ID2211 Datautvinning, grund-kurs 7,5 hp

Data Mining, Basic Course

När kurs inte längre ges har student möjlighet att examineras under ytterligare två läsår.

Fastställande

Kursplan för ID2211 gäller från och med VT19

Betygsskala

A, B, C, D, E, FX, F

Utbildningsnivå

Avancerad nivå

Huvudområden

Datalogi och datateknik

Särskild behörighet

Undervisningsspråk

Undervisningsspråk anges i kurstillfällesinformationen i kurs- och programkatalogen.

Lärandemål

I kursen studeras grunderna i datautvinning inkluderande Informationsnätverksanalys samt även grundläggande tekniker för utvinning och analys av textdata i naturligt språk.

Specifikt täcker kursen grunderna i grafteori, nätverksstruktur och länkanalys samt även grunderna i utvinning och analys av text i naturligt språk.

Efter denna kurs kan studenten utvinna och analysera informationsnätverk och texter i naturligt språk. Speciellt ska studenten kunna

- sammanfatta och beskriva de fundamentala begreppen i grafteori och tillämpa dem i praktiken för grafanalys
- sammanfatta och beskriva de fundamentala principerna i analys av naturligt språk och tillämpa dem i praktiken för att utvinna information ur texter
- elaborera runt och tillämpa algoritmer for massivt länkade dataproblem (till exempel grafklustring, identifiering av "communities" etcetera).

Kursinnehåll

- Grundläggande definitioner inom grafteori, starka och svaga band, graddistribution och klustringsmått.
- Erdos-Renyi, Wats-Strogatz, konfigureringsmodell, effekten av en "liten värld".
- Slumpmässig grafvandring, Page Rank.
- Kaskadformat beteende, epidemisk spridning.
- Algoritmen "Label Propagation", länkprediktion.
- Distributiv semantik, ordinbäddningar, sentimentanalys.
- Ämnesmodellering, documentsammanfattning, textsegmenteringsinlärning.

Kurslitteratur

Kursinnehållet är hämtat från följande läroböcker samt även från ett antal forskningsartiklar:

- John Hopcroft and Ravindran Kanna ”Foundations of Data Science” (2013).
- David Easley and Jon Kleinberg “Networks, Crowds, and Markets: Reasoning About a Highly Connected World” (2010).
- A. Rajaraman and J. D. Ullman, Mining of massive datasets. Cambridge University Press, 2012 (alternative: J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3-rd Ed., Morgan Kaufmann, 2012).

Examination

- PRO1 - Projekt, 3,0 hp, betygsskala: P, F
- TEN1 - Tentamen, 4,5 hp, betygsskala: A, B, C, D, E, FX, F

Examinator beslutar, baserat på rekommendation från KTH:s handläggare av stöd till studenter med funktionsnedsättning, om eventuell anpassad examination för studenter med

dokumenterad, varaktig funktionsnedsättning.

Examinator får medge annan examinationsform vid omexamination av enstaka studenter.

Etiskt förhållningssätt

- Vid grupparbete har alla i gruppen ansvar för gruppens arbete.
- Vid examination ska varje student ärligt redovisa hjälp som erhållits och källor som använts.
- Vid muntlig examination ska varje student kunna redogöra för hela uppgiften och hela lösningen.