Istraživanje podataka 1 - tekstovi zadataka, vežbe 8

1. Izvršiti klasifikaciju nad skupom o belonima (*balloons.csv*) primenom algoritma naivni Bajes za kategoričke atribute. Za test skup ispisati:

   - ukupnu preciznost za test skup
   - izveštaj o klasifikaciji (za svaku klasu izdvojiti: preciznost, odziv i f1) za test skupu

   Zadatak rešiti korišćenjem programskog jezika Python.

2. U programskom jeziku Python napisati kod kojim se

   - pravi term-matricu sa brojem pojavljivanja reči za trening skup:

Id teksta	**reči u dokumentu**	**klasa**
1	Chinese Beijing Chinese	yes
2	Chinese Chinese Shanghai	yes
3	Chinese Macao	yes
4	Tokyo Japan Chinese	no

   - pomoću algoritma multinomijalni naivni Bajes i trening skupa pravi model za klasifikaciju
   - klasifikuje test instanca X=*Chinese Chinese Chinese Tokyo Japan*.

3. Dat je skup sa podacima iz novinskih članaka - ebart. Članci su podeljeni prema klasi kojoj pripadaju u direktorijume: Ekonomija, HronikaKriminal, KulturaZabava, Politika i Sport. Svaki članak je obrađen: uklonjene su stop reči i svaka reč je zamenjena svojim korenom, a zatim je izvršeno prebrojavanje reči. Rezultat obrade svakog članka je sačuvan u zasebnoj datoteci. U dobijenoj datoteci koja odgovara jednom članku, u jednom redu su podaci o jednom korenu reči - koren reči i broj pojavljivanja tog korena u tom članku.

   U programskom jeziku Python napisati kod kojim se

   - pravi term-matrica sa brojem pojavljivanja reči za članke iz teksta
   - prave modeli za klasifikaciju primenom algoritama
     - multinomijalni naivni Bajes
     - K najbližih suseda i unakrsne validacije za odabir vrednosti parametara
     - drveta odlučivanja i unakrsne validacije za odabir vrednosti parametara
   - na standardni izlaz se za svaki napravljeni model ispisuje:
     - korišćene vrednosti parametara algoritma pri pravljenju modela, ukoliko je model napravljen korišćenjem unakrsne validacije
     - ukupna preciznost za trening i test skup

– matrica konfuzije za trening i test skup
– izveštaj o klasifikaciji (za svaku klasu izdvojiti: preciznost, odziv i f1) za trening i test skup