Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2010-06-08*
Sal	*VALMAT*
Tid	*14-18*
Kurskod	*TDDD01*
Provkod	*TEN1*
Kursnamn/benämning	*Språkteknologi*
Institution	*IDA*
Antal uppgifter som ingår i tentamen	*12*
Antal sidor på tentamen (inkl. försättsbladet)	*3*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*/2422 0703182422*
Besöker salen ca kl.	*14.45*
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

tisdag 8 juni 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. Vad är resultatet av att lemmatisera meningen *Alla människor är i grunden goda*?(2p)

2. När sammansättningar som har samma slutled samordnas i en text kortas ofta det första ordet och förses med ett bindestreck: *för- och nackdelar, jord- och skogsegendom*. (a) Skriv ett reguljärt uttryck som kan känna igen sådana samordningar i löpande text. (Du kan förutsätta att det inte finns någon mellanliggande radbrytning.) (2p)

3. I en korpus hittades 100 ord som slutar på 'het', bl.a. nyhet och glödhet. Av dessa befanns 60 vara substantiv, medan 40 var adjektiv. (a) Ange MLE-uppskattningen med användningen av denna korpus för sannolikheten att ett ord som slutar på 'het' är ett adjektiv; (b) Hur kan vi ändra denna uppskattning om vi vill ge utrymme för möjligheten att egennamn också skulle kunna sluta på 'het', trots att inga sådana egennamn setts i korpusen? (2p)

4. Ange en ändlig automat (FSA) som genererar samma språk som det reguljära uttrycket ba^+b^*. (2p)

5. Ett rättstavningsprogram kritiserades för att generera för många 'falska positiva'. (a) Vad betyder det? (b) Vilket eller vilka av måtten precision eller recall påverkas av antalet falska positiva? (2p)

6. Vid ordprediktion är man intresserad av sannolikheten $p(w_i| w_1 w_2 \ldots w_{i-1})$. (a) Vad innebär det att använda en bigrammodell för denna sannolikhet? (b) Vilket är det vanligaste måttet för att mäta prestanda för en ordprediktionsmodell? Ange namn och definition. (3p)

7. I modellen Den brusiga kanalen (eng. Noisy channel) förekommer en s.k. a priori-sannolikhet. (a) Vad representerar denna sannolikhet? (b) Hur uppskattas den vanligen om vår tillämpning är maskinöversättning? (2p)

Institutionen för datavetenskap
Linköpings universitet
LINKÖPINGS UNIVERSITET
e-post: lah@ida.liu.se

8. I ett visst läge i en chart-parser finns bland annat följande tillstånd/bågar:

   S → . NP VP [0,0]

   NP → DET . AP N [0,1]

   Ange två andra tillstånd/bågar som också måste finnas i charten. Motivera ditt svar. (2p)

9. Vad menas med redigeringsavstånd (Minimal edi distance, eller Levenshtein distance)? (2p)

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

11. (a) Förklara begreppen anafor och antecedent med exempel. (b) Varför är det viktigt att kunna känna igen anaforer och antecedenter exempelvis i en tillämpning som informationsextraktion? (3p)

12. Relationer, som den mellan en vara och ett pris kan uttryckas på många sätt i naturligt språk, t.ex med uttryck som *kosta, betinga ett pris, priset för ... är* . Beskriv en metod att hitta en stor mängd sådana relationsangivande uttryck i en större textkorpus. (2p)