

*Trajanje: 120 minuta. Rješenja zadataka na košuljici i po potrebi na zasebnim papirima. Zaokružiti na košuljici redni broj rješavanog zadatka (1-10).*

1. (3 boda) Pseudokodom opisati *Iterative Message Passing* predložak za *MapReduce* programski model. Neka je zadan neusmjereni graf u kojem postoje dva tipa čvorova: bijeli i crni čvorovi (**Napomena:** *vidi sliku 1*). Korištenjem opisanog predloška rješiti problem izračunavanja udaljenosti do najbližeg crnog čvora u grafu za sve čvorove u grafu. Potrebno je pseudokodom opisati implementacije *svih metoda i struktura podataka* koje se koriste u *Iterative Message Passing* predlošku za rješenje danog problema. Za zadani graf na slici 1 prikazati vrijednosti korištenih struktura podataka u svakom čvoru za sve iteracije algoritma.
2. (3 boda) Zadana su tri teksta  $A$ ,  $B$  i  $C$ :  $A = \text{padaju zvijezde}$ ,  $B = \text{ratovi zvijezda}$  i  $C = \text{kisa pada u maju}$ . Za svaki tekst izračunati sve *shingleove* duljine  $s = 3$  te prikazati tekstove u tablici kao stupce nula i jedinica gdje su redci tablice *shingleovi* (**Napomena:** *praznine ignorirati tj. izbaciti*). Jedinica u tablici označava da je neki *shingle* prisutan u tekstu dok nula označava da *shingle* nije prisutan u tekstu. Nakon toga, na skupovne reprezentacije tekstova potrebno je primijeniti algoritam *MinHash* korištenjem sljedeće četiri funkcije  $f_1(r) = r + 1 \bmod N$ ,  $f_2(r) = r + 3 \bmod N$ ,  $f_3(r) = r + 6 \bmod N$  i  $f_4(r) = r + 11 \bmod N$ , gdje je  $r$  redni broj retka u tablici, a  $N$  jest ukupan broj redaka odnosno *shingleova*. Primjenom četiriju funkcija hashiranja izgraditi sažetke tekstova te izračunati sličnost za sve parove tekstova. Konačno, izračunati prosječnu vrijednost apsolutnog odstupanja *MAE* (eng. *Mean Absolute Error*) sličnosti sažetaka u odnosu na sličnost tekstova. Sličnost sažetaka računati prema formuli  $\text{Sim}_{H_A, H_B} = \frac{K}{M}$ , gdje je  $K$  broj redaka u kojima se sažeci  $H_A$  i  $H_B$  poklapaju, a  $M$  je ukupan broj redaka tj. broj hash funkcija koje se koriste. Za stvarnu sličnost tekstova koristiti *Jaccardovu* sličnost skupova koji su pridruženi tekstovima.  
a.) Koliko iznosi vjerojatnost da par dokumenata koji ima sličnost 95% ne postane kandidat za izravnu usporedbu - lažno negativni *FN* (eng. *False Negative*) par?  
b.) Pretpostavimo da postoji neki par dokumenata kojem se sažeci poklapaju 15%, kolika je vjerojatnost da takav lažno pozitivan *FP* (eng. *False Positive*) par dokumenata postane kandidat za izravnu usporedbu?  
c.) Skicirati kako izgleda idealna funkcija sažimanja neosjetljivog na lokalne promjene te pokraj nje skicirati funkciju koja se dobije korištenjem metode *LSH*. Na grafu označiti područja koja predstavljaju lažno negativne i lažno pozitivne kandidate. Skicirati što će se dogoditi s grafom funkcije ako se koriste  $b = 32$  pojasa svaki veličine  $r = 8$  bita.
3. (3 boda) U sustavu postoji  $N = 10^9$  različitih tekstualnih dokumenata među kojima postoji određeni broj duplikata. Za identifikaciju duplikata koristi se metoda generiranja sažetaka pomoću neke funkcije sažimanja (npr. *Sim-Hash*). Zbog velikog broja dokumenata/sažetaka nemoguća je izravna usporedba svih parova dokumenata/sažetaka pa se koristi tehniku *sažimanja osjetljivog na lokalne promjene* (eng. *Locality Sensitive Hashing*) *LSH*. Veličina sažetka iznosi 256 bita, koristi se  $b = 16$  pojaseva svaki veličine  $r = 16$ . Cilj je identificirati sve parove kandidata za izravnu usporedbu kojima se sličnost sažetaka poklapa do uključivo 95%, tj. sažeci dokumenata se poklapaju u barem 95% bita.
4. (3 boda) Za zadani graf na slici 2 napisati jednadžbe toka ranga/utjecaja (eng. *rank*) za sve čvorove u grafu (eng. *Flow Equation Formulation*). Analitičkim postupkom rješiti sustav jednadžbi. Napisati jednadžbu u matričnom obliku (eng. *Matrix Formulation*). Metodom uzastopnog potenciranja izračunati vrijednost rang vektora  $r$  za prve tri iteracije algoritma. Rang vektor u početku treba inicijalizirati tako da svi čvorovi dobiju jednak rang/utjecaj. Izračunati prosječnu apsolutnu kvadratnu pogrešku *RMSE* (eng. *Root Mean Square Error*) metode uzastopnog potenciranja kroz tri iteracije u odnosu na egzaktno analitičko rješenje.
5. (3 boda) Za zadani graf na slici 3 napisati jednadžbe toka ranga/utjecaja (eng. *rank*) za sve čvorove u grafu (eng. *Flow Equation Formulation*).
  - a.) Primjeni li se metoda uzastopnog potenciranja s ciljem izračunavanja vektora ranga  $r$  na zadani graf, hoće li dobiveni rezultati biti vjerodostojni? Detaljno objasniti odgovor.
  - b.) Modificirati zadani graf uvođenjem teleportirajućih poveznica (eng. *teleports*). Napisati vrijednost *Google* matrice  $A$  ako vjerojatnost da će slučajni šetač (eng. *random walker*) slijediti poveznicu iz grafa iznosi  $\beta = 0.8$ .
6. (3 boda) Bloomov filter od  $3 \cdot 10^6$  bita je popunjén s  $5 \cdot 10^5$  elemenata. Odredite pogrešku (izraženu kao razlika vjerojatnosti pojavljivanja lažno pozitivnog rezultata) pri korištenju 10 funkcija sažimanja u odnosu na optimalni broj funkcija sažimanja.
7. (3 boda) Prikažite stanje struktura podataka nakon izvođenja prvog, međukoraka i drugog koraka algoritma PCY za zadani skup košara. Pretpostavite da algoritam redom obilazi košare  $B_1 - B_6$  te predmete unutar košara s lijeva

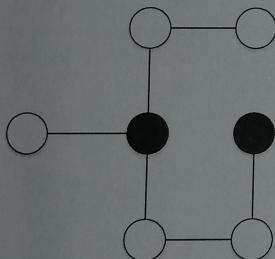
na desno. Tim redoslijedom algoritam dodjeljuje indekse predmetima. Prag potpore (engl. *support*) iznosi  $2/3$ . Prepostavite da algoritam u prvom koraku koristi funkciju sažimanja  $h(i, j) = (i \cdot 4 + j) \bmod 5$ , pri čemu su  $i$  i  $j$  indeksi predmeta i vrijedi:  $0 \leq i < j < N$ , pri čemu je  $N$  ukupan broj različitih predmeta.

$$\begin{array}{ll} B_1 = \{a, b, c, d\} & B_4 = \{b, c, d\} \\ B_2 = \{a, c, d\} & B_5 = \{a, d\} \\ B_3 = \{a, b, c\} & B_6 = \{b, c\}. \end{array}$$

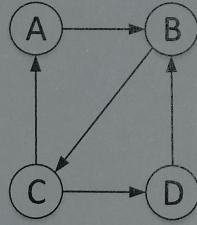
8. (3 boda) Za zadani skup košara odredite sve česte podskupove podataka. Prepostavite da je prag potpore (engl. *support*)  $1/3$ .

$$\begin{array}{ll} B_1 = \{a, b, c, d, e\} & B_4 = \{a, b, d, e\} \\ B_2 = \{a, c, d, e\} & B_5 = \{a, d\} \\ B_3 = \{b, c, e\} & B_6 = \{b, c, e\}. \end{array}$$

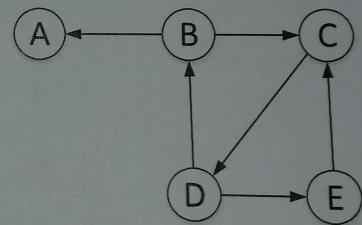
9. (3 boda) a.) Napišite formulu za *Jaccardovu sličnost* dvaju skupova.  
 b.) Kratko i jasno objasnite što radi funkcija sažimanja  $h_\pi = \min_\pi(D)$  koja se koristi u *MinHash* algoritmu.  
 c.) Dokažite da gore spomenuta funkcija sažimanja  $h_\pi(\cdot)$  zadržava sličnost dokumenata tj. da vrijedi  $p[h_\pi(A) = h_\pi(B)] \sim Jaccard\_Sim(A, B)$ , gdje su  $A$  i  $B$  dva dokumenta predočena skupovima elemenata.
10. (3 boda) a.) Matematički izrazite drugi moment toka podataka.  
 b.) Dokažite da izraz za procjenu drugog momenta toka podataka prema Alon-Matias-Szegedyju konvergira ka izrazu iz a.) dijela zadatka. Detaljno objasnite svaki korak dokaza!



Slika 1: Grafovi uz zadatak 1



Slika 2: Grafovi uz zadatak 4



Slika 3: Grafovi uz zadatak 5