



Introduktion till text mining

Introduction to text mining

7,5 högskolepoäng

7,5 credits

Ladokkod: C3IIT1

Version: 1.0

Fastställd av: Utskottet för utbildningar inom bibliotek, information och IT 2017-11-07

Gäller från: VT 2018

Nivå: Avancerad nivå

Huvudområde (successiv fördjupning): Biblioteks- och informationsvetenskap (A1N)

Utbildningsområde: Övrigt

Ämnesgrupp: Biblioteks- och informationsvetenskap

Förkunskapskrav: Kandidatexamen samt minst betyget Godkänt på delmomentet *Hemtamen 1: Klassifikation (IK11)* i kursen *Introduktion till kunskapsorganisation (32IIK1)* och delmomentet *Inlämningsuppgift: SKOS (IN01)* i kursen *XML och kontrollerade vokabulärer (32IXK1)*.

Betygsskala: Underkänd, Godkänd eller Väl godkänd

Innehåll

Kursen ger en introduktion till maskininlärning och s k "text mining" med särskild hänsyn tagen till klassifikation och klustring av text. Studenten förväntas utveckla en förståelse för och erfarenhet av strategiska val i förberedelsen för klassifikation, såsom tokenisering och val av särdrag, samt av olika modeller och algoritmer för textkategorisering och efterföljande evalueringsmetoder. Ett visst utrymme lämnas också för studier av olika visualiseringsmetoder.

Mål

Efter avslutad kurs ska studenten kunna, med avseende på:

Kunskap och förståelse

- 1.1 redogöra för olika valmöjligheter i och metoder för särdragsextraktion med hänsyn till syftet med kategoriseringen av text
- 1.2 redogöra för huvudprinciperna bakom de dominerande modellerna för klassifikation och klustring och karaktärisera vad som skiljer dem åt
- 1.3 redogöra för skillnader mellan och likheter i algoritmisk och mänsklig textkategorisering

Färdighet och förmåga

- 2.1 förbereda en samling texter för klassifikation och klustring med hjälp av därför avsedd programvara
- 2.2 tillämpa och utvärdera algoritmer för klassifikation med avseende på text
- 2.3 visualisera egenskaper hos en text, en samling av texter och/eller resultatet av en kategorisering med hjälp av tillgänglig programvara

Värderingsförmåga och förhållningssätt

- 3.1 värdera maskininlärningsmodeller och val av särdrag med hänsyn till syftet med kategorisering av text
- 3.2 identifiera algoritmernas begränsningar

Undervisningsformer

Undervisningen bedrivs i form av laborationer, föreläsningar, praktiska genomgångar och övningar.

Undervisningen bedrivs helt eller delvis på engelska

Examinationsformer

Kursen examineras genom följande examinationsmoment:

Hemtentamen

Lärandemål: 1.1-1.3, 3.1-3.2

Högskolepoäng: 4,5

Betygsskala: Underkänt, Godkänt eller Väl godkänt

Projektrapport

Lärandemål: 1.1, 2.1-2.3, 3.1

Högskolepoäng: 3

Betygsskala: Underkänt, Godkänt eller Väl godkänt

För betyget Godkänt på hel kurs krävs minst godkänt betyg på båda momenten. För betyget Väl godkänt gäller följande: Betyget Godkänt på Projektrapport ger värdet 0,6 medan Väl godkänt ger 1.0. Poängen på Hemtentamen normaliseras med avseende på maxpoäng och ger ett värde mellan 0 och 1. Värdena för Hemtentamen respektive Projektrapport normaliseras därefter med avseende på andelen kurspoäng och summeras vilket ger ett slutvärde mellan 0 och 1. För betyget Väl godkänt krävs 0,75 eller mer.

$$\text{Totalvärde} = ((\text{HemtentamensPoäng}/\text{Maxpoäng}) * 4,5 + (\text{ProjektrapportVärde} * 3)) / 7,5$$

Då kursplanen ändras kommer student som önskar slutföra rester från ett kurstillfälle att examineras utifrån kursens nya innehåll och upplägg. Då kursen har upphört kan student som önskar slutföra rester följa hela eller delar av annan likvärdig kurs.

Studentens rättigheter och skyldigheter vid examination är enligt riktlinjer och regelverk vid Högskolan i Borås.

Kurslitteratur och övriga läromedel

Kurslitteraturen är på engelska

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search*. Addison-Wesley: Harlow. Kap. 2, s. 21-56; Kap. 8, s. 281-336.

Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press. Kap. 8. (26 s.) [Tillgänglig elektroniskt]

Sebastiani, S. (2005). Text categorization. I A. Zanasi (Ed.), *Text mining and its applications* (s. 109-129). Southampton: WIT Press. [Tillgänglig elektroniskt]

Stavrianou, A., Andritsos, P. & Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *SIGMOD Record*, 36(3), 23-33.

Witten, I. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd. ed.). Amsterdam: Kaufman. (S. 3-17)

Dokumentation av och användarhandledning för den programvara som tillämpas i kursen tillkommer i en utsträckning som beror av studentens egna uppkomna behov.

Studentinflytande och utvärdering

Kursen utvärderas i enlighet med gällande riktlinjer för kursvärderingar vid Högskolan i Borås, där studenternas synpunkter ska inhämtas. Kursutvärderingsrapporten publiceras och återkopplas till deltagande och blivande studenter i enlighet med ovan nämnda riktlinjer, och ligger till grund för framtida utveckling av kurser och utbildningsprogram. Kursansvarig lärare ansvarar för att utvärdering enligt ovan genomförs.

Övrigt

För kursen rekommenderas kunskaper i matematik motsvarande Gymnasieskolans Matematik kurs C.

Kursen ingår i Masterprogram i Biblioteks- och informationsvetenskap, distansutbildning.