Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2011-04-29
Sal	U6, Hus C
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	*2422*
Besöker salen ca kl.	*14.40*
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

fredag 29 april 2011 kl. 14-18

Inga hjälpmedel är tillåtna.

Varje fråga är värd 3 poäng. Maximal poäng är 33. 16,5 poäng ger säkert godkänt. 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. 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 <0,0,1,1,0,0,0,0>. Ordna dokumenten efter relevans för sökfrågan. Motivera din rangordning.

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

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

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

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

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

5. I en text på sammanlagt 10000 ord finns 800 som börjar med stor bokstav. 500 av dessa inleder en mening och 600 är egennamn. Det finns inga meningar och inga egennamn som inte börjar med stor bokstav. Det finns 150 namnbigram, d.v.s. båda de ingående orden är egennamn. Använd Maximum Likelihood-uppskattning för att uppskatta följande sannolikheter: (a) sannolikheten för att ett ord som börjar med stor bokstav är ett egennamn, (b) sannolikheten för att ett egennamn följs av ett egennamn, (c) sannolikheten för att ett ord som inte står först börjar med stor bokstav.

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.

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

lars.ahrenberg@liu.se

7. Förklara modellen Naive Bayes för att klassificera ord eller dokument. Det går bra att göra det generellt eller utifrån ett konkret klassificeringsproblem med konkret angivna indikatorer.

8. Ange, med exempel, tre vanliga sätt att i text referera till en entitet som i den föregående texten introducerats med ett fullständigt namn.

9. I informationsextraktionssystem är man ofta intresserad av specifika entiteter och relationer mellan dem, t.ex. mellan personer och deras födelseår. Ett problem är då att de relationer man behöver modellera uttrycks på många olika sätt i naturlig text, ofta på sätt som är svåra att komma på, även för experter. En metod som har föreslagits för att hitta en stor mängd uttryckssätt för en given relation är bootstrapping. Beskriv vad denna metod går ut på.

10. Skriv ett reguljärt uttryck som identifierar fyrsiffriga tal, dvs naturliga tal i intervallet 1000-9999. Skriv också ett reguljärt uttryck som begränsar träffarna till tal i intervallet 1000-2499.

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.