Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2012-04-11
Sal	TER3
Tid	14-18
Kurskod	TDDD02
Provkod	TEN1
Kursnamn/benämning	Språkteknologi för informationssökning
Institution	IDA
Antal uppgifter som ingår i tentamen	*10*
Antal sidor på tentamen (inkl. försättsbladet)	*3*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*013-282422*
Besöker salen ca kl.	*14.45*
Kursadministratör (namn + tfnnr + mailadress)	*Helene Meisinger 281868, helene.meisinger@liu.se*
Tillåtna hjälpmedel	*Inga*
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	
Vilken typ av papper ska användas, rutigt eller linjerat	*valfritt*
Antal exemplar i påsen	

TENTAMEN

TDDD02 Språkteknologi för informationssökning

onsdag 11 april 2012 kl. 14-18

Inga hjälpmedel är tillåtna.

Varje fråga är värd 3 poäng. Maximal poäng är 30. 15 poäng ger säkert godkänt. Det går bra att besvara flera frågor på samma papper.

1. Antag att vi representerar dokument i en vektorrymd med hjälp av åtta förvalda termer och deras frekvens i dokumenten. Tre dokument är givna med sina representationer enligt nedan. En given sökfråga fick representationen <0,0,1,0,0,0,1,0>. Ordna dokumenten efter relevans för sökfrågan. Motivera din rangordning.

   D1: <0,4,4,1,0,3,1,0>

   D2: <1,0,0,0,3,0,0,2>

   D3: <0,0,1,4,2,0,0,1>

   D4: <3,1,0,0,2,1,0,3>

2. Ett frågebesvarande system testades med ett frågebatteri som omfattade 160 frågor. Systemet returnerade svar på 112 av dessa frågor, varav 96 stycken svar var korrekta. Beräkna systemets precision och recall. Motivera ditt svar.

3. Ange en reguljär substitution som stoppar in ett blanktecken i en sekvens av två tecken där det första är en bokstav och det andra är något av skiljetecknen komma, kolon eller punkt.

4. I en korpus hittas följande frekvenser för några olika ordsekvenser:

inga spår av	20
spår av den	10
spår av	90
av den	900
inga spår	50
spår	80
den	3600

   Beräkna uppskattningar för följande sannolikheter utifrån dessa data: *p(av | spår), p(av | den), p(av | inga spår), p(den | av)*, eller motivera varför sannolikheten inte går att beräkna.

Institutionen för datavetenskap
Linköpings universitet
LINKÖPINGS UNIVERSITET
lars.ahrenberg@liu.se

5. Definiera Levenshteinavstånd och ange Levenshteinavstånd för orden *spindel* och *spider* samt visa hur det kan beräknas.

6. (a) Förklara vad som menas med en språkmodell (eng. *language model*). (b) Förklara också hur man kan jämföra prediktionsförmåga hos två olika språkmodeller på en given texttyp.

7. Vilka är de centrala komponenterna i standardarkitekturen för ett frågebesvarande system (QA-system)?

8. Förklara termerna anaforiskt pronomen och antecedent och formulera en mening eller en kort text som visar att antecedenten inte alltid är den nominalfras som är närmast anaforen, trots att den har samma morfologiska egenskaper.

9. Klustring, eller gruppering av ord som har liknande betydelse, baseras ofta på antagandet att sådana ord uppträder i likartade kontexter. Ange någon metod som kursen tagit upp som kan tillämpas på problemet klustring.

10. Textsammanfattningssystem baseras ofta på extraktion av en delmängd av de meningar som ingår i texten. Ange tre vanliga, beräkningsbara indikatorer på att en mening är lämplig att ta med i en sammanfattning.