1 Organizacija invertovano-indeksnih datoteka

1. Za date indeksne termine formirati

   (a) binarno pretraživačko stablo;

   (b) balansirano B-stablo (kapaciteta 3).

   Indeksni termini su: MREŽA, RAČUNAR, TASTATURA, TABLET, MONITOR, KAMERA, RUTER, MIŠ, SLUŠALICE, EKRAN

2 Algoritmi za sravnjivanje teksta

1. Za tekst

   SVAKA SVRAKA SKAKALA NA DVA KRAKA

   i obrazac SVRAKE koji se traži u tekstu, pokazati rad

   (a) KMP algoritma;

   (b) BM algoritma.

3 Automatsko indeksiranje

1. U kolekciji od $N = 143,800$ dokumenata, date su frekvencije pojavljivanja sledećih reči (ukupan broj pojavljivanja reči u svim dokumentima):

termin	df_t
ljiljan	34,230
neven	14,430
jasmin	16,890
orhideja	23,150

   Dat je i apsolutan broj pojavljivanja reči unutar dokumenata D_1 i D_2 iz kolekcije:

termin	D_1	D_2
ljiljan	45	16
neven	0	34
jasmin	25	47
orhideja	23	15

   Broj reči po dokumentima: $|D_1| = 780, |D_2| = 920$

   (a) Izračunati indeks značaja ovih termina za indeksiranje (idf). Koji je najznačajniji termin za indeksiranje prema ovoj meri?

   (b) Izračunati združeni indeks ($tf-idf$) za ova tri dokumenta iz kolekcije. Koja reči su najznačajnije za indeksiranje svakog od tih dokumenata?

4 Hash tabele

1. Napraviti heš tabelu za termine: TASTATURA, RAČUNAR, MONITOR, FASCIKLA, KATALOG, ŠTAMPAČ i heš funkciju $f(s) =$ broj suglasnika u s $\%$ 5.

5 Automatsko indeksiranje i diskriminativna vrednost termina

1. Neka je data kolekcija dokumenata sa pridruženim terminima:

   D_1 T_1, T_3
   D_2 T_2, T_4, T_5
   D_3 T_1, T_2, T_4
   D_4 T_2, T_3, T_5

   (a) Izračunati gustinu prostora Q_{pre_6} ove kolekcije dokumenata koristeći Dajsov indeks sličnosti između dokumenata.

   (b) Neka se termin T_6 dodeli dokumentima D_1 i D_3. Odrediti gustinu prostora Q_{posle_6} posle ove dodele i diskriminatornu vrednost $dv_6 = Q_{pre_6} - Q_{posle_6}$ termina T_6.