Dugga 2018-12-18

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. korrekthet (*accuracy*)

ii. recall med avseende på klass B

b) Ge ett exempel på en situation där korrekthet (*accuracy*) är ett dåligt utvärderingsmått för en 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. $P(purple)$ ii. $P(snow \mid white)$

b) Skatta bigramsannolikheten $P(snow \mid purple)$ med hjälp av Maximum Likelihood-metoden med addera-k utjämning, $k = 0{,}01$. Svara med ett bråk.

c) Vi använder Maximum Likelihood-metoden med addera-k utjämning för att skatta två trigram-modeller på COCA: modell A med $k = 0$ och modell B med $k = 1$. Vi beräknar de tränade modellernas entropi på samma data. Förklara varför modell B har en högre entropi än modell A.

03 Ordklasstaggning (3 poäng)

a) Ge åtminstone fyra exempel på ordklasser i det svenska språket.

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 sannolikheterna för följande möjliga taggsekvenser. Svara med bråk. Vilken sekvens tilldelas den högre sannolikheten och föreslås därmed av modellen?

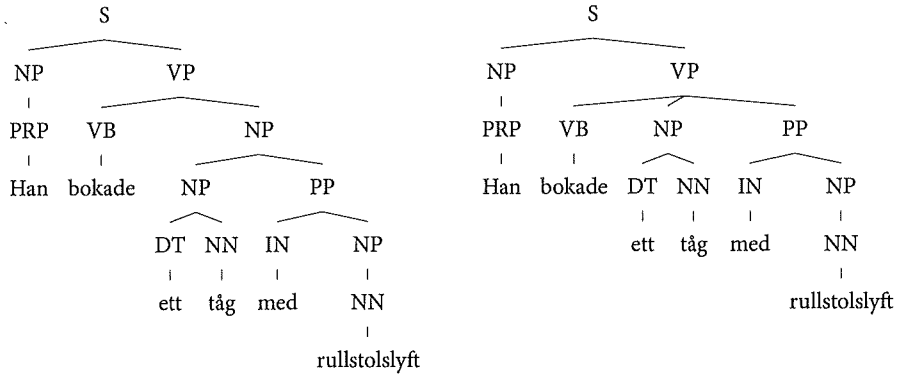
i. PN VB AB

ii. PN VB PP

c) Den andra metoden för ordklasstaggning som du lärt känna på kursen är "multiclass perceptron". Ange åtminstone tre skillnader mellan denna metod och metoden hidden Markov-modell.

04 Syntaktisk analys (3 poäng)

a) Du ska skatta en probabilistisk kontextfri grammatik utifrån den lilla trädbanken som du ser här nedan, bestående av två stycken frasstrukturträd. Ange sannolikheterna för alla NP-reglerna. Svara med bråk. Du behöver inte skatta sannolikheterna för de andra regeltyperna.



b) En transitionsbaserad dependensparser ska predicera ett dependensträd för en mening bestående av sex stycken ord. Den börjar i den initiala konfigurationen och utför följande transitioner:

SH SH LA SH SH LA SH RA SH RA RA

Rita det dependensträd som parsern predicerar. Representera orden genom att nummerera dem från 1 till 6.

c) Flera gånger på kursen har vi återkommit till begreppet "ambiguitet" (flertydighet). Förklara vad detta begrepp betyder i kontexten av syntaktisk analys och vilka utmaningar det medför för språkteknologi.

05 Semantisk analys (3 poäng)

a) Välj en semantisk relation: synonym, antonym, hyponym, hyperonym?

blomma	är ... till	ros
lysande	är ... till	klart
barn	är ... till	människa
ond	är ... till	god
maträtt	är ... till	ärtsoppa

b) En enkel metod för att bestämma ett ords betydelse är Lesks algoritm, som använder sig av semantiska lexikon. Följande betydelser hittar man när man slår upp det engelska ordet *course* i Wiktionary:

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.*

Vilken av betydelserna förutsäger Lesks algoritm för följande mening med de angivna definitionerna som underlag? Ignorera ordet *course*, skiljetecken och stoppord. Motivera ditt svar.

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 *qa'vIn*. Antag att semantisk likhet kan mätas med hjälp av cosinusmåttet.