LUNDS UNIVERSITET
Lunds Tekniska Högskola

Kursplan för läsåret 2007/2008

SPRÅKBEHANDLING OCH DATALINGVISTIK EDA171

Language Processing and Computational Linguistics

Antal högskolepoäng: 7,5. **Betygskala:** UG. **Nivå:** A (Avancerad nivå). **Undervisningsspråk:** Kursen kan komma att ges på engelska. **Valfri för:** C4, D4, E4, F4, Pi4. **Kursansvarig:** Professor Pierre Nugues, Pierre.Nugues@cs.lth.se, Inst f datavetenskap. **Förkunskapskrav:** EDA027 Algoritmer och datastrukturer. **Begränsat antal platser:** Ja. **Urvalskriterier:** Av platserna går högst 6 till utbytesstudenter. Återstående 52 platser fördelas i proportion till antalet sökande vid LTH och Nat fak. Urvalsgrund för studenter vid LTH: 1. Antal poäng som uppnåtts vid LTH. 2. Antal poäng i kurser i Datavetenskap. **Prestationsbedömning:** Godkända inlämningsuppgifter, projekt och godkända uppsatser krävs för godkänt betyg på kursen. Inlämningsuppgifter och projekt utförs normalt i grupper om två eller tre studenter men kan även utföras individuellt. De fem inlämningsuppgifterna behandlas under sammanlagt 6 laborationstillfällen. Projektet handleds av en lärare och redovisas i slutet av kursen. Om projektet inte godkänns vid detta tillfälle ges en möjlighet att inom en månad förbättra resultatet och presentera på nytt. **Hemsida:** http://www.cs.lth.se/EDA171.

Syfte

Under de senaste 15 åren har de språkteknologiska metoderna mognat avsevärt på grund av att mängden tillgänglig skriven och talad information har ökat kraftigt, vilket har lett till ökande behov av att behandla den automatiskt. Fastän de flesta datorsystem inte enbart ägnar sig åt språkbehandling så finns det numera många applikationer som i någon mån är "språkmedvetna" och har inbyggda språkteknologiska funktioner såsom stavnings- och grammatikkontroll, sökning och sammanfattning av information eller ett talbaserat dialoggränssnitt. Detta gör att fältet är av ökande betydelse för CS-ingenjörer.

Kursen ger en inledning till språkteknologins metoder. Den försöker täcka hela fältet från teckenkodning och statistiska språkmodeller till syntaktisk analys, semantik och dialogsystem. Kursen inriktar sig på välbeprövade tillämpningar i industriell skala eller på försöksnivå.

Mål

Kunskap och förståelse
För godkänd kurs skall studenten

- Förstå fältet av språkteknologiska metoder och viktiga applikationer som använder dem
- Känna till de viktigaste teknikerna, grundläggande algoritmer och allmänna arkitekturer

som används i applikationer
- Skapa och konstruera språkbehandlingsalgoritmer. Skriva, tolka, utvärdera och förbättra dem.
- Utforma, utveckla, utvärdera och beskriva en fullständig språkteknologisk prototyp

Färdighet och förmåga
För godkänd kurs skall studenten

- Förstå och utveckla annoteringsscheman, skapa och bearbeta strukturerade dokument genom att använda XML
- Förstå och skriva reguljära uttryck och använda dem i programmeringsspråk som Perl eller Java
- Använda logik och logikprogrammeringsspråk som Prolog- eller beskrivningslogik
- Förstå och använda maskininlärningsalgoritmer och statistiska tekniker
- Utveckla och utvärdera algoritmer i de viktiga fälten av språkteknologier, språkmodellering, partiell parsning, dependensparsning, semantiska nät och dialog, genom att använda verkliga data
- Konstruera en färdig språkteknologisk prototyp i ett grupprojekt med två eller tre deltagare (eventuellt en) - Söka litteratur och elektroniskt publicerat material inom språkteknologiområdet
  - Utvärdera och reflektera över existerande system och forskning inom området
  - Skapa och implementera prototypen
  - Skriftligt och muntligt presentera projektet

Värderingsförmåga och förhållningssätt
För godkänd kurs skall studenten

- Visa nyfikenhet, kreativitet och förmåga till problemlösning
- Visa förståelse för industriella och forsknings relaterade frågor i språkteknologiområdet

Innehåll

- *En översikt över NLP*: Presentation av NLP, tillämpningar, lingvistikens delområden, exempel.
- *Korpuslingvistik*: Reguljära uttryck, ändliga automater, introduktion till Perl, konkordanser, tokenisering, frekvenslistor, kollokationer.
- *Morfologi och ordklasstaggning*: Morfologi, transduktorer, ordklasstaggning.
- *Frasstrukturgrammatiker i Prolog*: Konstituenter, syntaxträd, DCG-regler, variabler, syntaktisk parsning, semantisk parsning med kompositionell analys.
- *Partiell parsning och informationsextrahering*
- *Syntaktiska formalismer*: Konstituens och dependens, funktioner.
- *Parsning*: Chart parsing, statistisk parsning, dependensparsning.
- *Semantik*: Formell semantik, lambda-kalkyl, kompositionalitet; substantiv, verb, determinerare, ord och betydelse, lexikal semantik, kasusgrammatiker, semantiska grammatikor.
- *Diskurs och dialog*: diskurs och retorik, anaforer, dialogstruktur, RST (rhetorical structure theory), dialog, ändliga automater, adjacency pairs, talhandlingar, multimodalitet.
- *Översikt över talsyntes och röstigenkänning*.

Litteratur
Nugues Pierre, An Introduction to Language Processing with Perl and Prolog. An

Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German. Series: Cognitive Technologies, Springer Verlag, 2006, ISBN: 3-540-25031-X.
Rekommenderad referenslitteratur: Manning and Schütze: Foundations of Statistical Natural Language Processing, MIT Press 1999. ISBN: 0-262-13360-1.