaj

d) **Rmd** - u ovoj mapi radimo R Markdown izvještaje. Ukoliko se radi o složenijem procesu analize podataka, preporučeno je slijedno numerirati izvještaje (`01-učitavanje.Rmd`, `02-transformacija.Rmd` itd.) Preporučuje se pažljivo dokumentirati cijeli postupak analize kako bi se u bilo kojem trenutku on mogao izvesti ponovo (nad istim podacima sa istim rezultatima za provjeru valjanosti, ili kao predložak za nove podatke).

3. Pored gore navedenih mapa i pripadnih datoteka, u osnovnu mapu projekta preporučuje se stvoriti dvije tekstualne datoteke:

a) `log.txt` - nakon svakog rada na podacima na vrh ove datoteke ukratko upišemo što se napravilo i koji su okvirni rezultati

b) `TODO.txt` - kratki opis planiranih zadataka za iduću "sesiju"

Ovakva struktura datoteka i mapa omogućiti će nam lakše provođenje i praćenje procesa analize. Ukoliko se ista pokaže nedovoljnom, preporučuje se istražiti resurse na Webu koji će pružiti dodatne informacije o nešto složenijim modelima organizacije ili konkretnoj softverskoj podršci koja izvjesne elemente organizacije procesa analize automatizira te omogućuje visokorazinsko upravljanje procesom.