Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2012-12-19
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.	*14.40*
Kursadministratör (namn + tfnnr + mailadress)	*Helene Meisinger 281868, helene.meisinger@liu.se*
Tillåtna hjälpmedel	*Inga 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	

TENTAMEN

TDDD02 Språkteknologi för informationssökning

onsdag 19 december 2012 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) Vad menas med lemmatisering? (b) Hur skiljer sig stemming från lemmatisering? (c) Vad är värdet av sådana operationer för dokumentsökning?

2. Antag att vi representerar dokument i en vektorrymd med hjälp av tio förvalda termer och deras frekvens i dokumenten. Tre dokument är givna med sina representationer enligt nedan. En given sökfråga fick representationen <0,0,1,1,0,0,0,0,0,0>. Ordna dokumenten efter relevans för sökfrågan. Motivera din rangordning.

   D1: <2,0,1,3,0,1,0,3,1,0>

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

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

3. System för namnigenkänning utvärderas vanligen med både precision och recall. Definiera dessa mått och ange minst två faktorer som gör att det är svårt att nå 100% för en namnigenkännare.

4. Anta att vi i givna svenska texter vill hitta två ord i följd som börjar på bokstaven D eller d, t.ex. *dagen D*, eller *de dygdiga*. (a) Skriv ett reguljärt uttryck som kan ges som argument till någon **grep**-funktion för detta ändamål. (b) Antag sedan att vi vill generalisera till godtyckligt långa sekvenser av ord som börjar på D eller d, t.ex. *din dumma dromedar*. Ange ett reguljärt uttryck som åstadkommer detta. (c) Beskriv hur uttrycket måste modifieras om vi vill generalisera till tvåordssekvenser som börjar på samma bokstav, godtyckligt vilken.

5. Om man skriver in ett felskrivet sökord i en sökmotor, t.ex. *julgrt*, kan man ofta få ett sökresultat för ett alternativt sökord, så säger exempelvis Google i detta fall "Visar resultat för *julgröt*". Beskriv någon metod från kursen som kan tillämpas på problemet att gissa vad användaren menar.

6. 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

LINKÖPINGS UNIVERSITET

lars.ahrenberg@liu.se

7. (a) Skriv ett uttryck som uppskattar trigramsannolikheten p(*annan* | *en eller*) med hjälp av bigramsannolikheter för de ingående orden; (b) Kommer uppskattningen att bli lägre eller högre än den korrekta trigramsannolikheten? Motivera ditt svar.

8. Utjämning (eng. *smoothing*) är en väsentlig komponent när man gör statistiska språkmodeller. Förklara varför utjämning behövs och beskriv kortfattat någon metod att göra det.

9. I informationsextraktionssystem är man ofta intresserad av specifika entiteter och relationer mellan dem, t.ex. mellan företag och deras vd:ar. 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. I artikeln "An Analysis of the AskMSR Question-Answering System" är författarnas grundidé att utnyttja vad de kallar *webbens redundans*. (a) Vad menas med det? (b) Beskriv översiktligt arkitekturen i systemet AskMSR.

11. Textsammanfattningssystem baseras ofta på extraktion av en delmängd av de meningar som ingår i texten. (a) Ange tre vanliga, beräkningsbara indikatorer på att en mening är lämplig att ta med i en sammanfattning. (b) Förklara det s.k. ROUGE-måttet för att utvärdera textsammanfattningssystem.