NTNU
Kunnskap for en bedre verden

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 12. august 2017

Eksamenstid (fra-til): 09.00-13.00

Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpemiddel tillatt.
Bestemt, enkel kalkulator tillatt.

Annen informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 2

Antall sider vedlegg: 0

Kontrollert av:

17.07.2017 Dato

Jon Olav Hauglid Sign

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig X	**2-sidig** ☐
sort/hvit X	**farger** ☐

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Oppgave 1 – Diverse – 20 % (alle deler teller likt)

a) I kontekst av web-bruk-gruvedrift, hva er sesjonering? Hvorfor kan dette være vanskelig? Forklar to heuristikker som kan brukes til å utføre sesjonering.

b) Forklar *støy* ("noise") og *outlier*.

c) *Curse of Dimensionality* kan føre til problem når man skal utføre klynging eller klassifisering på høy-dimensjonale datasett. Forklar minst to data-preprosessering-metoder som kan redusere problemene.

d) Forklar hvordan en likhetsmatrise ("similarity matrix") kan brukes til klyngingsvalidering.

Oppgave 2 – OLAP – 25 % (alle deler teller likt)

a) Forklar hva som menes med termene som er understreket i følgende definisjon av datavarehus: "data-warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data."

b) Forklar begrepene OLTP ("Online Transaction Processing") og OLAP ("Online Analytical Processing"). Legg vekt på å få frem forskjeller mhp. egenskaper og bruk.

c) Forklar *stjerne-skjema* og *snøflak-skjema.*

d) Forklar bitmap-indeks. I hvilke tilfeller er en bitmap-indeks egnet?

e) Forklar OLAP-operasjonene *slice* og *dice*.

Oppgave 3 – Klynging – 10 %

Forklar algoritmen for *DBSCAN.*

Oppgave 4 – Klassifisering – 15 % (5 % på a og 10 % på b)

a) Forklar *kryss-validering* ("cross validation").

b) Forklar *overtilpasning* ("overfitting"). Hva kan forårsake overtilpasning? Hva kan man gjøre for å redusere overtilpassing når man bruker beslutningstre?

Oppgave 5 – Assosiasjonsregler – 30 % (5 % på a, 10 % på b og 15 % på c)

a) Definer *maksimale* ("maximal") og *lukkede* ("closed") frekvente elementsett.

b) Anta handlekorg-data som er gitt under:

TransaksjonsID	Element
T1	ACD
T2	BCE
T3	ABCE
T4	BE

1) Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 2). Bruk $F_{k-1} \times F_{k-1}$-metoden for kandidat-generering.
2) Bruk apriori-algoritmen til å generere alle 3-elements assosiasjonsregler basert på resultatet i (1), gitt minimum konfidens på 100 %. Vis hvordan regler evt. kan "prunes".

c) Anta handlekorg-data som er gitt under:

TransaksjonsID	Element
T1	ABCD
T2	ABCD
T3	ABCEF
T4	ABEF
T5	ABDEFG
T6	AEF

Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støttetall (*minimum support count)* på 2.

1) Konstruer et FP-tre basert på datasettet.
2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgende kolonner for å vise resultatet:
   - Element
   - "Conditional pattern base"
   - "Conditional FP-tree"
   - Frekvente elementsett

   Forklar rekursivitet der dette er nødvendig.