Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2014-01-15
Sal	TER3, TER4
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	*013-282422*
Besöker salen ca kl.	*14.45*
Kursadministratör (namn + tfnnr + mailadress)	*Helene Meisinger* *281868, helene.meisinger@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	

TENTAMEN

TDDD02 Språkteknologi för informationssökning

onsdag 15 januari kl 14-18

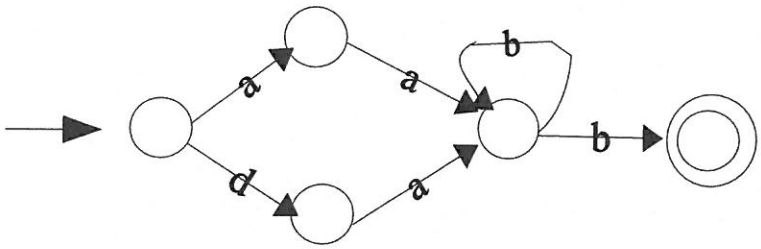
Inga hjälpmedel är tillåtna.

Varje fråga är värd 3 poäng. Maximal poäng är 33. Planerade betygsgränser är 16,5 (för trea), 23 (för fyra), 28 (för femma). Det går bra att besvara flera frågor på samma papper.

1. Förklara med exempel begreppen (a) *lemma*, (b) *stoppord*, (c) *token*.

2. I ett givet dokumentsökningssystem representeras sökfrågor och dokument med termvektorer. Varje termvektor är normaliserad till längden 1 medan elementen i vektorerna är proportionella mot termfrekvensen för ett visst ord i dokumentet. (a) Vad är syftet med att normalisera termvektorerna?, (b) Låt $\overline{Q}$ vara termvektorn för en given sökfråga. Hur bestäms det eller de dokument som är mest relevanta i förhållande till $\overline{Q}$?

3. Skriv ett reguljärt uttryck som genererar samma språk som automaten nedan. Gör uttrycket så kort som möjligt.



4. 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:

	Sanna positiva	Falska positiva
Ettordsnamn	420	60
Tvåordsnamn	200	40
Treordsnamn	44	12

(a) Vad är systemets precision på ettordsnamn?, (b) Vad är systemets recall på tvåordsnamn?, (c) Vad är systemets recall på alla namn?

5. Vad är ett bokstavsträd och vad används sådana till i språkteknologi och/eller informationssökning?

6. Ange tre olika metoder från kursen som kan tillämpas på problemet stavningskontroll, häri inräknat både problemet att känna igen ett felstavat ord och att gissa vilket ord som avsågs.

7. Förklara (a) vad som menas med en betingad sannolikhet, (b) hur ordsannolikheter betingas i en språkmodell baserad på trigram?

8. Vad menas med utjämning (eng. *smoothing*) av en språkmodell och varför tillämpar man det?

9. Vad innebär informationsutvinning (eng. information extraction) och vilka är de centrala komponenterna i ett informationsutvinningssystem?

10. Förklara modellen Naive Bayes för att klassificera objekt utifrån deras egenskaper och visa hur modellen kan appliceras på problemet att extrahera meningar ur en text för att skapa en sammanfattning.

11. Vad är Watson? Ange (minst) två egenskaper i Watsons arkitektur som skiljer det från tidigare frågebesvarande system.