729G17 Språkteknologi (2015)

Tentamen 2015-04-24

Marco Kuhlmann

Denna tentamen består av två delar: del A, som innehåller frågor 1–8, och del B, som innehåller frågor 9–12. Varje fråga är värd 3 poäng. För G krävs minst 12 poäng i del A. För VG krävs utöver detta minst 6 poäng i del B.

Viktig information! Du som skriver denna tentamen enligt den provuppsättning som gällde innan kursomgången 2015 (TEN1) ska även bearbeta extrauppgifterna 13 och 14 (efter del B). För G krävs då sammanlagt minst 15 poäng i del A och extrauppgifterna.

Del A

1. Ett textklassificeringssystem baserat på metoden Naive Bayes klassificerar nyhetstexter som antingen ”nyheter om Sverige” (S) eller ”nyheter om Norge” (N). Systemet använder följande sannolikheter:

   $$P(\mathtt{S}) = 3/4 \qquad P(\text{Stockholm} \mid \mathtt{S}) = 5/8 \qquad P(\text{Oslo} \mid \mathtt{S}) = 1/8$$
   $$P(\mathtt{N}) = 1/4 \qquad P(\text{Stockholm} \mid \mathtt{N}) = 1/3 \qquad P(\text{Oslo} \mid \mathtt{N}) = 2/3$$

   (a) Räkna ut de värden som systemet jämför för att avgöra om dokumentet ”Stockholm Stockholm Stockholm Oslo” är en nyhet om Sverige eller en nyhet om Norge. Ställ upp bråk. (b) Ange en dokumentsamling utifrån vilken man får de angivna sannolikheterna om man skattar med Maximum Likelihood-metoden.

2. Inom språkteknologin används ordet *ord* på flera olika sätt. Ange tre olika betydelser och ge exempel på sammanhang där dessa olika betydelser är relevanta.

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

(a) Ställ upp bråk för ML-skattningen (Maximum Likelihood) av unigramsannolikheten $P(\textit{är})$ och bigramsannolikheten $P(\textit{är} \mid \textit{det})$.

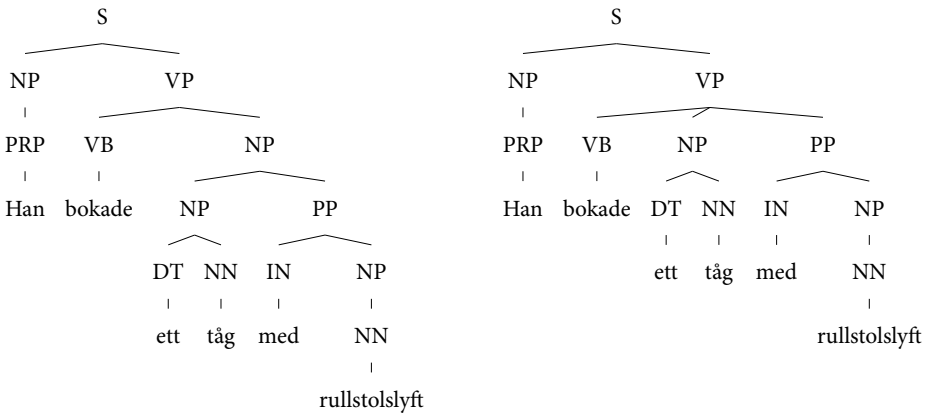
(b) Ställ upp ett bråk för ML-skattningen av bigramsannolikheten $P(\textit{nalkas} \mid \textit{det})$ med Add One-utjämning. Antag att vokabulären består av alla unika ord.

4. Här är två taggsekvenser för meningen *Jag skrev på utan att tveka*:

	Jag	skrev	på	utan	att	tveka
sekvens 1	PN	VB	PP	PP	IE	VB
sekvens 2	PN	VB	PL	PP	IE	VB

Du vet redan sannolikheten som en viss Hidden Markov-modell tilldelar sekvens 1 och vill nu veta sannolikheten som modellen tilldelar sekvens 2. Du kan dock endast fråga efter enstaka sannolikheter i modellen och varje sådan fråga kostar 10 kronor. Vilka frågor måste du ställa för att betala så lite som möjligt, och vilken formel måste du använda för att beräkna sekvenssannolikheten?

5. Nedanstående visas en liten trädbank bestående av två frasstrukturträd för meningen *Han bokade ett tåg med rullstolslyft*. Skatta sannolikheter för alla NP-regler och alla VP-regler med hjälp av Maximum Likelihood-metoden.



6. Tre frågor om WordNet: (a) Vad representerar noderna i WordNet? (b) Vad representerar länkarna? (c) Låt s_1 och s_2 vara två noder i WordNet och låt p stå för den kortaste vägen mellan dessa. Ange en formel för att räkna ut den semantiska likheten mellan s_1 och s_2.

7. Ett namnigenkänningssystem testades på en samling testdata innehållande 800 namnförekomster. Av dessa namn bestod 500 av ett ord, 250 av två ord och 50 av tre ord. Tabellen nedan anger systemets resultat.

	korrekta	falska
Ettordsnamn	420	60
Tvåordsnamn	200	40
Treordsnamn	44	12

Ställ upp bråk för följande:

(a) systemets recall (täckning) på ettordsnamn
(b) systemets precision på tvåordsnamn
(c) systemets recall på samtliga namn

8. Olika frågor är olika svåra för automatiska frågebesvarande system att hitta rätt svar på. Rangordna följande tre frågor från lättast till svårast och motivera rangordningen. Systemet antas inte ha någon egen kunskapsbas utan använda sig av Wikipedia-artiklar.

A. När föddes Einstein? B. Hur var han som människa? C. Var föddes Einsteins första fru?

Del B

9. Förklara kortfattat hur CKY-algoritmen för parsing med probabilistiska kontextfria grammatiker fungerar. Vad gör den? På vilken grundidé bygger dess effektivitet?

10. I flera typer av språkteknologiska system kan täckning (recall) inte mätas på det vanligaste sättet, dvs. genom att dividera antalet fall där system och facit överensstämmer med det totala antalet fall i facit. Ange två typer av system där detta inte fungerar så bra, förklara varför, och beskriv de utvärderingsmått som används i stället.

11. Du jobbar på ett företag som har utvecklat ett automatiskt system som kan klassificera filmrecensioner som antingen positiva eller negativa. Nu har ni blivit kontaktade av ett politiskt parti som vill använda ert system för att få fram attityder gentemot partiet som de reflekteras i Twitterinlägg. Ange skillnader mellan uppgifterna ”klassificera produktrecensioner” och ”analysera attityder i Twitterinlägg”. Beskriv sedan några konkreta tekniska problem som ni skulle behöva lösa för att anpassa ert system till den nya uppgiften.

12. Förklara modellen ”den brusiga kanalen” som används i samband med maskinöversättning. Hur kan man skatta de sannolikheter som ingår i denna modell?

Extrauppgifter

13. Ge tre exempel på tekniker som används för normalisering av textdokument (efter tokenisering) och förklara kort hur dessa tekniker fungerar.

14. For att utvärdera maskinöversättningssystem används BLEU som bygger på ett modifierat precisionsmått för n-gram. Förklara hur detta mått fungerar och varför man inte använder det vanliga precisionsmåttet. Illustrera dina svar med ett exempel.