Istraživanje podataka - teme za seminarske radove (š.g. 2020/2021.)

Bitne informacije:

- Student seminarski rad radi samostalno ili u paru. Jednu temu čini jedan skup podataka za obradu i metoda istraživanja podataka kojom se skup obrađuje (klasifikacija ili klasterovanje).

- Jednu temu može da radi samo jedan student (ili tim ako je tema za dva studenta).
- Student može da
  - izabere jednu od predloženih tema koja nije zauzeta
  - predloži skup podataka sa dole navedenih izvora koji je bi želeo da analizira i metodu istraživanja podataka koju bi primenio (klasifikacija ili klasterovanje). Pri prijavi svog skupa podataka obavezno navesti izvor (adresu skupa).

  Mogući linkovi za izvor materijala:

  1. SPMF open-source data mining library http://www.philippe-fournier-viger.com/spmf/
  2. https://www.kdnuggets.com/datasets/index.html
  3. https://www.kaggle.com/datasets
  4. http://archive.ics.uci.edu/ml
  5. http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html
  6. https://crawdad.org/all-byname.html
  7. http://socialcomputing.asu.edu/pages/datasets
  8. http://cnets.indiana.edu/resources/data-repository/
  9. https://www.visualdata.io/
  10. http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/
  11. https://datarepository.wolframcloud.com/category/Machine-Learning/

  Napomena: Uzeti samo teme čiji je materijal takav da ima više od 5000 slogova sa bar 15 atributa.

- Opšta struktura rada
  - Obavezno izvršiti preprocesiranje podataka.
  - Podatke obraditi traženom metodom (klasifikacija, klasterovanje) koristeći više algoritama.
  - Izvršiti analizu dobijenih rezultata; uporediti rešenja dobijena za različite algoritme.

  - Seminarski rad koji se predaje treba da bude zapakovan u jednu datoteku koja treba da sadrži
    * tekstualni deo (tekst zadatka, opis postupka, dobijena rešenja, objašnjenja, ...). Tekstualni deo treba da bude pisan u LaTeX-u ili MS-Word-u (isključivo .doc format). matu.
    * podatke (početnu verziju kao i verziju dobijenu preprocesiranjem)
    * konstruisani model (u dogovarjućem obliku, zavisi od korišćenog softvera).

    Materijal koji se šalje mora da sadrži sve što je potrebno za ponavljanje kompletnog postupka u lokalnom okruženju.
    Za rešavanje problema mogu se koristiti SPSS modeler, biblioteke programskog jezika python, kao i programi i skripte koje je napisao sam student. Ako se koristi neki deo koda koji je skinut sa mreže obavezno navesti njegov izvor.

- Odbrana seminarskog rada
  Student izlaže seminarski rad pred profesorom i asistentom u vidu prezentacije.

1 Predložene teme

1.1 Klasifikacija

1. Da li su pečurke jestive ili otrovne
   https://archive.ics.uci.edu/ml/datasets/Mushroom
2. Izvršiti klasifikaciju nad skupom podataka o načinu prevoza.
   https://www.kaggle.com/fschwartzer/tmd-dataset-5-seconds-sliding-window
3. Primenom klasifikacije predvideti da li će padati kiša korišćenjem podataka o prethodnom danu/danima.
   https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
4. Primenom klasifikacije predvideti da li će osoba dati otkaz.
   https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study#general_data.csv
5. Klasifikacija zvezda, galaksija i kvazara
   https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey
6. Klasifikacija drveća http://kdd.ics.uci.edu/databases/covertype/covertype.data.html (2 studenta)
7. Klasifikacija podataka o člancima
   http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity
8. Klasifikacija komentara o ženskoj odeći.
   https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews
9. Klasifikacija anketa o interesovanjima mladih ljudi.
   https://www.kaggle.com/miroslavsabo/young-people-survey
10. Klasifikacija Stack Overflow anketa.
    https://www.kaggle.com/stackoverflow/so-survey-2017 (2 studenta)
11. Klasifikacija anketa o programerskim veštinama.
    https://www.kaggle.com/hackerrank/developer-survey-2018 (2 studenta)

12. Klasifikacija podataka o istraživačima (ciljni atribute: role).
    https://www.kaggle.com/bmkramer/101-innovations-research-tools-survey

1.2 Klasterovanje

1. Klasterovati ispitanike prema odgovorima iz skupa *Kaggle Machine Learning & Data Science Survey 2017*
   https://www.kaggle.com/kaggle/kaggle-survey-2017. (2 studenta)

2. Izvršiti klasterovanje nad recenzijama zaposlenih.
   https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews

3. Izvršiti klasterovanje nad podacima o programerima.
   https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

4. Izvršiti klasterovanje nad anketama o mašinskom učenju i nauci o podacima.
   https://www.kaggle.com/kaggle/kaggle-survey-2018

5. Izvršiti klasterovanje nad komentarima o ženskoj odeći.
   https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

6. Izvršiti klasterovanje nad anketama o interesovanjima mladih ljudi.
   https://www.kaggle.com/miroslavsabo/young-people-survey

7. Izvršiti klasterovanje nad Stack Overflow anketa.
   https://www.kaggle.com/stackoverflow/so-survey-2017 (2 studenta)

8. Izvršiti klasterovanje nad anketama o programerskim veštinama.
   https://www.kaggle.com/hackerrank/developer-survey-2018 (2 studenta)

Napomena: Kolone koje sadrže tekstove (npr. komentari o aplikaciji) je potrebno zasebno obraditi - napraviti i term matricu.