AVSP - LJIR 2017.

1. Pretpostavimo da sustav Minhash stvara sažetke na ulaznim dokumentima duljine 6 redova. 4 od 6 redova sadrži 0, preostala 2 reda sadrže 1. Sveukupno onda postoji g! = 720 permutacija redova. Prilikom stvaranja Minhash sažetaka izabire se redni broj prvog reda koji u permutiranom poretku sadrži vrijednost 1.

a) Za koliko od ukupno 720 permutacija će vrijednost Minhash funkcije imati vrijednost 6?

b) Za koliko od ukupno 720 permutacija će vrijednost Minhash funkcije imati vrijednost 5?

c) Za koliko od ukupno 720 permutacija će vrijednost Minhash funkcije imati vrijednost 4?

2. Dva sustava za preporučivanje. R1 user-item tablica ima posjetitelje - države, a sadrži 1 ako je posjetitelj posjetio državu, inače sadrži 0. R2 user-item tablica ima posjetitelje - države, a sadrži ocjenu države ako ju je korisnik ocijenio.

a) Definirati matematičkim izrazom Jaccard, cosine i Pearson mjere sličnosti.

b) Koja od te 3 je najprimjerenija za R1 i zašto druge nisu?

c) Koja od te 3 je najprimjerenija za R2 i zašto druge nisu?

3. Sažetci duljine 12 bita, LSH, koriste se pojasi veličine 2, 4 i 6 bita.

a) Vjerojatnost da par dokumenata koji ima sličnost 85% ne postane kandidat za usporedbu

b) Vjerojatnost da par dokumenata koji ima sličnost 15% postane kandidat za usporedbu

c) Zadaci a) i b) ako se koriste 3 pojasa duljine 4 bita

d) Skicirati idealnu i stvarnu funkciju sažimanja, što se događa s vjerojatnosti iz c) ako sustav koristi 4 pojasa duljine 3 bita?

4. a) Opisati rad sustava za brzo pronalaženje sličnih sažetaka iz permutacijske tablice.

b) Duljina sažetka f 0 64, a broj sažetaka je $|F| = 2^{34}$. Sažetci su podijeljeni na blokove duljine 13, 13, 13, 13, 12 bitova, signifikantni dio je 2 bloka. Koliki je broj permutacijskih tablica i koliko će se sažetaka uspoređivati Hammingom?

5. Dvije inačice Bloomovog filtera. Inačica A je kanonska inačica (normalna). Inačica B ima dvije grupe funkcija – $f_i(x)$, domena [1, n] i $g_i(x)$, domena [1, m]. Prvo se za sve f_i postavljaju bitovi na 1, a potom za sve g_i na 0. Sve funkcije su uniformne na [0, m-1]

Prvo za inačicu A pa za inačicu B:

a) Vjerojatnost postavljanja nekog bita na 1 prilikom unosa novog elementa

b) Vjerojatnost da se pojavi false positive

c) Vjerojatnost da se pojavi false negative

6. CURE algoritam

Točke (2, 4), (3, 3), (3, 8), (5, 3), (5, 6), (7, 2), (6, 8), (9, 6). Reprezentativne točke su (3, 3) i (9, 6), pomak je 20%.

Kojoj grupi pripada točka (5, 6)?

7.

B_1 = (a, b, c, d)
B_2 = (a, c)
B_3 = (b, c)
B_4 = (a, b)
B_5 = (a, d)
B_6 = (a, b, c)
B_7 = (a, b, d)

a) Odrediti česte podskupove za support = 3

b) Odrediti interesantnost od a -> c

c) Na koji je način primjerenije pohranjivti podatke za zadani skup košara, trokutastom matricom ili raspršenim adresiranjem?