



INSTITUTIONEN FÖR DATA- OCH INFORMATIONSTEKNIK

DIT245 Maskininlärning för språkteknologi, 7,5 högskolepoäng

Machine learning for natural language processing, 7.5 credits

Avancerad nivå / Second Cycle

Fastställande

Kursplanen är fastställd av Institutionen för data- och informationsteknik 2019-12-04 och senast reviderad 2019-12-04. Den reviderade kursplanen gäller från och med 2020-08-31, höstterminen 2020.

Utbildningsområde: Naturvetenskapligt 100 %

Ansvarig institution: Institutionen för data- och informationsteknik

Inplacering

Kursen kan ingå i följande program: 1) Matematiska vetenskaper, masterprogram (N2MAT) och 2) Applied Data Science masterprogram (N2ADS)

Huvudområde

Datavetenskap

Data Science

Fördjupning

A1F, Avancerad nivå, har kurs/er på avancerad nivå som förkunskapskrav

A1F, Avancerad nivå, har kurs/er på avancerad nivå som förkunskapskrav

Förkunskapskrav

För tillträde till kursen krävs att studenten har en examen på kandidatnivå inom något ämne.

Utöver detta krävs:

- 7,5 hp avklarade kurser i programmering eller motsvarande,
- en kurs i sannolikheteori och statistik, t.ex. DIT862 Statistiska metoder för Data Science eller MSG810 Matematisk statistik och diskret matematik,
- en första kurs i maskininlärning, t.ex. DIT866 Tillämpad maskininlärning, DIT381 Algoritmer för maskininlärning och inferens, eller MSA220 Statistisk slutledning för stora datamängder.

Följande kunskapsnivå i Engelska krävs; Engelska 6/Engelska B eller motsvarande fränett erkänt internationellt test, t.ex. TOELF, IELTS

Lärandemål

Efter godkänd kurs ska studenten kunna:

Kunskap och förståelse

- beskriva fundamentala metoder för lagring av textdata för världens språk,
- beskriva de vanligaste typerna av språkteknologiska uppgifter,
- beskriva de vanligaste typerna av maskininlärning som används inom modern språkteknologi,
- förklara hur textdata kan annoteras för en språkteknologisk uppgift inför användning av maskininlärning.

Färdigheter och förmåga

- tillämpa mjukvarubibliotek som använder sig av maskininlärningsmetoder för vanliga språkteknologiska uppgifter,
- skriva programkod för att implementera maskininlärningsmodeller för språkteknologi,
- tillämpa utvärderingsmetoder för att mäta kvaliteten hos språkteknologiska system.

Värderingsförmåga och förhållningssätt

- beskriva fördelar och begränsningar hos olika maskininlärningsmodeller med avseende på en given språkteknologisk uppgift,
- resonera om vilka typer av data som skulle kunna vara användbara för en given tillämpning inom språkteknologi,
- välja en lämplig utvärderingsmetod för ett språkteknologiskt system och motivera detta val,
- resonera om etiska frågor rörande språkteknologiska system som bygger på metoder inom maskininlärning, exempelvis stereotyper och underrepresentation.

Innehåll

Kursen ger en introduktion till maskininlärningsmodeller och arkitekturer som används i moderna system inom språkteknologi.

Den snabba utveckling som sker inom maskininlärning har revolutionerat det språkteknologiska området, inklusive för kommersiellt viktiga tillämpningar som översättning, sammanfattning och informationsextraktion. Språkliga data uppvisar ett antal egenheter som gör dem mer utmanande att arbeta med i jämförelse med andra typer av data som förekommer inom maskininlärning: naturligt språk är diskret,

strukturerat och mångtydigt. Det finns en enorm språklig variation: inte bara finns det tusentals språk i världen, men även inom varje språk finns det stora variationer i stil och genre. Många språkliga fenomen har en "långsvansad" statistisk fördelning, vilket gör det mer kostsamt att producera träningsdata. På grund av dessa anledningar är maskininlärningsarkitekturer för språk teknologiska tillämpningar ofta uppbyggda på ett annorlunda sätt än för tillämpningar i andra området.

Kursen täcker följande breda områden:

- Att arbeta praktiskt med textdata, inklusive fundamentala uppgifter som orduppdelning och ordräkning;
- probabilistiska språkmodeller, t.ex. ämnesmodeller;
- översikt över de vanligaste typerna av språk teknologiska tillämpningar;
- representationsarkitekturer i språk teknologiska modeller, t.ex. ordinbäddningar, faltningsnät, rekurrenta nät och *attention models*;
- maskininlärningsmodeller för de vanligaste typerna av språk teknologiska problem, framför allt kategorisering, sekvenstagning, strukturprediktion samt generering;
- överföringsinläring inom språk teknologi.

Delkurser

1. Projekt (Project), 7,5 hp

Betygsskala: Väl godkänd (VG), Godkänd (G) och Underkänd (U)

Former för undervisning

Föreläsningar, datorlaborationer.

Undervisningsspråk: engelska

Former för bedömning

Kursen examineras genom obligatoriska skriftliga inlämningsuppgifter som redovisas genom rapportinlämning, och ett självdefinierat projekt som redovisas genom rapportinlämning samt muntlig presentation. Av inlämningsuppgifterna genomförs vissa individuellt och andra i grupper om normalt 2-4 studenter. Projektet utförs i grupper om 2-4 studenter.

Försenad inlämning av inlämningsuppgifter eller projekt innebär att lösningen får betyget Underkänd (U), om inte särskilda skäl föreligger. Underkänd inlämningsuppgift eller projekt omexamineras vid ett senare kurstillfälle.

Om student som underkänts två gånger på samma examinerande moment önskar byte av examinator inför nästa examinationstillfälle, ska sådan begäran inlämnas skriftligt

till institutionen och bifallas om det inte finns särskilda skäl däremot (HF 6 kap § 22).

I det fall en kurs har upphört eller genomgått större förändringar ska student garanteras minst tre examinationstillfällen (inklusive ordinarie examinationstillfälle) under en tid av minst ett år, dock som längst två år efter det att kursen upphört/förändrats. Vad avser praktik och VFU gäller motsvarande, men med begränsning till endast ett ytterligare examinationstillfälle.

Betyg

På kursen ges något av betygen Väl godkänd (VG), Godkänd (G) och Underkänd (U). För att bli godkänd på kursen krävs minst betyg godkänt på alla inlämningsuppgifter samt projektet. För att få ett högre betyg än godkänt för hela kursen krävs ett högre viktat genomsnitt av betygen från inlämningsuppgifterna och projektet.

Kursvärdering

Kursen utvärderas genom möten, både under och efter kursen, mellan lärare och studentrepresentanter. Ett anonymt skriftligt frågeformulär skickas även ut till studenterna efter kursens slut. Resultaten av utvärderingarna används för att förbättra kursinnehållet och som indikation till vilka delar som skulle kunna läggas till, tas bort, förbättras eller ändras.

Resultatet och eventuella förändringar i kursens upplägg ska förmedlas både till de studenter som genomförde värderingen och till de studenter som ska påbörja kursen.

Övrigt

Kursen är samläst med Chalmers.

Kurslitteratur kommer att publiceras senast 8 veckor innan kursstart.