SLO

Naloga 1 (vrednost naloge 30 točk)

Matrika sosednosti na sliki 1 opisuje povezave omrežja, po katerem se premika spletni pajek. V nekem trenutku je v frontierju stanje, kot ga prikazuje slika 2.

	1	2	3	4	5	6	7	8	9	10	11	
1		1	1					1				3
2							1				1	2
3					1				1	1	1	4
4												
5	1			1					1			2
6										1		1
7						1					1	2
8			1	1								2
9				1								1
10							1					1
11						1						1

Slika 1: matrika sosednosti

4	7	2	8						

Slika 2: stanje frontierja

a) Na kateri strani je bil spletni pajek pred tremi koraki, če vemo, da se premika v širino in da je začel s strani z največ izhodnimi povezavami?
b) Katero stran bo pajek iz točke a obiskal zadnjo?
c) Katera od spodnjih trajektorij ustreza pajku, ki je začel na isti strani kot pajek iz točke a, vendar se premika preferenčno in sicer tako, da iz frontierja vedno vzame stran z najmanjšim indeksom?
   Trajektorija A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 →10 → 11
   Trajektorija B: 11 → 10 → 9 → 8 → 7 → 6 → 5 → 4 → 3 →2 → 1
   Trajektorija C: 3 → 5 → 1 → 2 → 4 → 7 → 6 → 8 → 9 → 10 → 11
   Trajektorija D: 3 → 5 → 1 → 2 → 4 → 6 → 7 → 8 → 9 → 10 → 11

SLO

Naloga 2 (vrednost naloge 30 točk)

Uporabnik surfa po spletu, ki ga predstavlja omrežje iz naloge 1. Premika se naključno; začetno stran izbere naključno in prav tako vsako naslednjo stran, le da pri premiku na naslednjo stran naključno izbira izmed strani, ki so dostopne iz trenutne strani.

a) Kakšna je verjetnost, da bo uporabnik po treh korakih (prvi korak je, ko naključno izbere začetno stran) na spletni strani z indeksom 6?
b) Matriko sosednosti iz naloge 1 popravite tako, da bo ustrezala lastnostim, ki so pomembne za izračun PageRank-a. V primeru, da boste uporabili kakršenkoli parameter, zapišite kateri in kakšno vrednost ima.

SLO

Naloga 3 (vrednost naloge 20 točk)
Na voljo imamo 5 dokumentov, ki so vsi kategorizirani v kategorijo Šport.

a) Izračunajte, ali se dokument d6 uvršča v kategorijo šport, če:
   - za izračun uporabite Bayesovo teorijo,
   - nas kot pomembne zanimajo le besede z naslednjimi lemami: {igralec, kolesarski, nagrada, najboljši, niz, odličen, športnik},
   - vemo, da je meja za uvrstitev v kategorijo Šport glede na vrednosti klasifikatorjev drugih kategorij enaka 0,0621,
   - ne uporabljamo glajenja.

b) Iz slovarja besed, ki nas zanimajo, smo odstranili eno besedo in zaradi nje se je verjetnost razvrstitve v kategorijo Šport povečala. Katerih dveh besed zagotovo nismo odstranili?

Korpus:

d1. Slovenska prepoznavnost je predvsem zasluga naših odličnih športnikov.
d2. Slovenski športniki želijo nadaljevati uspešen niz proti Poljakom.
d3. Najboljši športniki so tisti igralci, ki jih ne zanimajo nagrade.
d4. Ti igralci so najboljši športniki in si zaslužijo nagrado.
d5. Športniki, potegujte se za odlično udeležbo in nagrade bodo vaše.

Nerazvrščeni dokument:

d6. Športnik leta prejel denarno nagrado.

SLO

Naloga 4 (vrednost naloge 20 točk)

V podjetju *SuperShoppster* prodajajo različne izdelke in redno menjujejo dobavitelje. Hkrati tudi sledijo spletnim virom, da ponujajo res najnižje cene izdelkov. V ta namen so pridobili korpus spletnih strani, ki vsebuje 1 milijon spletnih strani v HTML obliki. Zapišite naslednje postopke, pri čemer za parametre, ki jih postopki potrebujejo, razložite, kako jih pridobite.

a) Napišite psevdokodo algoritma, ki bo ugotavljal ali je spletna stran uporabna za nadaljnjo obdelavo podjetja *SuperShoppster* ali ne. Uporabite lahko tudi interne baze podatkov podjetja.

b) Ko izvajate avtomatsko ekstrakcijo vsebine iz podanih strani, se lahko zgodi, da namesto celotnih objektov (angl. data records), algoritem prepoznava atribute posameznih objektov kot samostojne objekte. Kako lahko omejite takšne napake, če uporabljate algoritem *Mining Data Records (MDR)*, ki smo ga spoznali na predavanjih? Razložite tudi zakaj prihaja do tega.

Algorithm: MDR($Node$, K, τ, d)
1 **if** TreeDepth($Node$) $\geq d$ **then**
2 CombComp($Node.Children$, K)
3 $DataRegions \leftarrow$ IdenDRs($Node$, K, τ)
4 **if** $(UncoveredNodes \leftarrow Node.Children$ - $\cup_{DR \in DataRegions}\ DR) \neq \emptyset$ **then**
5 **foreach** $ChildNode \in UncoveredNodes$ **do**
6 DataRegions $\leftarrow$ DataRegions $\cup$ MDR($ChildNode$, K, τ, d)
7 return $DataRegions$
8 **else**
9 return $\emptyset$