Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	12-05-15
Sal	TER 4
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)	*3*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*013-282422*
Besöker salen ca kl.	-
Kursadministratör (namn + tfnnr + mailadress)	*Åsa Kärrman 281868, asa.karrman@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

måndag 21 maj 2012 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.

Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla frågor i denna del.

1. (a) Skriv ett reguljärt uttryck som identifierar ordsekvenser eller rader som innehåller ordet 'ett', följt av ett godtyckligt ord, följt av ordet 'hus'. (b) Skriv också en reguljär substitution som extraherar ett godtyckligt ord mellan 'ett' och 'hus'. (2p)

2. I förväxlingsmatrisen nedan redovisas utdata från ett taggningssystem i form av absoluta frekvenser för korrekta och felaktiga kategoriseringar av ett antal ordformer i svenska. Rader anger systemets kategoriseringar, kolumner kategorierna i facit. Siffran 16 i raden Verb anger alltså att i 16 fall taggade systemet nomen som verb. (a) Vad är systemets precision för Verb? (b) Vilken ordklass har systemet svårast att tagga rätt? Motivera svaret! (3p)

	Nomen	Verb	Adjektiv	Adverb	Andra ordklasser
Nomen	600	14	6	10	56
Verb	16	455	2	6	1
Adjektiv	5	10	40	15	20
Adverb	0	2	18	50	18
Andra ordklasser	20	9	10	29	2400

3. I modellen 'den brusiga kanalen' (eng. *Noisy Channel*) ser formeln för att välja mest sannolika sträng ut så här:

$$S^* = \underset{S}{\operatorname{argmax}}\, p(S|O)$$

(a) Förklara formeln i ord. (b) Ofta skriver man om sannolikheten p(S|O) med hjälp av Bayes regel. Visa resultatet av omskrivningen. (2p)

4. Ändliga automater eller transduktorer har användning inom språkteknologin t.ex. för att lagra lexikon och få ord i löpande text morfologiskt analyserade. Ange minst två fördelar med denna typ av representation, jämfört med att ha orden med deras analyser lagrade i en textfil. (2p)

5. Förklara vad som menas med redigeringsavstånd och beskriv någon algoritm för att beräkna redigeringsavståndet mellan två ord. Tillämpa algoritmen på exempelorden *slott* och *koja* och beräkna deras redigeringsavstånd (3p)

6. I en större svensk korpus får vi fram följande frekvenser för några utvalda ord och ordpar:

   huller om: 71
   om buller: 73
   huller om buller: 71
   huller: 71
   buller: 402
   om: 175,580

   Vad blir den uppskattade sannolikheten för p(buller | huller om), om vi (a) använder trigramsannolikheter, (b) bigramsannolikheter? (2p)

7. (a) Vad innebär chunkning? (b) Hur kan chunkning hanteras som ett taggningsproblem? (2p)

8. Vid s.k. top-down chart-parsning av kontextfri grammatik används tre grundläggande operationer för att generera och bekräfta hypoteser: predicering, scanning och kombinering (eng. *completion*). Definiera dessa operationer eller beskriv vad de innebär. (3p)

9. Olika frågor är olika svåra för automatiska frågebesvarande system att hitta rätt svar på. Rangordna följande tre frågor från lättast till svårat för ett frågebesvarande system och motivera rangordningen. Systemet antas inte ha någon egen kunskapsbas utan använder sig av textdokument som dataresurs. (2p)

   Varför sjönk Titanic?
   Vilket år sjönk Titanic?
   När började Titanic spelas in?

10. Vad menas med typad precision (eng. *labelled precision*) och i vilket sammanhang kan det vara användbart? (2p)

11. Utjämning (eng. *smoothing*) är en teknik som används när man bygger statistiska språkmodeller (eng. *language models*). Förklara vad det innebär, varför man använder det och ge exempel på någon metod. (3p)