Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2014-08-19
Sal	G33, FOI hus G, Campus Valla
Tid	14-18
Kurskod	TDDD02
Provkod	TEN1
Kursnamn/benämning	Språkteknologi för informationssökning
Institution	IDA
Antal uppgifter som ingår i tentamen	*10*
Antal sidor på tentamen (inkl. försättsbladet)	*3*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*/2422*
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

tisdag 19 augusti kl 14-18

Inga hjälpmedel är tillåtna.

Varje fråga är värd 3 poäng. Maximal poäng är 30. Planerade betygsgränser är 15 (för trea), 20,5 (för fyra), 25,5 (för femma). Det går bra att besvara flera frågor på samma papper.

1. Förklara vad som menas med att tokenisera respektive normalisera en text.

2. Ange ett reguljärt uttryck som matchar orden i mängden {*is, isen, isar, isens, isars*} men inga andra bokstavssekvenser, (b) Rita upp en tillståndsautomat (FSA) som motsvarar det reguljära uttrycket, (c) Visa hur automaten kan utvidgas för att även hantera orden *isarna* och *isarnas*..

3. Ett namnigenkänningssystem testades på en samling testdata innehållande 500 namnförekomster. Av dessa namn bestod 200 av ett ord, 250 av två ord och 50 av tre ord. Tabellen nedan anger några av systemets resultat:

	Sanna positiva	Falska positiva
Ettordsnamn	170	40
Tvåordsnamn	200	10
Treordsnamn	20	22

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

4. Vad är Levenshteinavstånd? Ge exempel på två ord som har Levenshteinavståndet 2.

5. (a) Vad menas i sannolikhetsläran med att två händelser är oberoende? (b) Även i situationer där utfall inte (säkert) är oberoende antar man gärna att de är det. Varför det? (c) Ange en sådan modell som gåtts igenom i kursen och någon tillämpning av den.

6. Förklara vad som menas med en språkmodell baserad på bigram. Vad menas med att tillämpa utjämning (eng. *smoothing*) på en sådan modell och varför gör man det?

7. Vad innebär betydelsebestämning (eng. *word sense disambiguation*)? Ange kortfattat två metoder som används för detta ändamål.

lars.ahrenberg@liu.se

LINKÖPINGS UNIVERSITET

8. (a) Förklara vad som menas med begreppen anafor och antecedent och ge något exempel. (b) Varför är det viktigt att kunna identifiera antecedenter i samband med informationsutvinning?

9. Hur fungerar extraktionsmetoden för textsammanfattning?

10. Nedan anges fyra titlar på vetenskapliga artiklar. Tre av dem kan på goda grunder antas handla om samma sak, medan den fjärde ser ut att handla om någonting helt annat. Beskriv hur en vektorrumsmodell för titlar fångar sådana samband.

> Posoning by domestic vipers (Vipera berus and Vipera aspis): a retrospective study of 113 patients
>
> Antivenom treatment in Vipera berus envenoming – report of 30 cases
>
> A nationwide study of Vipera berus bites during one year – a study of 231 cases
>
> A block-sorting lossless data compression algorithm