Ekonomihögskolan

STAN49, Statistik: Analys av textdata, 7,5 högskolepoäng
Statistics: Analysis of Textual Data, 7.5 credits
Avancerad nivå / Second Cycle

Fastställande

Kursplanen är fastställd av Institutionsstyrelsen vid Statistiska institutionen 2021-11-29 att gälla från och med 2022-08-29, höstterminen 2022.

Allmänna uppgifter

Kurs på avancerad nivå som ingår som valfri kurs i en magister- eller masterexamen i statistik. Kursen kan även läsas som fristående eller inom andra magister- och masterprogram vid Lunds universitet.

Undervisningsspråk: Engelska

Huvudområde	*Fördjupning*
Statistik	A1F, Avancerad nivå, har kurs/er på avancerad nivå som förkunskapskrav

Kursens mål

Kunskap och förståelse
För godkänd kurs skall studenten

- kunna beskriva olika tekniker för förbehandling av textdata och kunna förklara när och varför dessa ska användas,
- kunna redogöra för olika sätt att representera textdata som vektorer,
- kunna förklara olika tekniker för klassificering och klustring av textdata,
- kunna förklara olika tekniker för ämnesmodellering och attitydanalys, och
- kunna förklara olika tekniker för informationsextraktion och textsammanfattning.

Färdighet och förmåga
För godkänd kurs skall studenten

- kunna tillämpa tekniker för att lösa typer olika av textdataanalysproblem,
- självständigt kunna identifiera och formulera en frågeställning relaterad till

textdata samt kunna presentera en lösning tilll denna, och
- skriftligt kunna redogöra klart för och diskutera sina slutsatser av olika textanalysproblem.

Värderingsförmåga och förhållningssätt
För godkänd kurs skall studenten

- kunna göra bedömningar av modellval utifrån frågeställning samt tillgängliga data och beräkningskapacitet.

Kursens innehåll

Kursen ger en introduktion till statistisk analys av text. Följande ämnen behandlas:

- Förbehandling av textdata
- Textrepresentation
- Textklassificering
- Klusteranalys av textdata
- Ämnesmodellering
- Attitydanalys
- Textsammanfattning

På kursen kommer både metoder som bygger på både klassiska statistiska ansatser (inklusive Bayesianska modeller) och moderna ansatser som djupinlärning och *recurrent neural networks* att presenteras.

Kursens genomförande

Kursen genomförs som en serie föreläsningar, datorövningar och seminarier. Kamratgranskning av andra studenters inlämningsuppgifter är en viktig och obligatorisk del av kursen.

Kursens examination

Examinationen utgörs av quiz och inlämningsuppgifter samt kamratgranskning av inlämningsuppgifterna. Slutbetyget på kursen bestäms som en sammanvägning av resultaten på quizen (33 %) och inlämningsuppgifterna (67 %).

Lunds universitet ser mycket allvarligt på fusk och kommer att vidta disciplinåtgärder mot alla slags försök till fusk i samband med tentamina och andra examinationsformer. Plagiering betraktas som ett mycket allvarligt akademiskt brott. Det straff som universitets disciplinnämnd kan utdela för detta, och för andra slags fusk i samband med olika former av examination, inkluderar avstängning från universitetet under en viss tidsperiod.

Om så krävs för att en student med varaktig funktionsnedsättning ska ges ett likvärdigt examinationsalternativ jämfört med en student utan funktionsnedsättning, så kan examinator efter samråd med universitetets avdelning för pedagogiskt stöd fatta beslut om alternativ examinationsform för berörd student.

Prov/moment för denna kurs finns i en bilaga i slutet av dokumentet.

Betyg

Betygsskalan omfattar betygsgraderna Underkänt, E, D, C, B, A.

A (Utmärkt) 85-100 poäng/procent. Ett framstående resultat som är utmärkt vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet.

B (Mycket bra) 75-84 poäng/procent. Ett mycket bra resultat som karakteriseras av mycket bra teoretiskt djup, praktisk relevans, analytisk förmåga samt självständighet.

C (Bra) 65-74 poäng/procent. Ett bra resultat som karakteriseras av bra teoretiskt djup, praktisk relevans, analytisk förmåga samt självständighet.

D (Tillfredsställande) 55-64 poäng/procent. Ett resultat som är tillfredsställande vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet.

E (Tillräckligt) 50-54 poäng/procent. Ett resultat som möter minimikraven vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet, men inte mer.

U (Otillräckligt/Underkänt) 0-49 poäng/procent. Ett resultat som är otillräckligt vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet.

För att få godkänt på en kurs måste studenten få betyg E eller högre.

Förkunskapskrav

STAN48 Statistik: Avancerad statistisk programmering samt STAN52 Statistik: Avancerade maskininlärning, eller motsvarande.

Prov/moment för kursen STAN49, Statistik: Analys av textdata

Gäller från V23

2301 Quiz, 2,5 hp
Betygsskala: Underkänd, Godkänd
2302 Inlämningsuppgifter, 5,0 hp
Betygsskala: Underkänd, Godkänd