

Tentamen i Språkteknologi 2D1418, 2006-10-23, kl. 14.00-18.00

Tentan består av 10 frågor. Totala poängsumman är 50 poäng. Inga hjälpmedel förutom boken *Speech and Language Processing* av Daniel Jurafsky & James H. Martin får användas. Eventuella bonuspoäng intjänade under 2006 kommer att adderas till dina tentapoäng. Betygsgränserna är preliminärt: 3: 25-34 poäng, 4: 35-44 poäng, 5: 45-50 poäng. Resultatet meddelas via Res och Studentexpeditionen inom tre veckor.

Exchange students are allowed to bring their bilingual lexicons and give answers and examples in English.

1. Stavningskontroll (6 poäng)

I riksdagsvalet 2026 vinner språkpolispartiet egen majoritet då väljarna tröttnat på förslappningen av språkbruket och alla anglicismer. Man genomför en språkreform och bestämmer att alla substantiv ska böjas enligt samma böjningsmönster så det blir ordning och reda (en bil, den bilen, flera bilar är som vanligt, medan bord blir ett bord, det borden, flera bordar). Däremot inför man några extra kasus, för att fjärma sig från engelskan. Rättstavningsprogrammet Stava måste nu uppdateras för den nya svenskan. Beskriv kortfattat vilka ändringar som behöver genomföras.

Fyra år senare har folk tröttnat på språkfascismen och röstar in ett anglofilt parti istället. Nu bestäms det att sammansättningar ska skrivas isär istället; är det gott nog för engelska så är det gott nog för oss, och de flesta gör ju dessutom redan fel ändå. Vilka ändringar kan det nu vara bra att göra i Stava?

2. Språkstatistik (6 poäng)

Så kallade kursledare på ett bluffuniversitet på Internet bestämmer sig för att det skulle se bra ut i reklamen med exempel på gamla tentor på hemsidan. Eftersom man varken kan eller orkar skriva egna tentor bestämmer man sig för att ladda ned tentor från fina universitet och med detta som träningskorpus statistiskt generera nya tentor till sin egen hemsida. Konstruera och beskriv en metod för att göra detta. Man vill använda så lite manuellt arbete som möjligt, och man har ingen större lingvistisk kompetens att tillgå. Däremot har man inget emot att använda färdiga programvaror om det skulle behövas. Beskriv även kortfattat problem, begränsningar och liknande med ditt förslag.

3. Maskininlärning (6 poäng)

Fjällföreningen Fjällhårdingarna har fått in för många ansökningar om medlemskap inför den nya säsongen. Man vill automatgallra så att alla "ovana" fjällvandrare sorteras bort automatiskt. Man tänker sig en maskininlärningslösning som läser av fritextdelen av ansökningarna som handlar om personernas fjällmeriter. Saker som "Jag gillar att göra små mysiga och lätta dagsturer från hotellet." ska gallras bort, men "Man kan räkna med att jag utsätter mig för livsfara åtminstone två gånger per ensamvandring och pannbensmarsch." ska vara kvar. Beskriv hur man kan gå tillväga för att konstruera ett system för detta ändamål. Använd viktiga ord som till exempel särdrag, träningsdata, överinlärning m.m. på ett sätt som visar att du vet mycket om maskininlärning i språkteknologisammanhang.

4. Informationssökning (5 poäng)

- a) Vad är term-dokument-matrisen? (1 poäng)
- b) Vad modellerar man med värdena som står i den? (1 poäng)
- c) Varför normaliserar man ofta kolumnerna i den? (1 poäng)
- d) Man normaliserar ofta något annat också. Vad? Vad menas med denna normalisering? Varför gör man den? (2 poäng)

Vänd!

5. Textklustring och textkategorisering (5 poäng)

Ett företag får massor av e-brev till sin kundtjänst. De har sparat alla manuellt i två mappar: en för positiva och en för negativa. Nu vill de ha en automatisk metod som lägger inkommande brev i rätt mapp.

- a) Ska de använda klustring eller kategorisering? Varför inte den ena eller den andra? (2 poäng)
- b) Förklara i stora drag hur det går till utgående från denna situation. (2 poäng)
- c) Vad skulle de kunna ha för nytta av den andra metoden? (1 poäng)

6. Textsammanfattning (4 poäng)

Du ska konstruera en automatisk textsammanfattare. Beskriv kortfattat minst två språkteknologiska verktyg, två statistiska metoder och en heuristisk tumregel (genrespecifik observation) som du skulle vilja inkludera i ditt system, samt vad de skulle vara användbara till i just textsammanfattning. Det står dig fritt att välja om du vill göra textextraktion eller textabstraktion, så länge du beskriver hur du skulle göra det automatiskt med en dator.

7. Utvärdering (6 poäng)

Du har en automatisk textsammanfattare som du vill utvärdera. För enkelhetens skull kan vi anta att du enbart ska utvärdera den på en text, som då följaktligen skall sammanfattas med ditt system. Hur skulle du gå till väga för att utvärdera denna textsammanfattare på denna enda text? Du ska alltså inte beskriva hur den automatiska sammanfattaren ifråga är uppbyggd eller hur den fungerar, bara hur själva utvärderingen av densamma ska gå tillväga. Tänk i termerna guldstandard, nedre gränser (baselines), övre gräns (human ceiling) och utvärderingsmetod(er). Tänk också på att det är svårt även för en människa att sammanfatta en text och hur detta påverkar din utvärdering.

8. Analys och generering (4 poäng)

Att analysera text har länge varit en central del inom språkteknologin. Språkgenerering har däremot hamnat i skymundan. Beskriv två tillämpningar och två språkliga nivåer (t.ex. morfologi, syntax, semantik, diskurs, pragmatik) där språkgenerering är nödvändigt om systemet skall vara användbart.

9. Språkteknologins mognad (4 poäng)

En del problem inom språkteknologin är mer eller mindre lösta medan andra problems lösning är avlägsna. Beskriv kortfattat ett problem som kan anses löst till den grad att det finns verkliga program som löser problemet så att det kan anses vara användbart. Beskriv också på ett kortfattat vis ett problem som vore önskvärt att lösa men som inget program idag kan lösa helt automatiskt.

10. Grammatikkontroll (4 poäng)

Inom området automatisk språkgranskning är det en ständig dragkamp mellan att hitta så många fel som möjligt och undvika att signalera korrekta konstruktioner som felaktiga. Beskriv denna problematik utifrån begreppen täckning (recall) och precision.