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: 9. august 2016

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:

_____

Dato 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) Hva er en *outlier*?
b) Forklar *web-bruk-gruvedrift* (Web usage mining), hva som er målet, og hva som er typiske data man bruker i denne prosessen.
c) Forklar hvordan man kan gjøre kontinuerlige attributtverdier om til kategorier.
d) Anta to bit-vektorer *p* og *q*:

p = 1 0 1 0 1 0 0 1 1 1
q = 1 0 0 0 1 0 1 1 0 1

Regn ut Jaccard-koeffisienten for bitvektorene *p* og *q*.

Oppgave 2 – Datavarehus og OLAP – 30 % (alle deler teller likt)

a) Forklar hva som menes med ordene som er understreket i følgende definisjon av datavarehus:
”A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data."
b) Forklar *enterprise-varehus, data mart, og virtuelle varehus.*
c) Forklar *stjerne-skjema* og *snøflak-skjema.*
d) Forklar *konsept-hierarki.*
e) Forklar *materialisering av kuboider,* hensikten med materialisering, og hvordan man kan velge hvilke kuboider som skal materialiseres.

Oppgave 3 – Klynging – 15 % (alle deler teller likt)

a) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør hierarkisk agglomerativ klynging på dette datasettet ved å bruke MIN (single link) og Manhattan-distanse. Vis det resulterende dendrogrammet.
b) Forklar "Bisecting K-means". Hva er en viktig fordel med denne sammenlignet med ordinær K-means?

X	Y
4	8
4	9
4	10
4	13
4	14
5	3
5	7
5	14
6	15
6	16
6	19
7	11
7	16
7	17
7	18
7	19

Oppgave 4 – Klassifisering – 10 % (alle deler teller likt)

a) Forklar *klassifisering.*
b) Forklar hvordan man konstruerer et *beslutningstre.*

Oppgave 5 – Assosiasjonsregler – 25 % (alle deler teller likt)

a) Anta handlekorg-data som er gitt under. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$-metoden for kandidat-generering.

TransaksjonsID	**Element**
T1	ABGH
T2	ADGHK
T3	ABC
T4	ACD
T5	ACGHK
T6	ACGHK
T7	BD
T8	ADGHK

b) Konstruer et FP-tre basert på datasettet ovenfor.
c) Vis hvordan man kan bruke ECLAT for å finne støtte (-tall) for elementsettet AG.