Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2011-08-25
Sal	TER2
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	*013-282422*
Besöker salen ca kl.	*14.40*
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	
Antal exemplar i påsen	

TENTAMEN

TDDD02 Språkteknologi för informationssökning

torsdag 25 augusti 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. Förklara varför man använder textnormalisering och ange två olika typer av normalisering. Typerna anges både med ett namn eller en beskrivning och en konkret illustration av hur ett ord förändras med den typen av normalisering.

2. Skriv ett reguljärt uttryck som endast matchar ord som innehåller tre vokaler och slutar på bokstaven g. Exempel: *dagstidning, kamratlig, olycklig.*

3. Ett namnigenkänningssystem uppgavs ha en precision på 85% när det testats på en dokumentsamling innehållande 500 namnförekomster. Går det med ledning av dessa uppgifter att avgöra hur många namn systemet hittade? Kan man ange ett intervall för antalet hittade namn? Svaren måste motiveras.

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

5. Förklara (a) generellt vad som menas med en språkmodell, (b) specifikt att en sådan modell är trigrambaserad och använder sig av back-off.

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

7. Förklara modellen Naive Bayes för att klassificera ord eller dokument. Det går bra att göra det generellt eller utifrån ett konkret klassificeringsproblem med konkret angivna indikatorer.

8. Vad menas med en boot-strapping metod? Förklara hur man kan använda boot-strapping för att skapa en mängd reguljära uttryck med hög prestanda för att känna igen relationer i textdokument, t.ex. relationen mellan ett land och dess huvudstad.

lars.ahrenberg@liu.se

LINKÖPINGS UNIVERSITET

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

10. I frågebesvarande system sker den initiala matchningen av den givna frågan mot meningar eller stycken i textdatabasen med användning av vektorrepresentationer. Förklara vad detta innebär och illustrera med följande meningar, givet frågan *Vilka länder gränsar till Belgien?*

    1. *Vi börjar redan planera nästa sommarsemester och är sugna på att ta bilen ner genom Holland, Belgien och Frankrike.*
    2. *Belgien gränsar i norr till Nederländerna, i söder till Frankrike och i öster till Tyskland och Luxemburg.*
    3. *Nederländerna gränsar bara till två länder: Belgien och Tyskland.*