1. U algoritmu grupiranja BFR skup odbačenih točaka (engl. discard set) predstavlja:
    - Točke koje su dovoljno blizu centroida grupe, a ne zadržavaju se u radnoj memoriji

2. Neka je zadana sljedeća "*Utiliity*" matrica u kojoj su znakom "-" označene ocjene koje nedostaju. Izračunajte sličnost između korisnika U_1 i U_2 korištenjem Cosine similarity mjere sličnosti.

	I_1	I_2	I_3	I_4
U_1	-	-	5	1
U_2	-	4	4	1

    - 0.72

3. U točki $C_{x,y}$ = (1, 2) zadan je centroid oko kojeg su podaci normalno distribuirani. Odredite Mahalanobisovu udaljenost točke $T_{x,y}$= (3, 4) od centroida, ako varijance po dimenzijama iznose: v_x = 0.01 i v_y = 0.02.
    - 24.49

4. Neka je zadan skup košara:

    B_1 = {1, 2, 3} B_4 = {1, 3, 4}

    B_2 = {1, 4} B_5 = {4}

    B_3 = {1, 3, 4} B_6 = {2, 3, 4}

    Pod pretpostavkom da prag potpore (engl. support threshold) iznosi 3, koliko podskupova podataka je često?
    - 6

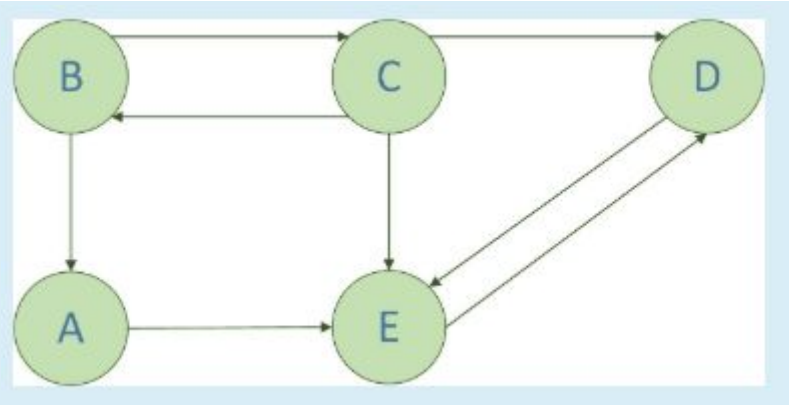
5. Promatramo računanje Rabinovog sažetka nad n-gram prozorom proizvoljnog teksta koristeći kodiranje polinomima. Ako je o_i broj operacija potrebnih za računanje sažetka i-tog n-gram prozora (i je zadan u [1, n]). Onda vrijedi:
    - $o_1 > o_i$, za svaki i > 1

6. Neka je zadana sljedeća "Utiliity" matrica u kojoj su znakom "-" označene ocjene koje nedostaju:

	I_1	I_2	I_3	I_4
U_1	-	-	5	1
U_2	-	4	4	1

    Izračunajte sličnost između korisnika U_1 i U_2 korištenjem Jaccard similarity mjere sličnosti.
    - ⅔, 0.67

7. Zadan je graf na slici za koji je potrebno izračunati vektor utjecaja (engl. rank vector). Primjeni li se metoda uzastopnog potenciranja s ciljem izračunavanja vektora ranga r na zadani graf, hoće li dobiveni rezultati biti vjerodostojni?



- Ne, u grafu postoji "paukova zamka" (engl. spider trap).

8. Prilikom stvaranja skupovnih reprezentacija tekstualnih dokumenata za potrebe algoritma MinHash, označite tvrdnju koja vrijedi za veličinu shingleova kada se razmatraju dulji tekstualni dokumenti.

- Uputno je koristiti duže shingleove

9. Pretpostavite da su provođenjem algoritma MinHash nad izvornim skupovnim reprezentacijama dobiveni sljedeći sažeci:

	s0	**s1**	**s2**	**s3**	**s4**
$\mathbf{b_0}$	2	5	4	7	3
$\mathbf{b_1}$	7	6	2	8	5
$\mathbf{b_2}$	4	3	3	5	6
$\mathbf{b_3}$	4	3	2	4	2

Nadalje, pretpostavite da se koristi algoritam sažimanja osjetljivog na bliskost (engl. Locality Sensitive Hashing) s veličinom pojasa b=2.
Unutar prvog pojasa koristi se sljedeća funkcija sažimanja:

$f_1 = (b_0 * 10 + b_1) \bmod 6$,

Unutar drugog pojasa koristi se sljedeća funkcija sažimanja:

$f_2 = (b_2 * 10 + b_3) \bmod 6$.

Npr. sažetak s_0 u prvom pojasu se raspršio u pretinac p=3:

$p = (2 * 10 + 7) \% 6 = 3$.

Koliko će biti parova kandidata za sličnost kada se završi algoritam za oba pojasa?

- 4

10. Neka su zadana dva dokumenta D_1="ABECEDA" i D_2="CEDAR". Izračunajte sažetke dokumenata koristeći MinHash algoritam uz duljinu shingleova L=3, koristeći dvije funkcije sažimanja f_1 i f_2 umjesto permutacija prema zadanoj tablici.

r	shingle	$f_1 = (r + 1) \bmod 6$	$f_2 = (r + 2) \bmod 6$
0	ABE	1	2
1	BEC	2	3
2	ECE	3	4
3	CED	4	5
4	EDA	5	0
5	DAR	0	1

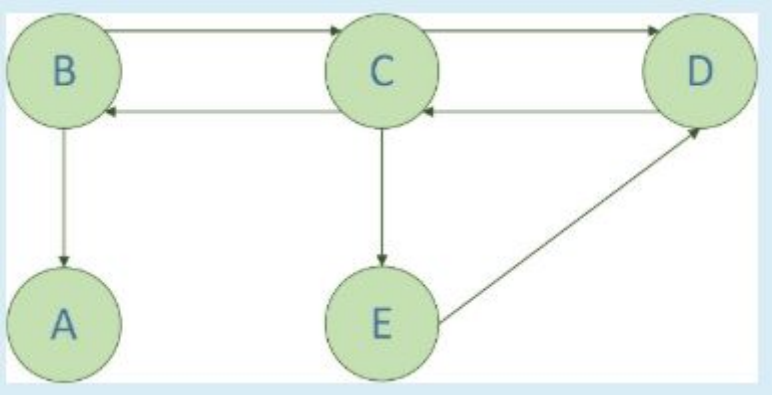
Izračunajte sličnost dobivenih sažetaka kao omjer: broj redaka u kojima su sažeci jednaki i ukupni broj redaka u sažecima!

- 0.5

11. Negativna granica u Toivonenovom algoritmu definira se kao skup podskupova podataka X za koji vrijedi:

- Niti jedan y iz X nije čest u uzorku, ali svi podskupovi od y nastali uklanjanjem točno jednog elementa jesu česti

12. Zadan je graf na slici za koji je potrebno izračunati vektor utjecaja (engl. rank vector).



Primijeni li se metoda uzastopnog potenciranja s ciljem izračunavanja vektora ranga r na zadani graf, hoće li dobiveni rezultati biti vjerodostojni?

- Ne, u grafu postoje "mrtvi čvorovi" (engl. dead end nodes).

13. U tablici ispod zadana je matrica ocjena korisnika za pojedine filmove (engl. user-item matrix, utility matrix). Prazna polja u matrici predstavljaju ocjene koje nedostaju. Korištenjem algoritama suradničkog filtriranja (engl. Collaborative Filtering) potrebno je izračunati ocjenu za korisnika U_4 i film M_2 ako se koristi User-User pristup suradničkog filtriranja.
Bitne napomene:

- Za računanje sličnosti među filmovima koristi se Pearson Correlation Coefficient.
- Sustav koristi najviše k = 2 najsličnijih filmova za izračun ocjene.
- Prilikom izračuna, filmovi čija sličnost je manja od 0 se ne uzimaju u obzir.

	U_1	U_2	U_3	U_4	U_5
M_1	2	3	4	1	3
M_2	-	3	3	-	4
M_3	-	2	3	1	5
M_4	1	2	5	2	3
M_5	2	5	4	-	5

Odgovor:

- 3.00

14. Algoritam simhash u svojoj kanonskoj inačici koristi 3-bitne sažetke, a jedinke su riječi (odvojene razmacima). Interna funkcija sažimanja je definirana s h(x) = duljina(x) % 8, gdje je x riječ, a duljina(x) broj znakova riječi. Decimalni simhash sažetak teksta "with or without you" jest:

- 7

15. Složenost algoritma hijerarhijskog grupiranja iznosi:

- $O(n^2 \log(n))$

16. U postupku generiranja pravila asocijacije $I \rightarrow j$ nad skupom predmeta I parametar pouzdanost (eng. confidence) definira se kao (potpora - eng. support):

- $conf(I \rightarrow j) = support(I \cup j) / support(I)$

17. Sustav za pretragu dokumenata je ispisao sljedeće dokumente za zadani upit: [D1, D3, D5]. Međutim, ručnom provjerom pokazano je da je sustav ipak trebao ispisati dokumente: [D1, D2, D5, D9, D10]. Kolika je ocjena uspješnosti F1 za navedeni upit?

- 0.5

18. Pretpostavite da se koristi algoritam sažimanja osjetljivog na bliskost (engl. **Locality Sensitive Hashing**). Ukupna duljina sažetaka jest N elemenata te se pritom koristi B pojaseva svaki duljine B elemenata. Što će se dogoditi ako se **broj pojaseva poveća** na 2B, a **duljina jednog pojasa smanji** na B/2 elemenata?

- Broj **lažno negativnih** (engl. false negative) parova kandidata za sličnost će se **smanjiti**, dok će se broj **lažno pozitivnih** (engl. false positive) parova kandidata za sličnost **povećati**.

19. Ulaz u algoritam simhash su "n-gram" jedinke. Ako je svaka n-gram jedinka dodatna dimenzija u višedimenzionalnom prostoru pojavljivanja jedinki, onda s **povećanjem parametra "n"** načelno vrijedi:

- **sažeci sličnih** dokumenata su sve **manje slični** i **dimenzionalnost** sustava **raste**

20. Koji od navedenih algoritama grupira skup podataka u samo jednom prolazu?

- CURE
- Aglomerativno hijerarhijsko grupiranje
- BFR
- k-means
- Divizivno hijerarhijsko grupiranje

21. Neka je dostupan skup simhash sažetaka i neka se koristi sustav brze paralelne pretrage bliskih sažetaka koristeći više permutiranih tablica. U takvom sustavu je potrebno provjeriti Hammingovu udaljenost k za određeni broj sažetaka. Ako je k unaprijed zadana udaljenost, onda **smanjenjem k**, načelno vrijedi:

- **Preciznost** se **povećava**, a **odziv** se **smanjuje**.

22. U postupcima pronalaska čestih podskupova podataka pojam **potpore podskupa (itemset support)** definira se kao:

- Ukupan broj košara u kojima se pojavljuje odgovarajući podskup predmeta

23. U sustavu za preporučivanje knjiga korisnicima, koristimo jednostavni (osnovni) model matrične faktorizacije s dvije skrivene značajke (K = 2). U tom modelu, knjizi Hobit pridružen je vektor skrivenih značajki [0.2, -0.75], a knjizi Pinokio vektor [-0.63, 0.4]. Model predviđa da će interes korisnika Ivice za Hobita biti 0.23, a interes Marice za Pinokija biti 0.77. Poznato je da Ivica i Marica knjige čitaju zajedno i jednako ih ocjenjuju pa su njihovi vektori skrivenih značajki jednaki. Vaš je zadatak izračunati taj vektor. Zbrojimo li elemente tog vektora, dobit ćemo (zaokruženo na dvije decimale):

- -2.47

24. Označiti točan odgovor ako se razmatraju prostorna i vremenska složenost algoritma Flajolet-Martin: Odaberite jedan odgovor:

- vremenska složenost je obrnuto proporcionalna broju korištenih funkcija sažimanja
- prostorna složenost je obrnuto proporcionalna broju korištenih funkcija sažimanja
- prostorna složenost ne ovisi o broju funkcija sažimanja
- prostorna složenost je proporcionalna broju korištenih funkcija sažimanja

25. Uz podatkovnu matricu A i glavne komponente v_1 i v_2 izračunajte sadržaj matrice A' koja je rekonstrukcija točaka u originalni podatkovni prostor nakon PCA transformacije.

$$A = \begin{bmatrix} 1 & 1 \\ 1.5 & 2 \\ 2 & 2 \\ 3 & 3 \\ 3.5 & 3 \\ 4 & 4 \\ 5 & 5 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$v_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$$

- $$A' = \begin{bmatrix} 1 & 1 \\ \frac{7}{4} & \frac{7}{4} \\ 2 & 2 \\ 3 & 3 \\ \frac{13}{4} & \frac{13}{4} \\ 4 & 4 \\ 5 & 5 \end{bmatrix}$$

26. Što od navedenog **ne vrijedi** za autoenkodere?

- Funkcija dekodera jest rekonstruirati ulaz iz njegove sažete, kodirane reprezentacije.
- Autoenkoder je građen od enkodera i dekodera.
- Regularizacijom kontraktivnog autoenkodera (engl. contractive autoencoder) nastoje se smanjiti iznosi derivacija težina u skrivenim slojevima.
- Ograničavanje preslikavanja ulaza na izlaz kod rijetkog se autoenkodera (engl. sparse autoencoder) s jednim skrivenim slojem postiže dodatnim kažnjavanjem težina neurona ulaznog sloja.
- Učenje konvolucijskih autoenkodera podrazumijeva učenje optimalnih filtera za detekciju specifičnih značajki u "rešetkastim" podacima.

27. Što od navedenog **nije mjera** primjenjiva za ocjenu preciznosti sustava za predviđanje i preporučivanje?

- Rank correlation
- Normalized Discounted Cumulative Gain (NDCG)
- Root Mean Square Error (RMSE)
- Hit ratio (HR)
- Alternating least squares (ALS)

28. Svojstvo lokalnosti društvenih mreža definira se na sljedeći način:

- Ako postoji brid između čvorova A i B te čvorova A i C, onda je vjerojatnost da su čvorovi B i C povezani natprosječna.

29. Kompetitivni omjer BALANCE algoritma za dva oglašivača iznosi:

- 0.75

30. Inačica m-bitnog Bloomovog filtera koristi dvije grupe funkcija sažimanja: $f_i(x)$, i je iz [1, N] i $g_j(x)$, j je iz [1, M] te su sve funkcija sažimanja uniformne na [0, m-1]. Kod ove inačice Bloomovog filtera potrebno je prvo postaviti na 1 bitove na pozicijama $f_1(x),f_2(x),...,f_N(x)$ te potom postaviti na 0 bitove na pozicijama $g_1(x),g_2(x),...,g_M(x)$. Bloomov filter vraća potvrdan odgovor ako su svi bitovi na pozicijama $f_1(x),f_2(x),...,f_N(x)$ jednaki 1, a svi bitovi na pozicijama $g_1(x),g_2(x),...,g_M(x)$ jednaki 0.
Za opisanu inačicu Bloomovog filtera vrijedi:

- Moguća je pojava lažno pozitivnog rezultata i lažno negativnog rezultata.

31. Zadana je matrica snage pripadnosti čvorova zajednicama za BigCLAM algoritam.
Postoje dvije zajednice A i B pri čemu su snage pripadnosti čvorova X i Y zajednici A redom $X_A = 0.8$ i $Y_A = 0.92$.
Odredite koliko iznosi snaga pripadnosti čvora Y zajednici B (Y_B = ?) ako je poznato da X_B iznosi 0.66, a ukupna vjerojatnost da su čvorovi X i Y povezani u grafu iznosi 0.708.

- 0.75

32. Ulazni parametri Adwords problema su:

- Skup ponuda oglašivača za pojmove pretraživanja, CTR, budžet oglašivača i limit broja oglasa koji se mogu prikazati korisniku

33. Osnovni parametri AGM modela su:

- broj čvorova mreže, broj zajednica mreže, vjerojatnost povezanosti čvorova u zajednici i pripadnost čvorova zajednicama

34. Društvena mreža sa dvije zajednice opisana je AGM modelom. Vjerojatnost da su dva čvora povezana unutar zajednice A iznosi $p_A = 0.72$, a vjerojatnost da su dva čvora povezana, a pripadaju istovremeno zajednicana A i B iznosi 0.93. Koliko iznosi vjerojatnosd da su čvorovi povezani unutar zajednice B (p_B = ?)

- 0.75

35. Odaberite tvrdnju koja **je istinita** za duboke mreže vjerovanja.

- Veze u najdubljem sloju duboke mreže vjerovanja usmjerene su prema plićem sloju.
- Učenje duboke mreže vjerovanja započinje predtreniniranjem ograničena Boltzmannova stroja najbližeg njezinu izlazu.
- Usmjerene veze u nižim slojevima duboke mreže vjerovanja usmjerene su prema dubljem sloju.
- Latentne varijable u skrivenim slojevima uglavnom poprimaju binarne vrijednosti.
- Ulazni sloj duboke mreže vjerovanja specijaliziran je za detekciju značajki iz podataka.

36. Koji se od navedenih blokova (košara, tj. bucketa) **ne može pojaviti** u osnovnom Datar-Gionis-Indyk-Motwani (DGIM) algoritmu?

- Blok s 0 nula i 8 jedinica
- Blok s 80 nula i 8 jedinica
- Blok s 2 nule i 14 jedinica
- Blok s 1 nulom i 8 jedinica
- Blok s 0 nula i 1 jedinicom

Unofficial hint - broj jedinica mora biti potencija broja 2. Jedino u bloku veličine 1 ne smije biti niti jedna 0.

37. Koliki je postotak bitova postavljen na vrijednost 1 u 1000-bitnom Bloomovom filteru koji koristi 2 funkcije sažimanja za ukupno 100 unesenih elemenata, ako pretpostavimo da su korištene funkcije sažimanja savršene (bez kolizija)?

- 20%

38. Tri oglašivača A, B i C postavljaju ponude na pojmove oglašavanja X, Y i Z.
Oglašivač A postavlja ponude na pojmove X i Y, oglašivač B na pojmove Y i Z, a oglašivač C na pojmove X i Z.
Oglasi se prikazuju primjenom BALANCE algoritma pri čemu je:

- Početni budžet svih oglašivača isti i iznosi 3
- Cijena svih oglasa je ista i iznosi 1
- U slučaju izjednačene situacije algoritam daje prednost oglašivačima po abecedi, tako da A ima najviši, a C najniži prioritet

Navedite kojim će se redosijedom prikazivati oglasi ako je ulazni niz upita: **X X X Y Z Y Z Z Y**

- A C A B B A C B -

39. PCA je _____ metoda, a LDA je _____ metoda:

- nenadzirana, nadzirana