Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2010-03-10*
Sal	*TER2*
Tid	*14-18*
Kurskod	*TDDD01*
Provkod	*TEN1*
Kursnamn/benämning	*Språkteknologi*
Institution	*IDA*
Antal uppgifter som ingår i tentamen	*11*
Antal sidor på tentamen (inkl. försättsbladet)	*4*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*ankn 2422, mobil 0703182422*
Besöker salen ca kl.	*kl 15*
Kursadministratör (namn + tfnnr + mailadress)	*Anna Grabska Eklund Ankn. 23 62, annek@ida.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

TDDD01 Språkteknologi

onsdag 10 mars 2010 kl. 14-18

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 26. 13 poäng ger säkert godkänt (betyg 3).

Besvara alla frågor. Det går bra att besvara flera frågor på samma papper.

1. Förklara vad som menas med *tokenisering* resp. *normalisering* av en textkorpus. Ge också konkreta exempel på hur dessa processer kan förändra originaltexten. (2p)

2. System för ordklasstaggning utnyttjar ords lokala kontext för att välja rätt tagg. Ange ett vanligt regelformat för sådan disambiguering och formulera någon eller några explicita regler som kan appliceras på ordet *skruvar* i dessa två meningar: (2p)

   Hon skruvar bollen i mål.

   Här hittar du specialtillverkade skruvar och bultar.

3. (a)Ange en ändlig automat eller transduktor som mappar om orden *skriva, skriver, skrev* på följande sätt:

   skriva → skriv V INF

   skriver → skriv V PRES

   skrev → skriv V PAST

   (b) Hur kan samma automat/transduktor utvidgas för att också hantera motsvarande böjningsformer för verbet *riva*? (3p)

4. Vad menas med redigeringsavstånd (Minimal edi distance, eller Levenshtein distance)? Hur räknar man ut redigeringsavståndet och hur kan man använda det i samband med stavningskontroll? Illustrera med paret *jäna ~ gärna.* (3p)

5. Många metoder för parsning av naturligt språk använder en datastruktur kallad chart. Vad är syftet med charten och vad för slags data lagras där? Svaret bör i detalj ange vad för information som finns i ett charttillstånd, gärna med exempel. (3p)

6. Vad menas med partiell parsning (eller chunkning)? Varför föredras partiell parsning framför fullständig parsning i många applikationssystem som t.ex. informationsutvinning? (2p)

7. Två system för namnigenkänning jämfördes med avseende på deras förmåga att känna igen och klassificera namn av olika slag. Avgör med hjälp av nedanstående tabell (a) Vilket system har bäst precision vad gäller igenkänning av personnamn? (b) Vilket system har bäst recall i fråga om namnigenkänning generellt? Du ska visa hur du räknat. (2p)

	System A		System B		Antal i facit
	Funna	Korrekta	Funna	Korrekta	
Personnamn	120	90	50	48	100
Organisationsnamn	85	60	40	32	80
Övriga namn	12	2	5	5	20

8. I en textkorpus hittades följande frekvenser för några vanliga ord och deras bigram och trigram:

du	400
och	2000
jag	500
du och	32
och jag	20
du och jag	4
jag och	48
och du	12
jag och du	2

(a) Vad är Maximum Likelihood uppskattningen av bigramsannolikheten p(och | du)?

(b) Vad är Maximum Likelihood uppskattningen av trigramsannolikheten p(jag|du och)?

(c) Vad blir sannolikheten för ordsekvensen 'du och jag' om den uppskattas med sannolikheter för de ingående bigrammen? Det räcker här att ange ett korrekt uttryck. (3p)

9. Vad är en statistisk språkmodell och vilken roll har sådana språkmodeller i statistisk maskinöversättning? (2p)

10. Relationer, som den "att vara gift med någon" kan uttryckas på många sätt i naturligt språk, i exemplet med ord och fraser som *gifte sig, var gift med, X's make, Y's hustru, ...* Beskriv en metod att hitta en stor mängd sådana relationsangivande uttryck i en större textkorpus. (2p)

11. I Lesks algoritm för bestämning av ordbetydelser används definitioner och exempel som hämtas från ordböcker. Beskriv kortfattat hur den metoden fungerar. För att illustrera metoden kan du använda meningen *Krakel Spektakel hängde och slängde i en gardin* och nedanstående data om ordet *hänga* hämtat från Svensk Ordbok (här i förkortad och något omskriven form):

    1. hållas uppe (mot tyngdkraften) endast genom att vara fästad i sin övre del vid t.ex. krok eller stång: *tavlan hänger över bokhyllan, i fönstret hängde dyrbara gardiner...*

    2. anbringa (ngt) i fäste ett stycke ovanför golvet eller marken så att föremålets nedre del är fri i förhållande till underlaget: *hon hängde tavlan över bokhyllan, han hängde ut kläderna på tork...*

    3. avliva (ngn) med hjälp av uppfästad snara e. d.; vanligen men ej nödvändigtvis efter dom: *mördaren hängdes i gryningen, folkhopen ville hänga hästtjuven i närmaste träd,...*

    (2p)