EXAMENSARBETE Investigating Hybrid Approaches for Name Matching of Points of Interest
STUDENT Lucy Albinsson, Tove Sölve
HANDLEDARE Dennis Medved (LTH), Hampus Londögård (AFRY)
EXAMINATOR Jacek Malec (LTH)

Ser kartan dubbelt?

POPULÄRVETENSKAPLIG SAMMANFATTNING **Lucy Albinsson, Tove Sölve**

För att karttjänster ska vara pålitliga är det viktigt att kunna avgöra om två närliggande intressepunkter med liknande namn är samma plats eller inte. Examensarbetet undersöker metoder för att avgöra detta genom att titta på olika aspekter av likheter mellan namnen.

Karttjänster har blivit en viktig del för att underlätta människors vardag. För att kartan ska vara användbar och representera den verkliga världen krävs det stora mängder geografisk information. Därför används ofta flera olika källor och att kombinera dessa kan leda till utmaningar. Tänk dig att du är på väg till universitetssjukhuset i Lund och tar fram din karttjänst för att hitta till rätt avdelning. Kartan visar två markörer som båda verkar representera avdelningen du letar efter, trots att de har något olika namn och ligger i olika delar av byggnaden. Att karttjänsten visar dubbletter av en och samma plats, är ett av problemen som kan uppstå när information hämtas från flera källor. Det finns ingen försäkran om att en plats har samma namn och geografisk information i olika databaser. Likaså är det svårt att avgöra om två närliggande platser med liknande namn faktiskt är olika, vilket gör det svårt att slå ihop rätt information.

I vårt examensarbete har vi undersökt om det går att avgöra om två närliggande platser är samma eller inte utifrån deras namn och geografiska position. Vi undersöker två aspekter av likhet mellan text, dels struktur och förekomst av tecken och dels semantisk betydelse. Våra metoder är baserade på kombinationer av traditionella algoritmer, vektorrepresentationer av semantisk likhet och maskininlärning.

Resultaten visar att det är positivt att kombinera algoritmer då de kan täcka upp för varandras brister. Dock visade det sig svårt att dra nytta av den semantiska likhethet mellan namnen eftersom att det ofta saknas sammanhang i namn. Maskininlärningsmetoderna var baserade på flera av de kombinerade algoritmerna och visade sig ge bäst resultat. Överlag visade resultaten att det kan vara positivt att kombinera metoder som tar hänsyn till olika aspekter av likhet för att avgöra om två intressepunkter representerar samma plats.