TENTAMEN

TDDD02 Språkteknologi för informationssökning

tisdag 20 augusti 2013 kl 14-18

Inga hjälpmedel är tillåtna.

Varje fråga är värd 3 poäng. Maximal poäng är 30. 15 poäng ger säkert godkänt. Det går bra att besvara flera frågor på samma papper.

1. I dokumentsökningssystem används ofta någon form av textnormalisering. Visa hur en mening som 'Sveriges äldsta motorväg, en del av E22, går mellan Malmö och Lund.' kan påverkas av textnormalisering.

2. När dokument representeras med termvektorer tillämpas ofta någon form av termviktning, t.ex. baserat på måtten ***tf*** och ***idf***. Förklara vad dessa mått innebär och ange något vanligt sätt de kombineras på i termviktningsmodeller.

3. Skriv ett reguljärt uttryck som matchar alla ord i en text som innehåller minst en vokal och minst tre konsonanter i följd.

4. Vad innebär namnigenkänning och hur mäter man prestanda för namnigenkänningssystem?

5. I nedanstående tabell visas frekvensdata för några ord och ordsekvenser från en svensk korpus med totalt 500,000 ord. Uppskatta med hjälp av dessa data följande ngram-sannolikheter: (a) p(en), (b) p(en | av), (c) p(pojkarna | en av).

en	6000
av	1200
en av	90
av en	60
av pojkarna	8
en av pojkarna	0

lars.ahrenberg@liu.se

LINKÖPINGS UNIVERSITET

6. (a) Vad menas med utjämning (eng. *smoothing*) av en språkmodell och varför tillämpar man det? (b) Hur bör perplexiteten ändras för en språkmodell om man använder utjämning?

7. Vilket är redigeringsavståndet mellan orden *broar* och *bryggor?* Var noga med att ange vilken definition du använder och visa hur man räknar ut det algoritmiskt.

8. Förklara vad som menas med ordbetydelsebestämning (eng. *word sense disambiguation*) och visa hur modellen Naive Bayes kan användas för detta syfte.

9. I standardarkitekturen för ett frågebesvarande system brukar man hitta en modul för frågeanalys. Förklara vad denna modul gör och hur dess utdata används av andra moduler i systemet.

10. Ange tre indikatorer som är användbara vid textsammanfattning baserad på extraktion.