Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2010-08-24*
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	*2422*
Besöker salen ca kl.	
Kursadministratör (namn + tfnnr + mailadress)	*Elisabeth Qvarnström 013-281706, eliqv@ida.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	
Antal exemplar i påsen	*5+2*

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

TDDD02 Språkteknologi för informationssökning

tisdag 24 augusti 2010

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. För att representera ett dokument görs ofta olika former av bearbetningar såsom stemming eller eliminering av stoppord. Förklara vad dessa båda bearbetningar innebär och motivera varför de används.

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

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

4. Ange ett reguljärt uttryck som känner igen prisangivelser upp till 10,000:-. Uttrycket ska kunna hantera öresangivelser med kolon som i 89:50 eller 25:- och tusental med eller utan kommatecken som i 5,500 eller 5500. Däremot kan du anta att det inte ingår något blanktecken i uttrycket.

5. Ange med ledning av frekvensangivelserna i tabellen nedan Maximum Likelihood-uppskattningar av (a) bigramsannolikheten p(inte|kan), (b) trigramsannolikheten p(komma|kan inte). c) Ange också en nackdel med Maximum Likelihood-uppskattningar.

inte	300
kan	250
komma	200
kan inte	40
inte kan	2
inte komma	5
kan inte komma	4
komma kan inte	1

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

Institutionen för
datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

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

7. Förklara begreppen hyponymi och antonymi med exempel.

8. Låt w_1^N stå för en sekvens av N ord w_1, w_2, ... w_N och $t_1^N = t_1, t_2 \ldots, t_N$ vara en motsvarande sekvens av ordklasstaggar. Med en sannolikhetsbaserad modell för ordklasstaggning anges den bästa taggningen t* med formeln

$$t^* = \underset{t_1^N}{\operatorname{argmax}}\, p(t_1^N \mid w_1^N)$$

(a) Uttryck ovanstående formel i ord. (b) Skriv om den med hjälp av Bayes regel som produkten av två sannolikheter, (c) Ge något skäl varför produkten $\Pi\, p(t_i|w_i)$ med $1 \leq i \leq N$ ger en dålig approximation av t*.

9. Antag att vi ur texter vill extrahera så många korrekta par <A, B> som möjligt, där A står för ett företag och B för företagets vd. Exempel: Om en text innehåller meningen "Olle Olsson, vd på Informatia AB, kandiderar för en riksdagsplats." vill vi extrahera paret <Informatia AB, Olle Olsson>. Beskriv hur man kan använda en s.k. bootstrapping-metod för att göra detta.

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