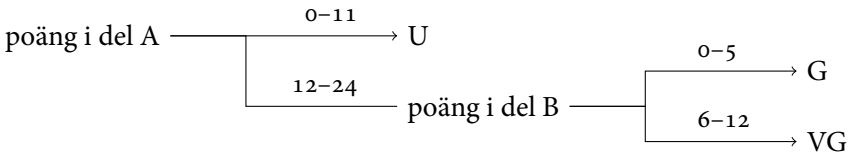
729G17 Språkteknologi (2015)

Tentamen 2015-08-21

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. Betyget sätts enligt följande schema:



Dina inlämningar till del B kommer endast att rättas om du har minst 12 poäng i del A. Rättningen av del A kommer då att avbrytas.

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 ska klassificera nyhetstexter som antingen ”nyheter om Sverige” (`S`) eller ”nyheter om Norge” (`N`). Systemet ska tränas på följande dokumentsamling:

	dokument	klass
1	Köpenhamn Stockholm	`S`
2	Oslo Stockholm Stockholm	`S`
3	Helsingfors Stockholm	`S`
4	Oslo Oslo Stockholm	`N`

(a) Ställ upp formler för de värden (Π-värden) som systemet jämför för att avgöra om dokumentet ”Stockholm Stockholm Stockholm Oslo” är en nyhet om Sverige eller en nyhet om Norge. (b) Skatta de sannolikheter som ingår i formlerna med hjälp av Maximum Likelihood-metoden.

2. Nedan ser du två versioner av ett (avformaterat och tokeniserat) dokument, en före och en efter normalisering. Identifiera de normaliseringstekniker som har tillämpats och beskriv dem kortfattat.

Före normalisering

```
Den liknar andra arter inom familjen
med böjd näbb , mönstrad brun
ovansida , vitaktig undersida och
långa styva stjärtpennor som den
använder för att kunna balansera
upprätt på trädstammar och grenar .
```

Efter normalisering

```
likna annan art familj böjd näbb
mönstrad brun ovansida vitaktig
undersida lång styv stjärtpenna
använda kunna balansera upprätt
trädstam gren
```

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 och sekvensen *det är* 927 gånger.

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

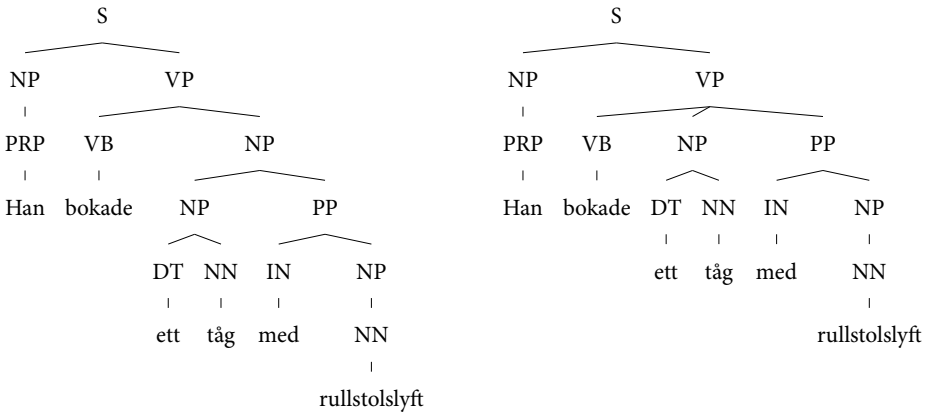
(b) Antar att sekvensen *det nalkas* inte förekommer i korpusen. Förklara hur man kan använda Add One-utjämning för att skatta bigramsannolikheten $P(nalkas \mid det)$.

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

	Jag	skrev	på	utan	att	tveka
sekvens 1	PN	VB	PL	PP	IE	VB
sekvens 2	PN	VB	PP	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 (tagg–tagg eller tagg–ord) 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. Nedanstående tabell visar hur ofta vissa målord (rader) förekommer tillsammans med vissa kontextord (kolumner) i en given textsamling.

	wheel	transport	passenger	tournament	goal	match
automobile	1	1	1	0	0	0
car	1	2	1	0	0	0
soccer	0	0	0	1	1	1
football	0	0	1	1	2	1

(a) Rita målorden som vektorer i ett koordinatsystem där x-axeln svarar mot det totala antalet förekomster tillsammans med ett av kontextorden *wheel*, *transport*, *passenger* och y-axeln svarar mot det totala antalet förekomster tillsammans med ett av kontextorden *tournament*, *goal*, *match*. (b) Förklara hur man med hjälp av sådana vektorrepresentationer kan mäta likhet mellan målorden.

7. Ett namnigenkänningssystem testades på en samling data 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 treordsnamn
(c) systemets precision på samtliga namn

8. 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 kortfattat och ge exempel på tekniker som kan användas för att lösa dessa. Använd relevant terminologi.

Del B

9. Förklara utförligt hur parsning med en girig dependensparser fungerar. Hur ser den initiala konfigurationen ut; vilka konfigurationer betraktas som slutkonfigurationer? Vilken roll har guiden? Utifrån vilken information ger guiden sina rekommendationer.

10. Ge en utförlig förklaring av modellen ”den brusiga kanalen” som den används i samband med maskinöversättning. Hur kan man skatta de sannolikheter som ingår i denna modell?

11. For att utvärdera maskinöversättningssystem används måttet BLEU. Förklara utförligt hur detta mått fungerar och varför man inte använder precision och täckning (recall) istället som i många andra tillämpningar. Illustrera ditt svar med ett exempel.

12. Du är konsult inom ett forskningsprojekt som ska analysera texter 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.

Extrauppgifter

13. Förklara begreppen token, lemma och lexem. Ge för varje begrepp ett exempel på en språkteknologisk tillämpning där begreppet är relevant.

14. Du ska organisera en tavling dar deltagarna ska bygga ordklasstaggare som tränas på data med hjälp av övervakad maskininlärning. (a) Beskriv hur datan för denna uppgift borde se ut. (b) Beskriv en lämplig baseline för tävlingen. (c) Varför bör testdatan för tävlingen hållas hemliga medan tävlingen pågår? Svara utförligt.