Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2011-12-15
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	*013282422*
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	

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

TDDD02 Språkteknologi för informationssökning

torsdag 15 december 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. I dokumentsökningssystem används ofta någon form av textnormalisering. Förklara varför man använder textnormalisering och ge två olika konkreta exempel på normaliseringsoperationer.

2. En substitution är defierad i UNIX-programmet sed på följande sätt:

```
s/\([\.|\?|,]\)/ \1/g;
```

Vad blir resultatet om denna substitution appliceras på textraden

```
Var fanns Ekman? Vart hade han tagit vägen?
```

3. (a) Ange tre strängar som matchar det reguljära uttrycket a*cb+. (b) Definiera en ändlig tillståndsautomat som definierar samma språk som detta reguljära uttryck.

4. (a) Förklara vad som menas med en språkmodell. (b) Två olika språkmodeller, A och B, jämfördes på testdata som tagits fram för en given domän. Modellen A hade en perplexitet på 105 medan B hade 380. Vilken modell bör man välja och varför?

5. Ange ett uttryck för sannolikheten av meningen *snart är det vår* givet att vi känner till bigramsannolikheterna för orden i meningen och sannolikheten för att ordet *snart* kommer först i en mening.

6. Ett namnigenkänningssystem testades med följande resultat. Av de 200 namnen i testkorpusen lyckades systemet hitta 180. Det föreslog 90 andra ord felaktigt som namn. Beräkna följande mått på systemets prestation som namnigenkännare, eller förklara varför de inte går att beräkna: (a) precision, (b) recall, (c) korrekthet.

7. Två komponenter i informationsutvinningssystem är koreferenlösning och relationsbestämning. Förklara vad dessa komponenter gör.

8. Det första steget i ett typiskt frågebesvarande system är frågeanalys. Vad är utdata från frågeanalysen och hur används dessa utdata i andra delar av systemet?

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

Tel: 013-282422
e-post: lah@ida.liu.se

9. Ett system för ämnesklassificering av inkommande texter från en nyhetsbyrå baseras på ordlistor med ämnesord för tio olika ämnen. För att avgöra om en text handlar om ett visst ämne, t.ex. ekonomi, eller inte, används följande indikatorer: 1. rubriken innehåller minst ett ämnesord, 2. den första meningen innehåller minst ett ämnesord, 3. minst 5% av alla ord i texten, oräknat stoppord, är ämnesord. Som beslutsmetod används Naive Bayes.

   (a) Inför lämpliga beteckningar för ämnet och för de tre indikatorerna och formulera beslutsregeln matematiskt.
   (b) Varför kallas Naive Bayes för 'naive' och vad innebär det för beslutsregeln?
   (c) Vad står apriori-sannolikheten för i detta fall och hur kan den lämpligen uppskattas?

10. Antag att vi underhåller en FAQ med hundratals frågor och vill tillåta användare att navigera i FAQ:n genom att formulera godtyckliga frågor i ett dialogfönster. Beskriv någon metod som kursen tagit upp som kan returnera den bäst matchande frågan från FAQ:n med sitt svar. För full poäng måste du ange (i) hur du vill representera frågor och svar, (ii) vad användarfrågan matchas mot, och (iii) hur man avgör vilken som är den bästa matchningen.

11. Om man skriver in sökordet **månbenet** i Google får man, förutom ett antal träffar, också responsen Menade du: **netnet**? (a) Vad är redigeringsavståndet mellan sökord och förslag i detta fall? (b) Motivera ditt svar på (a), (c) Hur kan man förklara att Google ger förslag med så pass stort avstånd till sökordet?