

Igenkänning av metadata

Igenkänningen av metadata består av tre separata delar: igenkänning av namnentiteter, automatisk ämnesordsindexering och igenkänning av dokumentets språk.

Igenkänning av namnentiteter

Bakgrund, fördelar och ändamål

Med namnentiteter är det möjligt att kombinera och söka efter arkivenheter och dokument som rör olika ämnen. Då kan slutanvändaren även hitta material som hen inte ursprungligen hade förmått söka.

Komponenten använder en finskspråkig eller engelskspråkig modell för igenkänning av namnentiteter baserat på det automatiskt igenkända språket i dokumentets textinnehåll. Den finskspråkiga modellen bygger på den finska versionen av BERT-språkmodellen (<https://github.com/TurkuNLP/FinBERT>) och har utbildats och utvecklats själv, medan den engelska igenkänningen av namnentiteter bygger på en app som tillhandahålls av Spacy-biblioteket (https://spacy.io/models/en#en_core_web_trf). De entitetskategorier som den finskspråkiga modellen har känt igen är

- datum: t.ex. 1.10.2016, 1 oktober, under 2016, på 1980-talet
- organisationer: t.ex. Apple, Åbo universitet, Samlingspartiet, Centralkriminalpolisen
- personnamn: t.ex. Sauli Niinistö, Julgubben, @digikim
- diarienummer: t.ex. SRK/123/45/1999, 1/23/56, 1000-1100/123/1988
- geopolitiska ortnamn: t.ex. Päijänne-Tavastland, Helsingfors, H:fors, Skatudden, Lappland
- andra ortnamn (ej geopol.): t.ex. Yosemite nationalpark, Mars, Atlanten, Kemi älv
- produkter: t.ex. iPhone 6, C++, Helsingin Sanomat, Patriot Act-lagen
- händelser: t.ex. Mobile World Event, Andra världskriget, covid-19
- FO-nummer: se <https://www.ytj.fi/sv/index/y-tunnus.html>
- nationaliteter, religiösa och politiska grupper: t.ex. finländare, finländshet, finskspråkiga, muslimer, Extinction Rebellion

Med undantag för diarienummer och FO-nummer känner den engelska modellen igen samma kategorier, fastän definitionen av innehållet i kategorierna inte i alla situationer helt motsvarar det ovannämnda.

Vad gör komponenten?

Med hjälp av igenkänningen av maskinskriven text (OCR) får komponenten ett igenkänt textinnehåll som indata från det digitaliserade dokumentet, där den söker metadata i de kategorier som anges ovan. Komponentens återställer en klassspecifik lista över metadata som hittats i dokumentet och från vilka eventuella upprepade uttryck i helt identisk form har raderats (t.ex. även om ordet Helsingfors upprepas flera gånger i texten, förekommer det endast en gång i resultatet).

Hur används komponenten?

Komponenten finns tillgänglig både som en del av Arkkiivi-gränssnittet (<http://www.arkkiivi.fi/>) och som en fristående app som tillhandahåller ett programmeringsgränssnitt (API) på DALAI-projektets GitHub-webbplats (https://github.com/DALAI-project/NER_API). Modellfilen kan laddas ner på webbplatsen HuggingFace (<https://huggingface.co/Kansallisarkisto/finbert-ner>). Närmare anvisningar om användningen av komponenten inom olika miljöer finns på den webbplatsen.

Utbildning av komponenten

För utbildningen av komponenten användes såväl Turku OntoNotes Entities Corpus-materialet (<https://github.com/TurkuNLP/turku-one>), den finskspråkiga delen av NewsEye-materialet (<https://zenodo.org/record/4694466#.YJR20qE6-bi>) och materialet som har sammanställts och annoterats från handlingar som Riksarkivet har digitaliserat. Det självproducerade utbildningsmaterialet har först körts via igenkänningen av maskinskriven text (OCR), varefter namnentiteter som ingår i kategorierna ovan har annoterats från texten manuellt. Totalt innehåller komponentens utbildningsdata cirka 105 000 entiteter.

Mängden och den tidsmässiga omfattningen av det material som använts i utbildningen (från mitten av 1800-talet till i dag) är begränsad, vilket vid sidan av kvaliteten på textigenkänningen bidrar till att det också uppstår fel i komponentens igenkänning.

Ämnesordsindexering

Bakgrund, fördelar och ändamål

Komponenten baserar sig på Annif-programvaran (<https://annif.org/>) som utvecklats av Nationalbiblioteket och som är föremål för kontinuerlig och aktiv utveckling. Ämnesordsindexeringen används för att söka information utifrån fyrkantstaggat utan att hela dokumentet behöver gås igenom. På så sätt kan man till exempel på ett effektivt sätt hitta dokument som innehåller ett utvalt ämnesord. Annif utvecklades ursprungligen för ämnesordsindexering av olika examensarbeten och vetenskapliga artiklar, men resultaten visar att det kan användas relativt allmänt. Annif-programvaran används av t.ex. Rundradion och Tyska nationalbiblioteket.

Vad gör komponenten?

Komponenten samlar ihop olika modeller för ämnesordsindexering när den använder Annif, som bygger på lösningar för maskininlärning och språkteknik. Slutresultatet är en beskrivning av dokumentet efter ämnesord.

Filformat och eventuella förbehandlings

Komponenten finns tillgänglig som en del av Arkkiivi-gränssnittet (<http://www.arkkiivi.fi/>), där följande filformat som stöds för närvarande är: .pdf, .jpg, .tif, .tiff, .xml och .txt. Om en fil tolkas som digitalt skapad, genomgår den inte en igenkänning av maskinskriven text (OCR). Om det är fråga om en bildfil, utförs OCR-behandlingen med programvaran Apache Tika (<https://tika.apache.org/>), eftersom Annif endast kan tolka textinnehåll.

Resultat

Komponenten återställer X ämnesord (10 som standard) som data i .json-format, som kan exporteras från gränssnittet även som en fil i .csv-format.

Material som använts i utbildningen av komponenten

I utbildningen av Annif har följande ordförråd använts: YSO-suomi, YSO-english och ALLFO-svenska.

Igenkänning av språk

Igenkänningen av språket i dokument bygger på det textinnehåll som textigenkänningen resulterar i. För igenkänning av språk används en funktion som har tagits fram för programvaran Apache Tika (<https://tika.apache.org/>).