TDDD02 Språkteknologi för informationssökning (2014)

# Tentamen 2015-01-14

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. Ge tre exempel på tekniker som används för normalisering av textdokument i samband med dokumentsökning och förklara kort hur dessa tekniker fungerar.

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 |

   (a) Ange termfrekvensen för *bil* i D1. (b) Ange den inversa dokumentfrekvensen för *bäst*. Använd en logaritm med bas 2. (c) Ange tf-idf-värdet för *försäkring* i D2.

3. Antag att `\v` är ett reguljärt uttryck som matchar alla vokaler i det svenska alfabetet (versaler och gemener). Ange reguljära uttryck för följande (svenska) ordformer:

   (a) ord som börjar på en versal (stor bokstav) och innehåller minst två tecken
   (b) ord som slutar på *orna* och därutöver innehåller minst en vokal
   (c) ord som innehåller minst två vokaler separerade av minst en konsonant

4. Antag att vi i en korpus som omfattar 100 000 ord hittar ordet *det* 1 500 gånger, ordet *är* 1 800 gånger, sekvensen *det är* 250 gånger, ordet *sägs* 10 gånger, och sekvensen *det sägs* 0 gånger. Ställ upp bråk för Maximum Likelihood-skattningen av

   (a) unigramsannolikheten $P(det)$
   (b) bigramsannolikheten $P(är \mid det)$
   (c) bigramsannolikheten $P(sägs \mid det)$

**5.** Definiera Levenshteinavstånd. Ange Levenshteinavståndet för orden *sus* och *brus*.

**6.** Ett textklassificeringssystem baserat på metoden Naive Bayes ska klassificera engelska nyhetstexter som antingen ”texter som handlar om Kina” (K) eller ”texter som handlar om Japan” (J). Systemet ska tränas på nedanstående dokumentsamling:

| | dokument | klass |
|---|---|---|
| 1 | Chinese Beijing Chinese | K |
| 2 | Chinese Chinese Shanghai | K |
| 3 | Chinese Tokyo | K |
| 4 | Tokyo Japan Chinese | J |

Antag att systemet ska predicera klassen för dokumentet ”Chinese Chinese Chinese Tokyo”. Skatta de för denna klassificering relevanta sannolikheterna med Maximum Likelihood-metoden. Ställ upp bråk.

**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 alla namn.

**8.** Vilka är de centrala delproblemen i en standardarkitektur för ett frågebesvarande system baserat på dokumentsökning? Förklara kortfattat vad delproblemen innebär.

**9.** Förklara begreppet informationsbehov och hur det kommuniceras vid (a) dokumentsökning och (b) informationsutvinning.

**10.** Ange tre metoder som används för att välja ut de viktigaste meningarna ur en text i ett extraktionsbaserat textsammanfattningssystem. Förklara dem kortfattat.