Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2014-04-23
Sal	T1, Hus C, Campus Valla
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.	*14.45*
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 23 april 2014 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. 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 <0,0,0,1,1,0,0,0>. Ordna dokumenten efter relevans för sökfrågan. Motivera din rangordning.

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

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

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

2. Ange för var och en av mängderna L1 och L2 ett reguljärt uttryck som matchar dem exakt. Rita också upp en automat (FSA) som matchar mängden L2.

   L1: {a, ab, abb, abbb, abbbb, ... }

   L2: {b, c, bb, cc, bbb, ccc, ... }

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 *stoft* och *kofta* och visa hur man räknar fram det algoritmiskt, t.ex. genom att ställa upp en matris för avstånden mellan delsträngar.

5. (a) Förklara vad som menas med en språkmodell (eng. *language model*). (b) Vad innebär det att en språkmodell använder trigram? (c) Vad innebär det att en språkmodell använder back-off?

6. 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.

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

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

8. 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.

9. Frågebesvarande system utvecklas för att kunna ge svar på godtyckliga frågor. Beskriv vilken information ett sådant system kan få från en fråga som *När började första världskriget?* och hur denna information används för att hitta möjliga svar.

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.

11. Förklara vad som menas med hyponymi och ange något sätt att identifiera hyponymer i löpande text.