Istraživanje podataka 1 - zadaci

1. Python
   - Učitati skup podataka *20newsgroups*.
   - Koliko ima različitih klasa?
   - Ako je potrebno, podeliti skup podataka na deo za obučavanje i deo za testiranje modela.
   - Da li je skup balansiran? Ako je potrebno, primeniti slučajno uzorkovanje skupa.
   - Pretvoriti tekstualne podatke u matricu terma u dokumentima (term frequency, TF matrica).
   - Primeniti Naivni Bajesov algoritam za klasifikaciju.
   - Proceniti kvalitet dobijenog modela koristeći matricu konfuzije i tačnost (accuracy) modela. Da li je tačnost adekvatna metrika za ovaj skup podataka?
   - Pretvoriti TF matricu u TF-IDF i ponoviti ceo postupak. Uporediti novi model sa starim.

2. Python
   - Dato je šest sintetičkih skupova podataka.
   - Na prva 2 skupa treba pronaći 2 klastera, a na preostala 4 skupa 3 klastera.
   - Na svakom od skupova podataka izvršiti hijerarhijsko klasterovanje sa 3 različite metode spajanja najbližih klastera – single, average i complete. Ako je potrebno, pre primene algoritama izvršiti odgovarajuće pretprocesiranje.
   - Nacrtati mrežu scatter plotova za svaki skup i svaki model. Instance obojiti na osnovu toga kom klasteru pripadaju. Kao naslov svakog od grafika postaviti metodu spajanja koja je korišćena, silueta koeficijent i vreme izvršavanja.
   - Uporediti rezultate različitih metoda spajanja na različitim skupovima podataka. Da li su rezultati, uključujući i vreme izvršavanja, očekivani? Da li je silueta koeficijent adekvatna metrika za sve skupove? Ako ne, zbog čega?
   - Odabrati jedan model i jedan skup podataka gde rezultati deluju dobro. Nacrtati dendrogram za taj model.

3. SPSS
   - Učitati skup podataka *marketbasket.csv*. Da li ima nedostajućih vrednosti?
   - Koristeći algoritam Apriori pronaći pravila pridruživanja o artiklima koji se zajedno kupuju u prodavnici. Postaviti uslove da je najmanja podrška za telo pravila 20%, a najmanja pouzdanost 80%. Omogućiti da u svakom telu pravila bude najviše 4 stavke.
   - Koliko pravila ima ukupno? Koliko zanimljivih pravila ima po Lift meri?
   - Koliko ima pravila u kojima se u telu zajedno javljaju pasta za zube (eng. toothpaste) i beli hleb (eng. white bread). Izdvojiti ih u poseban model.