TDDD02 Språkteknologi för informationssökning (2014)

Tentamen 2015-08-18

Marco Kuhlmann

Denna tentamen består av 10 frågor; varje fråga är värd 3 poäng. Planerade betygsgränser är 15 (för 3), 20 (för 4), 25 (för 5). Besvara gärna flera frågor på samma papper. Lycka till!

1. 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.

```
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
```

2. Nedanstående tabell visar antalet förekomster av tre söktermer (*bil*, *försäkring*, *bäst*) i en dokumentsamling bestående av fyra stycken textdokument (D1–D4).

	D1	D2	D3	D4
bil	27	4	24	12
försäkring	0	33	29	0
bäst	14	0	17	0

   Ange följande. När det behövs, använd en logaritm med bas 2.

   (a) termfrekvensen för *bil* i D3
   (b) idf-värdet för *försäkring*
   (c) tf-idf-värdet för *försäkring* i D2

3. En textfil innehållande en samling svenska ord är formaterad så att varje ord står på en egen rad. Ange reguljära uttryck som matchar ord enligt följande. Du får skriva `\v` för att matcha alla svenska vokaler (versaler och gemener).

(a) ord som börjar på en versal
(b) ord som innehåller minst två vokaler
(c) ord som slutar på *arna*

4. 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 ML-skattningen (Maximum Likelihood) 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 $P(nalkas \mid det)$.

5. Definiera Levenshteinavstånd. Ange Levenshteinavståndet för orden *kul* och *kurs*.

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

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

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

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 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 och ge exempel på tekniker som kan användas för att lösa dessa.

9. 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.

10. 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.