Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	2014-03-17
Sal	G33
Tid	8–12
Kurskod	TDDD01
Provkod	TEN1
Kursnamn/benämning	Språkteknologi
Institution	*IDA*
Antal uppgifter som ingår i tentamen	9 + 2
Antal sidor på tentamen (inkl. försättsbladet)	4
Jour/Kursansvarig	Marco Kuhlmann
Telefon under skrivtid	4644
Besöker salen ca kl.	09:00
Kursadministratör (namn + tfnnr + mailadress)	Åsa Kärrman 013-285760, 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	

TDDD01 **Språkteknologi**

Tentamen

Marco Kuhlmann, Institutionen för datavetenskap, Linköpings universitet
marco.kuhlmann@liu.se

17 mars 2014

Inga hjälpmedel är tillåtna. Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentamen är 24; 12 poäng ger säkert godkänt. Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla frågor i denna del. Varje fråga ger 2 poäng.

1. Vid språkteknologiskt korpusarbete tokeniserar man helst texter så att skiljetecken som komma och punkt utgör egna token. (a) Varför gör man det? (b) Varför resonerar man annorlunda när det gäller förkortningspunkter?

2. I en svensk korpus finner vi följande frekvenser för några utvalda ord och ordsekvenser:

   med: 32 900; *tanke*: 400; *på*: 24 500; *med tanke*: 260; *tanke på*: 270; *med tanke på*: 250

   Vad är den Maximum Likelihood-uppskattade sannolikheten $P(\text{på} \mid \text{med tanke})$ om vi använder (a) trigramsannolikheter, (b) en omskrivning till bigramsannolikheter?

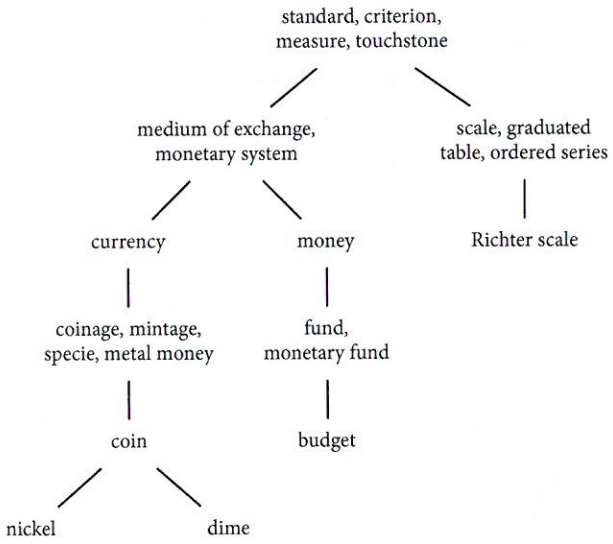
3. (a) Vilka typer av sannolikheter ingår i en Hidden Markov-modell för ordklasstaggning? (b) Vilka konkreta sannolikheter måste man ha skattat för att i en sådan modell kunna räkna ut den kombinerade sannolikheten för följande taggade mening?

   Jag/PN äter/VB

4. Nedanstående tabell visar reglerna i en probabilistisk kontextfri grammatik. Rita två olika frasstrukturträd (parseträd) enligt denna grammatik och ange deras sannolikheter.

S → NP VP	1,00
NP → *Lotta*	0,25
NP → *cykeln*	0,75
VP → V NP	1,00
V → *lånar*	1,00

5. Nedanstående graf visar en liten del av WordNet. (a) Förklara vad noderna och bågarna representerar. (b) Bestäm avståndet (eng. *pathlength*) mellan "nickel" och "budget" och utifrån detta avstånd räkna ut den semantiska likheten mellan de två orden.



6. Ett automatiskt system för författaridentifikation ska hitta texter som är skrivna av författaren A. En utvärdering av systemet på en guldstandard ger följande resultat, där + betyder att texten är skriven av A och – betyder att texten inte är skriven av A. Räkna ut systemets precision och recall.

	system +	system –
guldstandard +	1	3
guldstandard –	0	19

7. En enkel metod för att tagga filmrecensioner med polariteter är att använda en Naive Bayes-klassificerare. Ange klassificerarens beslutsregel och förklara den.

8. (a) Vad innebär entitetsextraktion (eng. *named entity recognition*)? (b) Hur kan entitetsextraktion hanteras som ett taggningsproblem?

9. En central modul i ett frågebesvarande system är en analysator som bestämmer frågans svarstyp. Förklara vad som menas med en svarstyp och ge några exempel på möjliga svarstyper.

Del B

Välj en fråga och besvara den utförligt. Varje fråga kan ge maximalt 6 poäng.

1. En probabilistisk parser ska räkna ut den mest sannolika syntaktiska analysen för en given mening. Förklara varför denna uppgift är beräkningsmässigt utmanande. Beskriv därefter två metoder för att bemöta denna utmaning.
2. Det finns för närvarande ett stort intresse inom både akademin och industrin i metoder för att analysera språk i sociala medier. Ange några skäl till detta. Diskutera därefter några av de utmaningar som språkteknologin ställs inför när den ska tillämpas på texter från Twitter och Facebook snarare än t.ex. tidningar och lexikon.