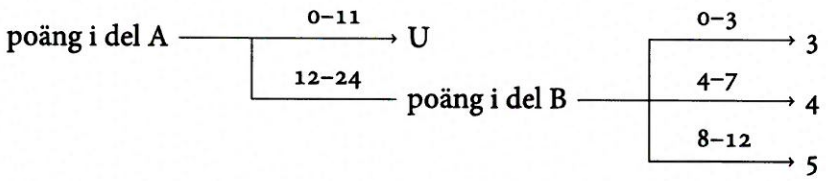
Försättsblad till skriftlig tentamen vid Linköpings universitet

Datum för tentamen	2016-06-07
Sal (1)	TER2
Tid	14-18
Kurskod	TDDD01
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Språkteknologi Skriftlig tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	12
Jour/Kursansvarig Ange vem som besöker salen	Marco Kuhlmann
Telefon under skrivtiden	013-284644
Besöker salen ca klockan	Besöker salen ca. kl 15. Finns tillgänglig via telefon under hela skrivningstiden.
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Elin Brödje, 013-284767, elin.brodje@liu.se
Tillåtna hjälpmedel	inga
Övrigt	-/-
Antal exemplar i påsen	

Tentamen 2016-06-07

Marco Kuhlmann

Tentamen består två delar, A och B. Varje del omfattar ett antal frågor à 3 poäng. **Del A** omfattar 8 frågor som kan besvaras kortfattat. Det krävs minst 12 poäng på denna del för att del B ska rättas. **Del B** omfattar 4 frågor som kräver utförliga svar med korrekt terminologi. Betyget sätts enligt följande:



Lycka till!

Del A

01 **Korrekthet, precision och täckning (recall).**

Vid utvärderingen av en ordklasstaggare fick man ut nedanstående förväxlingsmatris. Den markerade cellen (rad `NN`, kolumn `JJ`) anger antalet gånger systemet klassade ett ord som adjektiv (`JJ`) medan det enligt guldstandarden var ett substantiv (`NN`).

	NN	JJ	VB
NN	58	6	1
JJ	5	11	2
VB	0	7	43

a) Ställ upp ett bråk för taggarens precision på substantiv.

b) Ställ upp ett bråk för taggarens täckning (recall) på verb.

c) Ange en annan förväxlingsmatris där taggarens korrekthet är samma som i matrisen ovan men där respektive värden för a) och b) är 0%.

02 Textklassificering

Ett system för textklassificering baserat på metoden Naive Bayes ska avgöra om dokumentet "Tokyo Tokyo Peking" är en nyhet om Japan (klass J) eller en nyhet om Kina (klass K). Systemet använder vokabulären V = {Tokyo, Peking, Seoul} och bl.a. följande sannolikheter

$P(\mathsf{J}) = 3/4$	$P(\text{Tokyo} \mid \mathsf{J}) = 4/6$	$P(\text{Peking} \mid \mathsf{J}) = 1/6$
$P(\mathsf{K}) = 1/4$	$P(\text{Tokyo} \mid \mathsf{K}) = 1/2$	$P(\text{Peking} \mid \mathsf{K}) = 1/2$

a) Ange de två sannolikheter som saknas, $P(\text{Seoul} \mid \mathsf{J})$ och $P(\text{Seoul} \mid \mathsf{K})$.

b) Utifrån de angivna sannolikheterna, vilken klass predicerar systemet? Redovisa hur du räknat. Visa tydligt att du förstått Naive Bayes-klassificeringsregeln.

c) Ange en dokumentsamling från vilken man får de angivna sannolikheterna om man skattar med Maximum Likelihood-metoden.

03 Ordpredicering

I en text innehållande 1 215 396 ord och 105 436 unika ord hittas ordet *det* 13 694 gånger, ordet *är* 13 700 gånger, ordet *nalkas* 2 gånger, bigrammet *det är* 927 gånger och bigrammet *det nalkas* 0 gånger.

a) Skatta unigramsannolikheten $P(det)$ och bigramsannolikheten $P(är \mid det)$ med Maximum Likelihood-metoden. Ställ upp bråk.

b) Vad händer när man skattar bigramsannolikheten $P(nalkas \mid det)$ med Maximum Likelihood-metoden? Varför kan detta vara ett problem?

c) Skatta bigramsannolikheten $P(nalkas \mid det)$ med Maximum Likelihood-metoden och addera-k-utjämning med $k = 0{,}05$ (**inte** addera-1). Antag att vokabulären består av mängden av alla unika ord. Ställ upp bråk.

04 Ordklasstaggning

Följande matriser specificerar en Hidden Markov-modell. Istället för sannolikheter anges kostnader (negativa log-sannolikheter).

	PL	PN	PP	VB	EOS
BOS	11	2	3	4	19
PL	17	3	2	5	7
PN	5	4	3	1	8
PP	12	4	6	7	9
VB	3	2	3	3	7

	hen	vilar	ut
PL	17	17	4
PN	3	19	19
PP	19	19	3
VB	19	8	19

När man använder Viterbi-algoritmen för att beräkna den mest sannolika (minst kostsamma) taggsekvensen för meningen "hen vilar ut" enligt denna modell får man ut följande matris. Notera att matrisen saknar tre värden (markerade celler).

	hen	vilar	ut
BOS	0		
PL	A	B	21
PN	5	28	35
PP	22	27	20
VB	23	14	36
EOS			C

a) Beräkna värdet A. Redovisa hur du räknat.

b) Beräkna värdena B och C. Redovisa hur du räknat.

c) Ange den mest sannolika taggsekvensen för meningen. Förklara hur den kan fås från (den fullständiga) matrisen.

05 Syntaktisk analys

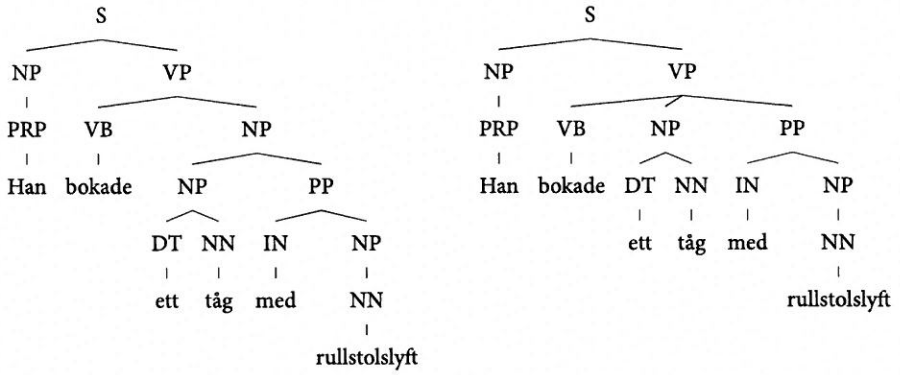
Nedan anges alla NP-regler och alla VP-regler i en viss probabilistisk kontextfri grammatik. En av NP-reglerna saknar ett sannolikhetsvärde.

$$\text{NP} \to \text{PRP}\ \tfrac{2}{7} \qquad \text{NP} \to \text{NP PP}\ ? \qquad \text{NP} \to \text{DT NN}\ \tfrac{2}{7} \qquad \text{NP} \to \text{NN}\ \tfrac{2}{7}$$

$$\text{VP} \to \text{VB NP}\ \tfrac{1}{2} \qquad \text{VP} \to \text{VB NP PP}\ \tfrac{1}{2}$$

a) Vilket sannolikhetsvärde saknas?

b) Nedan anges två träd som genererats av grammatiken. Ställ upp bråk för deras sannolikhetsvärden. Antag att alla regler som inte anges ovan har sannolikhet 1.



c) Hur måste man ändra sannolikheterna för VP-reglerna så att det vänstra trädet får högre sannolikhet än det högra? (Du behöver inte ange de exakta värdena.)

06 Semantisk analys

Betrakta följande (normaliserade) dokumentsamling:

(1) automobile wheel motor vehicle transport passenger

(2) car form transport wheel capacity carry five passenger

(3) transport London game spectator advise avoid use car

(4) London soccer tournament begin goal match

(5) Giggs score goal football tournament Wembley London

(6) Bellamy passenger football match play part goal

a) Komplettera följande term–term matris utifrån dokumentsamlingen.

	wheel	transport	goal	match
automobile	☐	☐	☐	☐
car	☐	☐	☐	☐
soccer	☐	☐	☐	☐
football	☐	☐	☐	☐

b) Rita in målorden som vektorer i ett koordinatsystem där x-axeln svarar mot det totala antalet förekomster tillsammans med ett av kontextorden *wheel* eller *transport* och y-axeln svarar mot det totala antalet förekomster tillsammans med ett av kontextorden *goal* eller *match*.

c) Förklara hur man med hjälp av sådana vektorrepresentationer kan mäta likhet mellan målorden. Vilka resultat skulle denna metod ge för de angivna målorden?

07 Informationsextraktion

Ett system för informationsextraktion är tränat på att hitta tre typer av namngivna entiteter: personer (PER), organisationer (ORG) och svenska tätorter (LOC).

a) Vilken av dessa tre typer är lättast att hitta med hjälp av namnlistor?

b) Entitetsextraktion kan ses som uppgiften att tagga varje token i en mening med en så kallad BIO-tagg. Sätt ut BIO-taggarna för följande mening. (Observera att meningen består av 20 stycken token.)

Astrid Lindgren , född den 14 november 1907 i Vimmerby , utsågs till hedersdoktor vid Linköpings universitet år 2000 .

c) System som använder BIO-taggningen kan utvärderas på taggnivå eller entitetsnivå. Ändra en av dina taggar så att den nya taggningen har 95% korrekthet på taggnivå men 0% precision och recall med avseende på entitetstypen ORG. Använd din ursprungliga taggning som guldstandard.

08 Frågebesvarande system

Följande frågor ställdes till ett frågebesvarande system:

A Vilket datum förliste *S/S Per Brahe*?

B När föddes den konstnär som drunknade då S/S Per Brahe förliste?

C Vad hette denna konstnärs hustru?

a) Ange svarstyper för de tre frågorna.

b) Olika frågor är olika svåra för frågebesvarande system att hitta rätt svar på. Rangordna de tre frågorna från lättast till svårast. Motivera din rangordning. Antag att systemet inte har någon egen kunskapsdatabas utan bygger på dokumentsökning i svenska Wikipedia.

c) Ferrucci et al. (2010) undersökte ett slumpmässigt urval av 20 000 Jeopardy-frågor och hittade 2 500 distinkta svarstyper. De fann att de 200 mest frekventa svarstyperna täckte mindre an 50% av fragorna. Förklara varför detta är ett problem för frågebesvarande system som använder sig av maskininlärning.

Del B

09 **Deidentifiering**

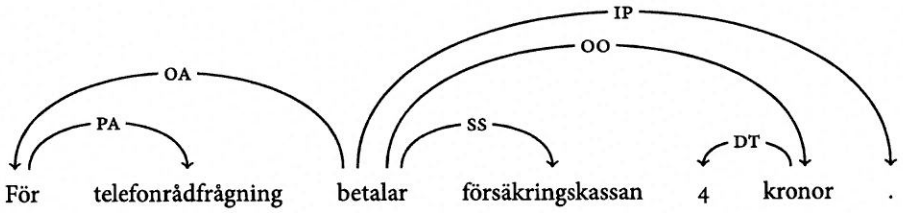
Du är konsult inom ett forskningsprojekt som ska analysera läkaranteckningar i patientjournaler. För att få etikprövning krävs att texterna deidentifieras, dvs. att all information som kan användas för att spåra datan till enstaka patienter tas bort. Exempel på sådan känslig information är namn, personnummer, telefonnummer och adress. Beskriv hur deidentifiering skulle kunna implementeras med hjälp av tekniker från kursen. Föreslå och motivera även ett relevant utvärderingsmått.

10 **Ordklasstaggning**

Diskutera likheter och skillnader mellan Viterbi-algoritmen och den perceptron-baserade algoritmen för ordklasstaggning. Vilka fördelar och nackdelar finns? För vilka tillämpningar passar den ena bättre än den andra?

11 **Transitionsbaserad dependensparsning**

Förklara utförligt hur parsning med en girig dependensparser fungerar. Illustrera ditt svar med hänvisningar till nedanstående dependensträd. Ange en fullständig transitionssekvens som skapar bågarna i detta träd.



12 **Frågebesvarande system**

Rita ett diagram över den standardarkitektur för frågebesvarande system baserat på dokumentsökning som vi lärt känna under kursen. Förklara de olika deluppgifterna i denna arkitektur och ge exempel på tekniker som kan användas för att lösa dessa. Använd relevant terminologi.