Tentamen 2016-01-13

Marco Kuhlmann

Denna tentamen består av 10 frågor. Frågorna 8–10 ligger på en högre kunskapsnivå än de övriga och kräver utförliga redovisningar. Varje fråga kan ge maximalt 3 poäng. Planerade betygsgränser är 16 (för 3), 20 (för 4) och 24 (för 5). Lycka till!

01 Det finns egenskaper hos naturligt språk som gör att många språkteknologiska problem är svåra; en av dessa egenskaper är flertydighet.

a) Ge ett exempel som illustrerar att naturligt språk kan vara flertydigt.

b) Flertydighet kan leda till så kallad kombinatorisk explosion. Förklara vad som menas med detta och ge ett konkret exempel som illustrerar problemet.

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

Ursprunglig version	**Normaliserad 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.`	`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.

03 Ett system för textklassificering baserat på metoden Naive Bayes ska avgöra om dokumentet ”Stockholm Stockholm Oslo” är en nyhet om Sverige (klass `S`) eller en nyhet om Norge (klass `N`).

a) Lista alla sannolikheter som systemet behöver ha tillgång till för att predicera dokumentets klass.

b) Skatta dessa sannolikheter med Maximum Likelihood-metoden utifrån följande dokumentsamling. Ställ upp bråk.

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

c) Beräkna de värden som systemet jämför för att avgöra dokumentets klass. Vilken klass predicerar systemet?

04 I en text innehållande 1 215 396 löpord och 105 436 unika ord hittas ordet *det* 13 694 gånger, ordet *är* 13 700 gånger, ordet *nalkas* 2 gånger, bigrammet *det är* 927 gånger och bigrammet *det nalkas* 0 gånger.

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

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

c) Skatta bigramsannolikheten $P(\textit{nalkas} \mid \textit{det})$ med en annan metod än Maximum Likelihood. Ställ upp bråk.

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

	NN	JJ	VB
NN	60	6	3
JJ	6	12	3
VB	0	6	42

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

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

c) Ange en annan förväxlingsmatris där taggarens korrekthet är samma som i matrisen ovan men täckning på adjektiv är 0%.

06 Förklara den standardarkitektur för frågebesvarande system som vi gått genom på kursen. Beskriv de olika delproblemen och ge exempel på tekniker som kan användas för att lösa dem.

07 Det vanliga precisionsmåttet kan generaliseras för att utvärdera system där resultatet inte är ett enda svar utan en rankad lista av svarsalternativ.

a) Ge exempel på tillämpningar där denna generalisering av precision är relevant och ange de specifika måtten som används i dessa sammanhang.

b) Förklara skillnaderna mellan dessa mått.

08 När man googlar efter ett felstavat ord som t.ex. *rätstavning* så tar det endast några bråkdelar av en sekund och Google frågar:

Menade du: *rättstavning*

I den här uppgiften ska du fundera på hur denna teknik kan implementeras.

a) Skissa på en algoritm som tar ett ord w och beräknar alla ord vars Levenshtein-avstånd till w är exakt ett.

b) Hur skulle man kunna kombinera denna algoritm med en unigrammodell för att föreslå det mest sannolika rättstavade ordet?

c) Hur skulle man kunna generalisera ansatsen till ord vars Levenshtein-avstånd till det felstavade ordet är större än ett? Vilket beräkningsmässigt problem uppstår?

09 Du är konsult i ett forskningsprojekt som vill analysera texter i patientjournaler. För att få etikprövning för projektet krävs att texterna deidentifieras, dvs. att all information som kan användas för att spåra data till enstaka patienter tas bort. Exempel på sådan känslig information är namn, personnummer och adress.

Beskriv hur deidentifieringen skulle kunna implementeras med hjälp av tekniker från kursen och vad för sorts resurser detta skulle kräva. Föreslå och motivera även ett relevant utvärderingsmått för det implementerade systemet.

10 Ditt företag har utvecklat ett framgångsrikt system för klassificering av nyhetstexter. Nu blir ni kontaktade av en kund som undrar om systemet även kan användas för predicering av attityder gentemot kundens produkter utifrån yttranden på sociala medier som Facebook och Twitter.

Diskutera likheter och skillnader mellan de två tillämpningarna. Vilka utmaningar ser du med den nya tillämpningen? Vilka resurser skulle ni behöva om ni skulle bestämma er att utveckla en anpassad version av ert system?