Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2010-04-08*
Sal	*TER2*
Tid	*14-18*
Kurskod	*TDDD02*
Provkod	*TEN1*
Kursnamn/benämning	*Språkteknologi för informationssökning*
Institution	*IDA*
Antal uppgifter som ingår i tentamen	*11*
Antal sidor på tentamen (inkl. försättsbladet)	*3*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*0703182422*
Besöker salen ca kl.	*-*
Kursadministratör (namn + tfnnr + mailadress)	*Elisabeth Qvarnström 013-281706, eliqv@ida.liu.se*
Tillåtna hjälpmedel	*inga*
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	
Vilken typ av papper ska användas, rutigt eller linjerat	*valfritt*
Antal exemplar i påsen	*12+2*

TENTAMEN

TDDD02 Språkteknologi för informationssökning

torsdag 8 april 2010 kl. 14-18

Inga hjälpmedel är tillåtna.

Varje fråga är värd 3 poäng. Maximal poäng är 33. 16,5 poäng ger säkert godkänt. Det går bra att besvara flera frågor på samma papper.

1. I dokumentsökningssystem används ofta någon form av textnormalisering, t.ex. via *stemming* eller *lemmatisering*. Förklara vad dessa bearbetningar går ut på och ge exempel som visar skillnaden mellan dem.

2. Förklara vad som menas med termviktning enligt tf-idf-modellen, dvs förklara vad tf står för, vad idf står för och hur de kombineras vid termviktning.

3. Ange en reguljär substitution eller en ändlig transduktor (eng. *Finite-state transducer*) som kan utföra mappningar som de följande. Din lösning ska generalisera till andra substantiv som slutar på -el, -er eller -en.

   sedel → sedel N SG

   sedlar → sedel N PL

   utter → utter N SG

   uttrar → utter N PL

   öken → öken N SG

   öknar → öken N PL

4. System för namnigenkänning utvärderas vanligen med både precision och recall. (a) Förklara varför det inte är tillräckligt att bara ange hur många namn systemet hittat, (b) Ange ett mått, med definition, som väger ihop precision och recall.

5. I en text på sammanlagt 2000 ord finns 150 som börjar med stor bokstav. 80 av dessa inleder en mening och 90 är egennamn. Det finns inga meningar och inga egennamn som inte börjar med stor bokstav. Det finns 30 ordbigram, d.v.s. båda de ingående orden är egennamn. Använd Maximum Likelihood-uppskattning för att uppskatta följande sannolikheter: (a) sannolikheten för att ett ord som börjar med stor bokstav är ett egennamn, (b) sannolikheten för att ett egennamn följs av ett egennamn, (c) sannolikheten för att ett ord som inte står först börjar med stor bokstav.

Institutionen för datavetenskap
Linköpings universitet
lars.ahrenberg@liu.se

6. Ange (a) ett uttryck för sannolikheten av ordsekvensen *men han kom inte* givet att vi känner till bigramsannolikheterna för orden i meningen, (b) samma fråga men under förutsättning att vi känner till de ingående ordens trigramsannolikheter.

7. System för ordklasstaggning utnyttjar ords lokala kontext för att välja rätt tagg. Ange ett vanligt regelformat för sådan disambiguering och formulera någon eller några explicita regler som kan appliceras på ordet *visa* i dessa två meningar:

   Eva ville visa oss sina leksaker.

   Jag ville sjunga en visa om kärleken.

8. För att hitta ord med specifika semantiska relationer till varandra kan man använda mönster kodade som reguljära uttryck. Tre exempel på sådana mönster ges nedan. Ange för vart och ett av dem vilken eller vilka semantiska relationer man är ute efter. Svaret ska ange den relation som ordet X har till ordet Y.

   (a) ... andra X så som Y, Z och W ...

   (b) ... är X eller Y spelar ingen roll ...

   (c) ... en/ett X är en/ett Y som ...

9. I modellen den brusiga kanalen (eng. *Noisy Channel*) används beslutsregeln

$$S^* = \underset{S}{\operatorname{argmax}}\, p(S|O)$$

   (a) Förklara denna regel i ord. (b) Beräkningen av p(S|O) baseras vanligtvis på en omskrivning som leder till två olika sannolikhetsfördelningar. Visa hur omskrivningen ser ut och ange vad de resulterande sannolikheterna kallas.

10. Ange fyra vanliga komponenter i ett informationsextraktionssytem och förklara vad de gör. Du kan förutsätta att indata till den första komponenten är en tokeniserad ordklasstaggad text som är relevant för systemets domän.

11. I extraktionsmetoden för textsammanfattning används olika kriterier för att bestämma om en mening ska inkluderas i sammanfattningen eller inte. Nämn tre sådana kriterier.