Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2013-03-25
Sal	TER4
Tid	14-18
Kurskod	TDDD02
Provkod	TEN1
Kursnamn/benämning	Språkteknologi för informationssökning
Institution	IDA
Antal uppgifter som ingår i tentamen	*11*
Antal sidor på tentamen (inkl. försättsbladet)	*3*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*013-282422*
Besöker salen ca kl.	*Nej, (använd telefon om det finns behov)*
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

måndag 25 april 2013 kl. 14-18

Inga hjälpmedel är tillåtna.

Besvara alla frågor. Varje fråga är värd 3 poäng. Maximal poäng är 33. Planerade betygsgränser är 16,5 (betyg 3), 22 (betyg 4) och 27 (betyg 5). Det går bra att besvara flera frågor på samma papper.

1. (a) Förklara vad som menas med stoppord och ge några svenska exempel på vanliga stoppord. (b) Ange något sätt, att givet en dokumentmängd identifiera lämpliga stoppord.

2. Vilka är de centrala komponenterna i en sökmotor?

3. (a) Ange ett reguljärt uttryck som matchar orden i mängden {ro, ros, rosa, rost, rosta} men inga andra teckensekvenser, (b) Ange en tillståndsautomat (FSA) som motsvarar det reguljära uttrycket, (c) Utvidga automaten så att ordet 'tro', men inga andra nya teckensekvenser, också kan matchas.

4. (a) Vad menas i sannolikhetsläran med att två händelser är oberoende? (b) Det finns modeller som något oegentligt antar att alla intressanta händelser är oberoende; beskriv en sådan modell och tala om varför antagandet om oberoende är väsentligt.

5. Antag att vi i en korpus som omfattar 100 000 ord hittar ordet *vi* 120 gånger, ordet *har* 1500 gånger, sekvensen *vi har*, 30 gånger och sekvensen *har vi* 20 gånger. Vad är Maximum Likelihood-uppskattningen av (i) sannolikheten för att *vi* kommer efter *har* utifrån denna korpus, (ii) unigramsannolikheten för ordet *har*?

6. Anta att namnet Lundbergs förekommer i en text med syftning ibland på en person ibland på ett företag. Hur skulle vi kunna skilja de olika fallen åt?

7. System för informationssökning utvärderas vanligen med både precision och recall. Definiera dessa mått och ange därutöver något mått som väger ihop precision och recall.

8. Ange Levenshteinavståndet mellan orden *stort* och *strut* 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
lars.ahrenberg@liu.se

9. (a) Ange tre olika kriterier som kännetecknar en nominalfras som är koreferent med ett anaforiskt pronomen. (b) Ange någon språkteknologisk tillämpning där koreferensresolution är en väsentlig funktion.

10. I många system är man intresserad av specifika relationer mellan olika entiteter, t.ex. mellan personer och deras födelseår eller släktskap. Ett problem är då att de relationer man behöver modellera uttrycks på många olika sätt i naturlig text. En metod som har föreslagits för att hitta en stor mängd uttryckssätt för en given relation är bootstrapping. Beskriv denna metod.

11. 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.