TDDE17 Introduktion till språkteknologi (2017)

# Dugga 2018-08-21

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. precision med avseende på klass C ii. recall med avseende på klass B

b) Skriv klart klassificeringsregeln för Naive Bayes-klassificeraren under antagandet att man använder log-sannolikheter. Förklara din notation.

$$\hat{c} = \arg\max_{c \in C} \ldots$$

c) Förklara varför praktiska implementationer av Naives Bayes ofta använder log-sannolikheter.

02 **Ordpredicering** **(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 sannolikheterna $P(snow)$ och $P(snow \mid purple)$ med hjälp av Maximum Likelihood-metoden (utan utjämning). Svara med bråk.

b) Skatta bigramsannolikheten $P(snow \mid purple)$ med hjälp av Maximum Likelihood-metoden med addera-$k$ utjämning, $k = 0{,}1$. Svara med ett bråk.

c) Vi använder maximum likelihood-skattning med addera-$k$ utjämning för att träna $n$-gram modeller på COCA, med $n = \{1, \ldots, 5\}$ och $k \in \{0, 0.1, 1\}$. Nedanstående tabell visar entropin för varje modell på träningsdatan. Vilken rad svarar mot vilket $k$-värde? Förklara ditt svar.

| | $n=1$ | $n=2$ | $n=3$ | $n=4$ | $n=5$ |
|---|---|---|---|---|---|
| $k=a$ | 7.3376 | 5.9834 | 6.7332 | 6.9556 | 7.0555 |
| $k=b$ | 7.3376 | 3.4269 | 1.4290 | 0.5436 | 0.4171 |
| $k=c$ | 7.3376 | 7.3837 | 8.4573 | 8.6577 | 8.7350 |

## 03 Ordklasstaggning (3 poäng)

a) 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 |

Vilken sannolikhet tilldelar denna modell följande taggade mening (ord/tagg)? Svara med ett bråk.

up/AB she/PN got/VB

b) Den andra metoden för ordklasstaggning som du lärt känna på kursen är "multiclass perceptron". En skillnad mellan denna och hidden Markov-modellen är att den sistnämnda är mera begränsad när det gäller att definiera särdrag (*features*). Vilka av nedanstående särdrag har en hidden Markov-modell tillgång till? Förklara ditt svar.

i. aktuellt ord

ii. ordet till vänster om det aktuella ordet

iii. ordet till höger om det aktuella ordet

iv. ordklasstaggen av ordet till vänster av det aktuella ordet

c) Ange åtminstone två andra skillnader mellan metoderna hidden Markov-modell och multiclass perceptron, förutom att man i multiclass perceptron har större friheter att definiera särdrag.

**04** **Syntaktisk analys** **(3 poäng)**

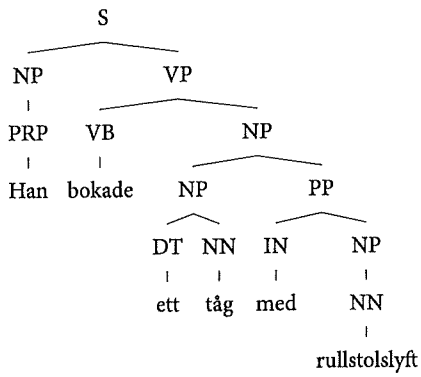
Nedan anges alla NP-regler och alla VP-regler i en viss probabilistisk kontextfri grammatik. Några värden saknas.

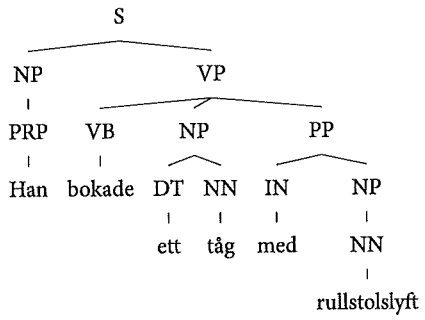
$$\text{NP} \rightarrow \text{PRP}\ \tfrac{2}{7} \quad \text{NP} \rightarrow \text{NP PP}\ \tfrac{1}{7} \quad \text{NP} \rightarrow \text{DT NN}\ \tfrac{2}{7} \quad \text{NP} \rightarrow \text{NN}\ a$$

$$\text{VP} \rightarrow \text{VB NP}\ \tfrac{1}{b} \quad \text{VP} \rightarrow \text{VB NP PP}\ \tfrac{1}{c}$$

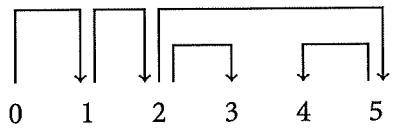
a) Ange värdena $a$, $b$, $c$.

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) Ange en transitionssekvens som låter en transitionsbaserad dependensparser återskapa följande dependensträd:



05 **Semantisk analys** **(3 poäng)**

Betrakta följande 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) Fyll i samförekomstmatrisen. Varje cell ska innehålla antalet gånger målordet (rad) förekommer i samma dokument (1)–(6) som kontextordet (kolumn).

| | passenger | transport | goal | match |
|---|---|---|---|---|
| automobile | ☐ | ☐ | ☐ | ☐ |
| car | ☐ | ☐ | ☐ | ☐ |
| soccer | ☐ | ☐ | ☐ | ☐ |
| football | ☐ | ☐ | ☐ | ☐ |

b) Rita målorden som vektorer i ett koordinatsystem där $x$-värdena svarar mot summan av antalet förekomster i kontexterna *passenger* och *transport* och $y$-värdena svarar mot summan av antalet förekomster i kontexterna *goal* och *match*.

c) Hur kan man med hjälp av sådana vektorrepresentationer mäta semantisk likhet mellan målorden? Vilka resultat skulle detta ge för målorden i exemplet?