SIM

Avancerade metoder för automatisk kodning

Avancerade metoder för automatisk kodning förekommer främst internationellt, men även i någon mån på SCB. Metoderna går ut på att kunna sätta koder även då benämningarna inte direktmatchar, utan överensstämmer ungefärligen. En sådan inexakt matchning bygger ofta på empiriska data och annan tilläggsinformation, ofta kombinerade med sannolikheter.

På senare år har metoder för textkategorisering med övervakad maskininlärning blivit populärt i kodningssammanhang. Metoderna går ut på att modellera sambandet mellan benämningar och beskrivningar och kod utifrån tidigare observerade data, för att sedan klassificera nya observationer. Supportvektormaskiner (SVM) och k-närmsta grannar (k-nearest neighbour, kNN) är exempel på modeller. Arbete med att ta fram metoder och verktyg för automatisk kodning med maskininlärning pågår.

En vedertagen metod är N-gram-metoden, som bygger på att alla bokstavsföljder med N bokstäver (i rad) bildas för den aktuella benämningen som ska kodas. Mängden av dessa bokstavsföljder jämförs sedan med mängder av följder för benämningar samlade i ett lexikon. Vanligen utnyttjas överlappande bokstavsföljder för bigram, där N = 2. Då är andra bokstaven i första följden lika med första bokstaven i andra följden o.s.v. Väsentligt vid tillämpningen av denna metod är vilket mått som används vid den maskinella jämförelsen. N-gram-metoden kan även användas för att korrigera vanliga stavfel före en direktmatchning. Denna metod används i nuläget inte på SCB.