Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2012-08-21
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	*282422*
Besöker salen ca kl.	*14.50*
Kursadministratör (namn + tfnnr + mailadress)	*Helene Meisinger 281868, helene.meisinger@liu.se*
Tillåtna hjälpmedel	
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	

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

TDDD02 Språkteknologi för informationssökning

tisdag 21 augusti 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. Ett dokumentsökningssystems prestanda mäts ofta med måtten precision och recall. (a) Ge definitioner av dessa två mått, (b) Förklara varför båda måtten behövs.

2. Förklara (a) vad som menas med en term-dokumentmatris och (b) måttet tf*idf.

3. Ange ett reguljärt uttryck som känner igen substantiv som innehåller avledningssuffixet 'ning'. Uttrycket ska utformas så att alla böjningsformer av sådana substantiv kan hittas och ha hög precision.

4. Ange med ledning av frekvensangivelserna i tabellen nedan Maximum Likelihood-uppskattningar av (a) bigramsannolikheten p(på|tänker), (b) trigramsannolikheten p(dig| tänker på). c) Ange också någon nackdel med Maximum Likelihood-uppskattningar.

tänker	100
på	5000
dig	400
tänker på	30
på dig	40
dig på	5
på tänker	1
tänker på dig	3

5. Ange ett uttryck för sannolikheten av sekvensen '*<s> nu börjar skolan igen*' givet att vi känner till bigramsannolikheterna för orden i sekvensen. <s> anger meningsstart.

6. Ange Levenshteinavståndet mellan orden *sommar* och *romans* och visa hur man räknar fram det algoritmiskt, t.ex. genom att ställa upp en matris för avstånden mellan delsträngar.

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

Tel: 282422
e-post:
lars.ahrenberg@liu.se

7. Antag att vi ur texter vill extrahera så många korrekta par <A, B> som möjligt, där A står för en stad och B för det land som staden ligger i. Exempel: Om en text innehåller meningen "Sent på kvällen landar vi i Quito, Equadors huvudstad." vill vi extrahera paret <Quito, Equador>. Beskriv hur man kan använda en s.k. bootstrapping-metod för att göra detta.

8. (a) Vad skiljer informationsutvinning (eng. information extraction) från informationssökning? (b) Ange minst fyra vanliga komponenter i ett informationsutvinningssystem.

9. Beskriv en vanlig arkitektur för ett frågebesvarande system.

10. Förklara metoden Naive Bayes och hur den kan tillämpas på problemet att bestämma vilken betydelse förekomsten av ett visst ord har i en given kontext.