Text Mining

Programkurs

6 hp

Text Mining

TDDE16

Gäller från: 2018 VT

Fastställd av
Programnämnden för data- och medieteknik, DM

Fastställandedatum

Huvudområde
Informationsteknologi, Datateknik, Datavetenskap

Utbildningsnivå
Avancerad nivå

Fördjupningsnivå
A1X

Kursen ges för
- Datavetenskap, masterprogram
- Civilingenjör i datateknik
- Civilingenjör i informationsteknologi
- Civilingenjör i mjukvaruteknik
- Computer Science, masterprogram

Förkunskapskrav
OBS! Tillträdeskrav för icke programstudenter omfattar vanligen också tillträdeskrav för programmet och ev. tröskelkrav för progression inom programmet, eller motsvarande.

Rekommenderade förkunskaper
Matematisk analys; Linjär algebra; Sannolikhetslära och statistisk; Maskininlärning; Grundläggande programmering.

Lärandemål
Kurses övergripande mål är att ge en introduktion till kvantitativ analys av text, med speciell fokus på maskininlärningsmetoder för textdokument. Kursdeltagarna ska lära sig de huvudsakliga stegen i kvantitativ textanalys: effektiv utvinning av text, lingvistisk processing av text till en form som lämpar sig för statistiska maskininlärningsmetoder som bl.a. används för textprediktion.

Efter genomgången kurs ska den studerande kunna:

1. använda grundläggande metoder för information extraction och information retrieval av textuella data
2. använda textbehandlingsmetoder för att förbereda textdokument för statistisk modellering
3. använda relevanta maskininlärningsmetoder för textanalys och korrekt tolka resultaten från en sådan analys
4. använda maskininlärningsmodeller för textprediktion
5. utvärdera maskininlärningsmodeller för textanalys

Kursinnehåll

Introduktion och översikt av kvantitativ textanalys med tillämpningar. Informationsutvinning, Web crawling, Information retrieval, Tf-idf, Vektorrumsmodeller, Textbehandling, Bag of words, N-grams, Gleshet och utjämning för texttillämpningar, dokumentklassificering, sentiment analysis, Modelutvärdering, Topicmodeller.

Undervisnings- och arbetsformer

Undervisningen består av föreläsningar, datorlaborationer och ett individuellt projektarbete. Föreläsningar används för att introducera begrepp och teori som studenterna sedan använder i praktisk problemlösning vid och datorlaborationer och i projektarbetet.

Examination

PRA1	Projekt	3 hp	U, 3, 4, 5
LAB1	Datorlaborationer	3 hp	U, G

UPG1 består av datorlaborationer som prövar studenternas förmåga att omsätta teoretisk kunskap till praktisk problemlösning inom text mining.
PRA1 är ett individuellt projektarbete där kursdeltagaren löser ett verkligt problem med textanalys. Projektet dokumenteras och utvärderas i form av en skriftlig projektrapport.

Betygsskala

Fyrgradig skala, LiU, U, 3, 4, 5

Övrig information

Påbyggnadskurser

Språkteknologi

Institution

Institutionen för datavetenskap

Studierektor eller motsvarande

Ann-Charlotte Hallberg

Examinator
Marco Kuhlmann

Kurshemsida och andra länkar
http://www.ida.liu.se/~TDDE16/

Undervisningstid
Preliminär schemalagd tid: 28 h
Rekommenderad självstudietid: 132 h

Generella bestämmelser

Kursplan

För varje kurs finns en kursplan. I kursplanen anges kursens mål och innehåll samt de särskilda förkunskaper som erfordras för att den studerande skall kunna tillgodogöra sig undervisningen.

Schemaläggning

Schemaläggning av kurser görs efter, för kursen, beslutad blockindelning. För kurser med mindre än fem deltagare, och flertalet projektkurser läggs inget centralt schema.

Avbrott på kurs

Enligt rektors beslut om regler för registrering, avregistrering samt resultatrapportering (Dnr LiU-2015-01241) skall avbrott i studier registreras i Ladok. Alla studenter som inte deltar i kurs man registrerat sig på är alltså skyldiga att anmäla avbrottet så att kursregistreringen kan
tas bort. Avanmälan från kurs görs via webbformulär, www.lith.liu.se/for-studenter/kurskomplettering?l=sv.

Inställd kurs

Kurser med få deltagare (< 10) kan ställas in eller organiseras på annat sätt än vad som är angivet i kursplanen. Om kurs skall ställas in eller avvikelse från kursplanen skall ske prövas och beslutas detta av programnämnden.

Föreskrifter rörande examination och examinator

Se särskilt beslut i regelsamlingen:
http://styrdokument.liu.se/Regelsamling/VisaBeslut/622678

Examination

Tentamen

Skriftlig och muntlig tentamen ges minst tre gånger årligen; en gång omedelbart efter kursens slut, en gång i augustiperioden samt vanligtvis i en av omtentamensperioderna. Annan placering beslutas av programnämnden.

Principer för tentamensschemat för kurser som följer läsperioderna:

- kurser som ges Vt1 förstagångstenteras i mars och omtenteras i juni och i augusti
- kurser som ges Vt2 förstagångstenteras i maj och omtenteras i augusti och i oktober
- kurser som ges Ht1 förstagångstenteras i oktober och omtenteras i januari

och augusti
- kurser som ges Ht2 förstagångstenteras i januari och omtenteras i påsk och i augusti

Tentamensschemat utgår från blockindelningen men avvikelser kan förekomma främst för kurser som samläses/samtenteras av flera program samt i lägre årskurs.

- För kurser som av programnämnden beslutats vara vartannatårskurser ges tentamina 3 gånger endast under det år kursen ges.
- För kurser som flyttas eller ställs in så att de ej ges under något eller några år ges tentamina 3 gånger under det närmast följande året med tentamenstillfällen motsvarande dem som gällde före flyttningen av kursen.
- Har undervisningen upphört i en kurs ges under det närmast följande året tre tentamina samtidigt som tentamen ges i eventuell ersättningskurs, alternativt i samband med andra omtentamina. Dessutom ges tentamen ytterligare en gång under det därpå följande året om inte programnämnden föreskriver annat.
- Om en kurs ges i flera perioder under året (för program eller vid skilda tillfällen för olika program) beslutar programnämnden/programnämnderna gemensamt om placeringen av och antalet omtentamina.

Anmälan till tentamen

För deltagande i tentamina krävs att den studerande gjort förhandsanmälan i Studentportalen under anmälningsperioden, dvs tidigast 30 dagar och senast 10 dagar före tentamensdagen. Anvisad sal meddelas fyra dagar före tentamensdagen via e-post. Studerande, som inte förhandsanmält sitt deltagande riskerar att avvisas om plats inte finns inom ramen för tillgängliga skrivningsplatser.

Teckenförklaring till tentaanmälningssystemet:
** markerar att tentan ges för näst sista gången
* markerar att tentan ges för sista gången

Ordningsföreskrifter för studerande vid tentamensskrivningar

Se särskilt beslut i regelsamlingen: http://styrdokument.liu.se/Regelsamling/VisaBeslut/622682

Plussning

Vid Tekniska högskolan vid LiU har studerande rätt att genomgå förnyat prov för högre betyg på skriftliga tentamina samt datortentamina, dvs samtliga provmoment med kod TEN och DAT. På övriga examinationsmoment ges inte möjlighet till plussning, om inget annat anges i kursplan.

Regler för omprov

För regler för omprov vid andra examinationsformer än skriftliga tentamina och datortentamina hänvisas till LiU-föreskrifterna för examination och examinator,

http://styrdokument.liu.se/Regelsamling/VisaBeslut/622678.

Plagiering

Vid examination som innebär rapportskrivande och där studenten kan antas ha tillgång till andras källor (exempelvis vid självständiga arbeten, uppsatser etc) måste inlämnat material utformas i enlighet med god sed för källhänvisning (referenser eller citat med angivande av källa) vad gäller användning av andras text, bilder, idéer, data etc. Det ska även framgå ifall författaren återbrukat egen text, bilder, idéer, data etc från tidigare genomförd examination.

Underlåtelse att ange sådana källor kan betraktas som försök till vilseledande vid examination.

Försök till vilseledande

Vid grundad misstanke om att en student försökt vilseleda vid examination eller när en studieprestation ska bedömas ska enligt Högskoleförordningens 10 kapitel examinator anmäla det vidare till universitetets disciplinnämnd. Möjliga konsekvenser för den studerande är en avstängning från studierna eller en varning. För mer information
se https://www.student.liu.se/studenttjanster/lagar-regler-rattigheter?l=sv.

Betyg

Företrädesvis skall betygen underkänd (U), godkänd (3), icke utan beröm godkänd (4) och med beröm godkänd (5) användas. Kurser som styrs av tekniska fakultetsstyrelsen fastställt tentamensschema skall därvid särskilt beaktas.

1. Kurser med skriftlig tentamen skall ge betygen (U, 3, 4, 5).
2. Kurser med stor del tillämpningsinriktade moment såsom laborationer, projekt eller grupparbeten får ges betygen underkänd (U) eller godkänd (G).

Examinationsmoment

1. Skriftlig tentamen (TEN) skall ge betyg (U, 3, 4, 5).
2. Examensarbete samt självständigt arbete ger betyg underkänd (U) eller godkänd (G).
3. Examinationsmoment som kan ge betygen underkänd (U) eller godkänd (G) är laboration (LAB), projekt (PRA), kontrollskrivning (KTR), muntlig tentamen (MUN), datortentamen (DAT), uppgift (UPG), hemtentamina (HEM).
4. Övriga examinationsmoment där examinationen uppfylls framför allt genom aktiv närvaro som annat (ANN), basgrupp (BAS) eller moment (MOM) ger betygen underkänd (U) eller godkänd (G).

Rapportering av den studerandes examinationsresultat sker på respektive institution.

Regler

Universitetet är en statlig myndighet vars verksamhet regleras av lagar och förordningar, exempelvis Högskolelagen och Högskoleförordningen. Förutom lagar och förordningar styrs verksamheten av ett antal styrdokument. I Linköpings universitets egna regelverk samlas gällande beslut av regelkaraktär som fattats av universitetsstyrelse, rektor samt fakultets- och områdesstyrelser.

LiU:s regelsamling angående utbildning på grund- och avancerad nivå nås på http://styrdokument.liu.se/Regelsamling/Innehall/Utbildning_pa_grund-_och_avancerad_niva.