Försättsblad till skriftlig tentamen vid Linköpings universitet

Datum för tentamen	2019-08-20
Sal (1)	TER1(5)
Tid	14-18
Utb. kod	TDDE17
Modul	KTR1
Utb. kodnamn/benämning **Modulnamn/benämning**	Introduktion till språkteknologi Kontrollskrivning
Institution	IDA
Antal uppgifter som ingår i tentamen	5
Jour/Kursansvarig Ange vem som besöker salen	Marco Kuhlmann
Telefon under skrivtiden	013-284644
Besöker salen ca klockan	endast telefonjour
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Veronica Kindeland Gunnarsson, 013-285634, veronica.kindeland.gunnarsson@liu.se
Tillåtna hjälpmedel	inga
Övrigt	-/-
Antal exemplar i påsen	

Dugga 2019-08-20

Examinator: Marco Kuhlmann

Denna dugga består av 5 uppgifter som prövar din förståelse av de grundläggande begrepp och procedurer som behandlas på kursen. Dessa uppgifter kräver endast kompakta redogörelser, t.ex. en kort text, en uträkning eller ett diagram. Varje uppgift är värd 3 poäng. För att bli godkänd på duggan krävs minst 12 poäng totalt.

Lycka till!

01 Textklassificering (3 poäng)

a) Vid utvärderingen av en textklassificerare fick man ut nedanstående förväxlingsmatris. Den markerade cellen anger antalet gånger systemet klassade ett dokument som klass C medan det enligt guldstandarden tillhörde klass A.

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

Ange följande värden. Svara med bråk.

i. recall med avseende på klass A

ii. precision med avseende på klass C

b) Du ska bygga en textklassificerare som ska avgöra om ett inkommande dokument är en nyhet om Kina eller Japan. Träningsmängden innehåller 90% nyheter om Kina och 10% nyheter om Japan, och du räknar med samma fördelning även i framtiden. Förklara varför korrekthet (*accuracy*) är ett dåligt utvärderingsmått för din textklassificerare.

c) Här ser du klassificeringsregeln för Naive Bayes-klassificeraren under antagandet att man använder standardsannolikheter. Skriv om den till log-sannolikheter. Förklara din notation.

$$\hat{c} = \arg\max_{c \in C} \left(P(c) \cdot \prod_{w \in V} P(w \mid c)^{\#(w)} \right)$$

02 Språkmodellering (3 poäng)

Datamängden *Corpus of Contemporary American English* (COCA) består av ca. 560 miljoner token och innehåller 1 254 193 unika ord. Vi hittar följande frekvenser av unigram och bigram:

snow	*white*	*white snow*	*purple*	*purple snow*
38,186	256,091	122	11,218	0

a) Skatta följande sannolikheterna med hjälp av Maximum Likelihood-metoden utan utjämning. Svara med bråk.

i. $P(white)$ ii. $P(snow \mid white)$

b) Skatta följande sannolikheter med hjälp av Maximum Likelihood-metoden och addera-k utjämning, $k = 0.01$. Svara med bråk.

i. $P(purple)$ ii. $P(snow \mid purple)$

c) Vi använder Maximum Likelihood-metoden med addera-k utjämning för att skatta n-gram-modeller på COCA, där $n \in \{1, \ldots, 5\}$ och $k \in \{0, 0.1, 1\}$. Nedanstående tabell visar entropin för varje modell när den utvärderas på samma data som den tränats på. Vilken rad svarar mot vilket k-värde, och varför?

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
rad 1	7.3376	5.9834	6.7332	6.9556	7.0555
rad 2	7.3376	3.4269	1.4290	0.5436	0.4171
rad 3	7.3376	7.3837	8.4573	8.6577	8.7350

03 Ordklasstaggning (3 poäng)

a) Det finns många ord som kan tillhöra flera olika ordklasser. Ange ett exempel och förklara hur detta faktum kan leda till *kombinatorisk explosion.*

b) Den första metoden för ordklasstaggning som du lärt känna på kursen är "hidden Markov-modellen". Följande matriser specificerar (delar av) en sådan modell. Den markerade cellen anger övergångssannolikheten från BOS till AB.

	AB	PN	PP	VB	EOS
BOS	1/11	1/10	1/12	1/11	1/25
AB	1/11	1/11	1/11	1/10	1/14
PN	1/11	1/12	1/12	1/10	1/16
PP	1/13	1/11	1/12	1/14	1/18
VB	1/11	1/10	1/10	1/13	1/15

	she	got	up
AB	1/25	1/25	1/14
PN	1/13	1/25	1/25
PP	1/25	1/25	1/13
VB	1/25	1/14	1/19

Vi använder modellen för att tagga meningen "she got up". Beräkna sannolikheten för följande möjliga taggsekvens. Svara med bråk.

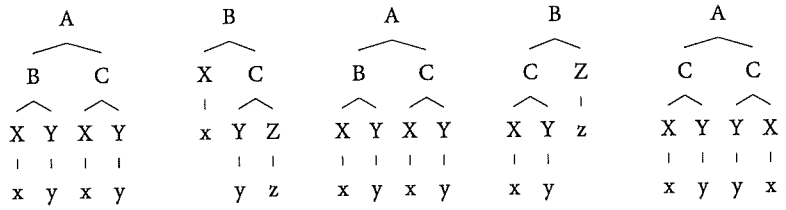
PN VB PP

c) Den andra metoden för ordklasstaggning som du lärt känna på kursen är "multiclass perceptron". Nedanstående tabell visar en jämförelse mellan de två metoderna. Fyll i de tre luckorna i denna jämförelse.

Hidden Markov-modellen	Multiclass Perceptron
A	använder vikter (godtyckliga reella tal)
B	girig sökning med lokalt optimala beslut
särdrag: aktuellt ord, föregående ordets tagg	C

04 Syntaktisk analys (3 poäng)

a) Du ska skatta en probabilistisk kontextfri grammatik utifrån den lilla trädbank som du ser här nedan. Skriv ner alla regler vars vänsterled är antingen B eller C och skatta deras sannolikheter med hjälp av maximum likelihood-metoden (utan smoothing). Svara med bråk.



b) Du summerar alla regelsannolikheter i en viss probabilistisk kontextfri grammatik. Vilka (noll eller flera) av följande resultat kan du *inte* få, och varför? Svara med en kort text.

i. 0.42 ii. 1 iii. 4.2 iv. 42

c) Rita det dependensträd som skapas av en transitionsbaserad dependensparser när den utför följande transitionssekvens:

SH SH SH LA SH RA SH SH RA RA RA

05 Semantisk analys (3 poäng)

a) Rita ett partiellt WordNet-träd för följande synsets:

1 university
2 institution, establishment
3 educational institution
4 hospital
5 kindergarten
6 medical institution

b) Här är tre signaturer (glossor och exempel) från Wiktionary för olika betydelser av ordet *course*:

1 A normal or customary sequence. **2** A learning program, as in university. *I need to take a French course.* **3** The direction of movement of a vessel at any given moment. *The ship changed its course 15 degrees towards south.*

Utifrån dessa signaturer, vilken av betydelserna prediceras av Simplified Lesk-algoritmen i nedanstående mening? Ignorera ordet *course*, stoppord och skiljetecken.

In the United States, the normal length of a course is one academic term.

c) Vi läser av ordvektorer för ord på klingonska från nedanstående samförekomstsmatris. (Målord svarar mot rader, kontextord svarar mot kolumner.)

	HuSHa'	*Ha'DIbaH*
qa'vIn	5	1
qurgh	5	5
jonta'	1	0
Dargh	1	4

Sortera de fyra målorden i avstigande grad av likhet till ordet *jonta'*. Antag att semantisk likhet kan mätas med hjälp av kosinusmåttet.