

# Mašinsko učenje

## Prvi domaći zadatak

Twitter sentiment analysis

### Opis zadatka

Treba napraviti model koji je prepoznaje ton (negativan ili pozitivan) nekog tweet-a. Negativan ton je obeležen klasom 0, a pozitivan klasom 1. Model trenirati na **train.csv** podacima, a testirati ih na **test.csv** podacima. Oba skupa podataka nalaze se na materijalima (Domaći1/data). Pošto se radi o realnim tekstualni podacima, podatke je neophodno procesirati i pripremiti za neki algoritam mašinskog učenja.

### Način bodovanja

Napomena: Neophodno je uraditi pretprocesiranje podataka primenom tehnika obrade prirodnih jezika, rađenih na času. Mogući algoritmi mašinskog učenja:

- Primena nekog algoritma klasifikacije rađenog na času (do 20 poena)
- Primena unapređenih algoritma klasifikacije rađenih na času (do 25 poena):
  - Obavezno: Izbor modela na osnovu validacionog set-a i analiza tačnosti na test setu
  - Algoritmi:
    - Logistic Regression sa L2 regularizacijom ili
    - Weighted k-NN
    - Naive Bayes

Modeli koji imaju tačnost ispod 50% neće biti razmatrani. Tačnost se analizira po formuli rađenoj na časovima vežbi (zbir true positive i true negative kroz ukupan broj test podataka).

**Domaći rade najviše 2 studenta!**

### Predaja domaćeg

Domaći se šalje arhiviran na [mcerovic@raf.rs](mailto:mcerovic@raf.rs) uz napomenu o imenu i prezimenu studenta, ili uz broj indeksa. Rok za predaju je ponoć petak – subota, 13 – 14 april. Odbrana domaćeg zadatka je obavezna.

Odbrana prvog domaćeg zadatka će se odvijati u kolokvijumskoj nedelji. Ako ne možete tada da prisustvujete odbrani, javite se putem mail-a, kako bismo se dogovorili o prethodnoj ili naknadnoj odbrani.