Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2010-08-20*
Sal	*TER1*
Tid	*14-18*
Kurskod	*TDDD01*
Provkod	*TEN1*
Kursnamn/benämning	*Språkteknologi*
Institution	*IDA*
Antal uppgifter som ingår i tentamen	*11*
Antal sidor på tentamen (inkl. försättsbladet)	*3*
Jour/Kursansvarig	*Maria Holmqvist*
Telefon under skrivtid	*Ank. 1466 eller 0709-736834*
Besöker salen ca kl.	*15:00*
Kursadministratör (namn + telnr. + mailadress)	*Anna Grabska Eklund 013-282362, annek@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	

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

TDDD01 Språkteknologi

fredag 20 augusti 2010

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 26. 13 poäng ger säkert godkänt.

Det går bra att besvara flera frågor på samma papper.

1. Vad menas med att tokenisera en text? (2p)

2. Definiera måtten precision och recall, t.ex. för ett namnigenkänningssystem och förklara varför det inte räcker med bara ett av dessa mått för att beskriva systemets kvalitet. (3p)

3. (a) Vad menas med en trigrammodell för att predicera ord i kontext? (b) Vilket är det vanligaste sättet att uppskatta trigramsannolikheter? (2p)

4. Skriv ett reguljärt uttryck som kan känna igen prisangivelser upp till 10,000:-. Uttrycket ska klara av öresangivelser som i 99:90, 850:- och 75:00 men du kan anta att alla tecken är sammanskrivna, dvs att det inte finns något blanktecken inuti prisangivelsen. (3p)

5. Vad gör en Porter-stemmer och varför vill man använda en sådan? (2p)

6. I chartparsning används s.k. punkterade regler. Ett exempel är 'S → NP . VP'. Vad anger punkten? (2p)

7. Förklara principerna för en kaskadparser och ange minst två centrala komponentr i en sådan. (3p)

8. Språkteknologi använder sig ofta av rudimentära semantiska representationer (som många inte tycker förtjänar epitetet 'semantiska'), som t.ex. ordpåsar (eng. *bag-of-words*) eller mallar (eng. *templates*). Förklara vad dessa båda representationer står för och ange för var och en av dem någon typ av system som använder dem. (2p)

9. Ange två olika slags särdrag (eng. *features*) som är användbara i samband med ordbetydelsebestämning. (2p)

Institutionen för
datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post: lah@ida.liu.se

10. Förklara vad som menas med redigeringsavstånd och beskriv någon algoritm för att beräkna redigeringsavståndet mellan två ord. Tillämpa algoritmen på något exempel och ange slutligen vilka för- och nackdelar algoritmen har i samband med stavningskontroll. (3p)

11. En grundläggande modell för statistisk maskinöversättning är den brusiga kanalen (eng. The Noisy Channel). Beskriv denna modell och hur den är tänkt att ge oss en lösning för den bästa översättningen. (2p)