Tentamen 2016-03-30

Marco Kuhlmann

Tentamen består av två delar, A och B. Varje del omfattar ett antal frågor à 3 poäng. **Del A** omfattar 7 frågor som kan besvaras kortfattat. **Del B** omfattar 3 frågor av essäkaraktär; dessa kräver utförliga redovisningar i form av sammanhängande text. Planerade betygsgränser är 15 (för 3), 20 (för 4) och 25 (för 5).

Lycka till!

Del A

01 Flertydighet

Nedan visas möjliga ordklasstaggar för orden i meningen *jag bad om en kort bit*.

jag	bad	om	en	kort	bit
PN	VB	PP	DT	JJ	NN
NN	NN	SN	PN	AB	VB
		PL	RG	NN	
		AB	NN		

a) Vilken av de två möjliga ordklasserna har ordet *bad* i den aktuella meningen, verb (`VB`) eller substantiv (`NN`)?

b) Ange en mening där ordet *bad* har den andra ordklassen.

c) Flertydighet kan leda till så kallad kombinatorisk explosion. Förklara vad som menas med detta. Ta hjälp av bilden.

02 Korrekthet, precision och täckning (recall).

Vid utvärderingen av en ordklasstaggare fick man ut nedanstående förväxlingsmatris. Den markerade cellen anger antalet gånger systemet klassade ett ord som adjektiv (`JJ`) medan det enligt guldstandarden var ett substantiv (`NN`).

	NN	JJ	VB
NN	58	6	1
JJ	5	11	2
VB	0	7	43

a) Ställ upp ett bråk för taggarens precision på adjektiv.

b) Ställ upp ett bråk för taggarens täckning (recall) på verb.

c) Ange en annan förväxlingsmatris där taggarens korrekthet är samma som i matrisen ovan men där respektive värden för a) och b) är 0%.

03 Dokumentsökning

Innan ett textdokument förs in i ett dokumentindex genomgår det normalisering:

Ursprunglig version

```
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.
```

Normaliserad version

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

a) Identifiera de tekniker som har tillämpats på den ursprungliga versionen av dokumentet för att skapa den normaliserade versionen. Illustrera varje teknik med ett konkret exempel från texterna.

b) Några normaliseringstekniker kräver mer språkvetenskaplig kunskap eller mera avancerade språkvetenskapliga data än andra. Ordna de tekniker som du identifierat med avseende på denna skala. Motivera din rangordning kortfattat.

c) Vilken av dessa normaliseringstekniker kan till viss del simuleras genom att vikta söktermer med hjälp av tf–idf? Förklara!

04 Textklassificering

Ett system för textklassificering baserat på metoden Naive Bayes ska avgöra om dokumentet ”Tokyo Tokyo Peking” är en nyhet om Japan (klass J) eller en nyhet om Kina (klass K).

a) För att predicera dokumentets klass använder systemet bl.a. sannolikheterna $P(\texttt{J})$ och $P(\text{Tokyo} \mid \texttt{J})$. Lista de övriga sannolikheter som är relevanta.

b) Skatta de relevanta sannolikheterna med Maximum Likelihood-metoden utifrån följande dokumentsamling. Ställ upp bråk.

	dokument	klass
1	Tokyo Tokyo	J
2	Tokyo Peking	J
3	Tokyo Seoul	J
4	Peking Tokyo	K

c) Utifrån de skattade sannolikheterna, vilken klass predicerar systemet? Redovisa hur du räknat. Visa tydligt att du förstått Naive Bayes-klassificeringsregeln.

05 Ordpredicering

I en text innehållande 1 200 300 löpord och 105 400 unika ord hittas ordet *det* 13 600 gånger, ordet *är* 13 700 gånger, ordet *nalkas* 2 gånger, bigrammet *det är* 900 gånger och bigrammet *det nalkas* 0 gånger.

a) Skatta unigramsannolikheten $P(det)$ och bigramsannolikheten $P(är \mid det)$ med Maximum Likelihood-metoden. Ställ upp bråk.

b) Vad händer när man skattar bigramsannolikheten $P(nalkas \mid det)$ med Maximum Likelihood-metoden? Varför kan detta vara ett problem?

c) Skatta bigramsannolikheten $P(nalkas \mid det)$ med Add One-utjämning. Utgå ifrån att vokabulären består av alla ord i texten. Ställ upp bråk.

06 Informationsextraktion

Ett system för informationsextraktion är tränat på att hitta tre typer av namngivna entiteter: personer (PER), organisationer (ORG) och svenska tätorter (LOC).

a) Vilken av dessa tre typer är lättast att hitta med hjälp av namnlistor?

b) Entitetsextraktion kan ses som uppgiften att tagga varje token i en mening med en så kallad BIO-tagg; ett exempel är B-PER. Sätt ut BIO-taggarna för följande mening. (Observera att meningen består av 20 stycken token.)

   Astrid Lindgren , född den 14 november 1907 i Vimmerby , utsågs till hedersdoktor vid Linköpings universitet år 2000 .

c) System som använder BIO-taggningen kan utvärderas på taggnivå eller entitetsnivå. Ändra en av dina taggar så att den nya taggningen har 95% korrekthet på taggnivå men 0% precision och recall med avseende på entitetstypen PER. Använd din ursprungliga taggning som guldstandard.

07 Textsammanfattning

För utvärdering av textsammanfattningssystem används måttet ROUGE (Recall-Oriented Understudy of Gisting Evaluation).

Referenstext	**Systemtext**
Möbeljätten Ikea planerar att bli dubbelt så stort år 2020. På åtta år innebär det att mellan 160 och 200 nya varuhus ska öppnas.	Möbeljätten Ikea ska bli dubbelt så stort år 2020. Tusentals människor kommer att behöva anställas.

a) Ställ upp ett bråk för systemets ROUGE-1-värde (avser unigram).

b) Ställ upp ett bråk för systemets ROUGE-2-värde (avser bigram).

c) Hur skulle du kunna fuska i en utvärdering baserad på ROUGE-1?

Del B

08 **Centralitet**

En teknik som används för att välja ut de viktigaste meningarna ur en text i ett extraktionsbaserat textsammanfattningssystem bygger på idéen att mäta hur ”central” en kandidatmening är för dokumentet. Förklara denna teknik utförligt. Illustrera din förklaring med bilder.

09 **Kunskapsglappet**

I sin bok *Megatrends* (1982) myntade John Naisbitt aforismen: ”We are drowning in information but starved for knowledge.” Konkretisera detta citat med några exempel som har anknytning till kursen. Diskutera frågan om språkteknologi kan hjälpa oss att övervinna ”kunskapsglappet”.

10 **Attitydanalys**

Ditt företag har utvecklat ett framgångsrikt system för klassificering av nyhetstexter i kategorier såsom *politik*, *kultur* och *sport*. Systemet bygger på modellen Naive Bayes. Nu blir ni kontaktade av en kund som undrar om ert system även kan användas för predicering av attityder gentemot kundens produkter utifrån yttranden på sociala medier som Twitter och Facebook.

Diskutera: Vad skulle du svara kunden? Vilka likheter och skillnader ser du med den nya tillämpningen? Vad skulle ni behöva göra för att anpassa ert system?