Analitička obrada podataka, I kolokvijum, 19. 4. 2021.

1. (20 bodova) Dato je 100 item-a. Označeni su brojevima od 1 do 100. Baza se sastoji od 100 transakcija. Označene su brojevima od 1 do 100. Item i pripada transakciji b akko je i djelilac b. Na primjer, 1 pripada svakoj transakciji a transakcija sa rednim brojem 12 sastoji se od 1, 2, 3, 4, 6 i 12.
    a. Ako je min_support = 5 koji su item-i frekventni?
    b. Ako je min_support = 5 koji su parovi item-a frekventni?
    c. Kolika je suma |t1|+|t2|+…+|t100|, gdje || označava broj item-a u transakciji?
    d. Koja transakcija sadrži najveći broj elemenata?
    e. Koliko iznosi confidence za pravila {5, 7} -> 2 i {2, 3, 4} -> 5?
    f. Opisati sva asocijativna pravila sa confidence = 100%.

2. (20 bodova) Za korak spajanja u Apriori algoritmu predlaže se metoda $C_k = F_{k-1} \bowtie F_1$ po kojoj kandidatski k-skup nastaje proširivanjem frekventnog (k-1)-skupa sa frekventnim 1-skupom. Na primjer, ako je $k = 3$, spajanjem $\{2, 3\} \in F_2$ i $\{1\} \in F_1$formira se $\{2, 3, 1\} \in C_3$.
    a) Ako je $F_3 = \{123, 124, 125, 134, 135, 234, 235, 345\}$ napisati skup $C_4 = F_3 \bowtie F_1$. Obrazložiti odgovor.
    b) Dokazati da je prethodna metoda kompletna.

3. (20 bodova) Data je šema relacione baze podataka:
    time(time_key, day, day_of_week, month, quarter, year)
    item(item_key, item_name, brand, type, supplier_type)
    branch(branch_key, branch_name, branch_type)
    location(location_key, street, city, province_or_state, country)
    sales(time_key, item_key, branch_key, location_key, number_of_units_sold, price)
    a) Nacrtati odgovarajuću šemu zvijezda koja sadrži dvije mjere: dollars_sold i units_sold
    b) Nacrtati mrežu kuboida koja čini 4-D kocku podataka za dimenzije: time, item, branch i location
    c) Napisati MDX upit koji generiše osnovni kuboid za 4-D kocku podataka iz zadatka pod b)
    d) Napisati MDX upit da se napravi lista sa ukupnom zaradom za svaku prodavnicu iz Dalasa za godinu 2021.

4. (20 bodova) Data su dva binarna vektora X=0101010001 i Y=0100011000. Izračunati Hamingovo i Jakardovo rastojanje između ova dva vektora.
    a) Napisati vezu između Hamingovog rastojanja i SMC;
    b) Izračunati kosinus mjeru sličnosti za vektore X i Y. Objasniti šta je zajedničko za Jakardovo rastojanje i kosinus mjeru sličnosti.

5. (20 bodova) Dokazata da važi Pr(*minhash*(C1) = *minhash*(C2)) = JACCARDSIM(C1 , C2), gdje su C1 i C2 proizvoljne dvije kolone u karakterističnoj matrici.