

Rezultat D3.4: Demonstracijska aplikacija za semantično iskanje

Vodilni partner: **Semantika, d.o.o.**

Sodelujoči partnerji:

Fakulteta za računalništvo in informatiko UL

Inštitut za novejšo zgodovino

Avgust 2025

Uvod

Programska oprema GalisOnline je celovit dokumentacijski sistem za muzeje in arhive, ki že vrsto let omogoča napredno upravljanje in iskanje podatkov o predmetih, vendar se je doslej zanašala predvsem na besedilne vnose in ročno dodane metapodatke. Velik izziv pri digitalizaciji kulturne dediščine predstavljajo slikovni materiali (npr. skenirani arhivski dokumenti, fotografije napisov ali časopisov), iz katerih besedilo ni neposredno dostopno za iskanje. V okviru raziskovalno-razvojnega projekta PoVejMo ("Prilagodljiva obdelava naravnega jezika z velikimi jezikovnimi modeli") smo zato razvili novo storitev optičnega prepoznavanja znakov (OCR), ki je posebej prilagojena za slovenski jezik in integrirana v GalisOnline. S tem smo nadgradili obstoječo digitalizacijsko infrastrukturo - uporabnikom omogočamo avtomatsko pretvorbo besedila iz slik v računalniško berljivo obliko, kar odpira možnost iskanja po vsebini slikovnih gradiv in bogatenja metapodatkov zbirk. Nova rešitev, ki jo opisujemo v tem poročilu, je bila v sklopu projekta že preizkušena v realnem okolju GalisOnline (dosežena tehnološka zrelost TRL 6) in predstavlja pomemben korak k prijaznejšemu dostopu do informacij v slovenskem kulturnem in arhivskem gradivu.

Cilj C3.2- OCR za slovensko gradivo

Projekt PoVejMo je usmerjen v raziskave in razvoj novih rešitev na področju umetne inteligence in jezikovnih tehnologij s poudarkom na slovenščini. Znotraj projekta poteka več delovnih sklopov med katerimi je delovni sklop C3.2 osredotočen prav na nadgradnjo digitalizacijske infrastrukture z optičnim prepoznavanjem znakov. Glavni cilj sklopa C3.2 je razviti in integrirati storitev OCR, ki bo prilagojena slovenskemu jeziku ter izboljšana s postopkom kontekstno občutljivega črkovanja z uporabo modela *Micka*. To pomeni, da se iz izvornih slik (npr. skenov dokumentov, fotografij napisov ipd.) samodejno izlušči besedilo, nato pa sistem s pomočjo jezikovnega modela po trebi popravi prepoznane besede glede na kontekst in stopnjo zaupanja prepoznave - s tem se odpravijo morebitne OCR napake in starinske oblike zapisa pretvorijo v sodobno obliko. Rezultat C3.2 neposredno podpira širši cilj projekta PoVejMo, da z uporabo velikih jezikovnih modelov izboljša dostopnost in uporabnost slovenskih besedilnih virov v digitalni obliki.

V konzorciju PoVejMo je podjetje Semantika d.o.o. (razvijalec GalisOnline) vodilni partner za ta delovni sklop, v sodelovanju z raziskovalci z UL FRI in strokovnjaki z Inštituta za novejšo zgodovino. Skupaj smo zasnovali rešitev, ki muzejem, arhivom in drugim uporabnikom omogoča, da obogatijo obstoječe podatkovne zbirke z avtomatsko prepoznanim besedilom iz slikovnega gradiva. To dopolnjuje tradicionalno katalogizacijo - namesto da bi se morali zanesti le na ročne prepise ali opombe, lahko sedaj sistem sam prebere besedilo z digitalnih slik in ga priloži predmetom v bazi. S tem projekt naslavlja eno od pomembnih področij (digitalizacija in OCR) v sklopu širših prizadevanj PoVejMo, ki vključujejo še vrsto drugih demonstratorjev (od dialogov in ukazov v slovenščini do semantičnega iskanja, strojnega prevajanja in govornega vmesnika).

Razvoj rešitve in uporabljene tehnologije

Pri razvoju OCR rešitve smo izhajali iz uveljavljenega odprtokodnega orodja PaddleOCR (različica PP-OCRv5), ki omogoča tako zaznavo besedila na sliki kot prepoznavo znakov. Za potrebe slovenskega jezika smo model dodatno prilagodili - uporabili smo osnovni model PP-OCRv5 in ga dotrenirali na približno 500 primerih slovenskega besedila iz dejanskih gradiv Semantike. Ti primeri so vključevali različne vrste slik (npr. skenirane tipe pisav, starinske tiskovine, tabelne dokumente ipd.) skupaj s pripadajočimi preverjenimi prepisi. S pripravo takšne učne množice (vsaka enota je vsebovala izvorno sliko, prvotni izpis osnovnega OCR modela ter ročno korigirano besedilo) smo model naučili posebnosti slovenske pisave - vključno z določenimi starimi črkovnimi zapisi, diakritičnimi znamenji in pogostimi besednimi zvezami v slovenskem jeziku. Rezultat prilagoditve je novi model PP-OCR-5-SL, ki bistveno bolje prepozna slovensko besedilo, še posebej na zgodovinskih dokumentih, kot osnovna različica.

Pomemben del rešitve je tudi komponenta za kontekstualno črkovanje, imenovana *Micka*. Gre za velik jezikovni model, razvit v sklopu PoVejMo (izhaja iz družine odprtokodnih modelov *GaMS*), ki je usposobljen za slovenski jezik. Njegova vloga v našem OCR sistemu je post-procesiranje: besedilo, ki ga prepozna model PP-OCR-5-SL, se posreduje modelu *Micka*, ta pa v kontekstu celotnega stavka popravi morebitne tipkarske napake ali neobičajne besede, pri čemer upošteva tudi zanesljivost prepoznave, da morebitne originalne napake (prisotne v izvirnem besedilu) ohrani.

Tehnološko gledano, je storitev zasnovana kot neodvisen REST API znotraj infrastrukture GalisOnline. Za izvajanje OCR se uporablja Python strežniška aplikacija, ki preko PaddleOCR prebere sliko in vrne prepoznano besedilo (v JSON obliki skupaj s koordinatami zaznanih blokov besedila). V naslednjem koraku API izvede še klic modulov za kontekstno popravljanje besedila in končni izpis optimizira za vnos v podatkovno bazo (npr. odstrani odvečne prelome vrstic, normalizira posebne znake). Celoten proces je dovolj hiter, da lahko deluje v interaktivnem času za posamezne dokumente - za običajno stran teksta sistem potrebuje manj kot sekundo za prepoznavo in nekaj sekund za jezikovni popravek na standardni strojni opremi, kar je sprejemljivo za neposredno uporabo v aplikaciji.

Implementacija kot del GalisOnline

Razvito OCR rešitev smo tesno povezali z obstoječim sistemom GalisOnline, tako da bo v prihodnosti uporabnikom na voljo brez potrebe po zunanjem orodju. Arhitektura integracije sledi načelu modularnosti: GalisOnline aplikacija po potrebi pošlje zahtevo na OCR API (ko uporabnik naloži novo sliko dokumenta ali ko sproži akcijo "preberi besedilo" za obstoječo sliko). API vrne prepoznano in korigirano besedilo, ki se nato shrani v sistem kot del podatkov o digitalnem objektu. S tem lahko to besedilo sodeluje v indeksiranju za iskanje, prav tako pa ga lahko katalogizator pregleda in po potrebi uredi.

Ena od prednosti razvite rešitve je, da je bila neposredno preizkušena in integrirana v obstoječo aplikacijo GalisOnline - tehnično smo GalisOnline nadgradili z novim modulom za prepoznavo besedila.

The screenshot displays the GalisOnline interface. On the left, a 'MEDIA' sidebar lists files: 'sword.png', 'ocr-test2 1.png', 'ocrocr.png', and 'unnamed.jpg'. The main area shows a large image of a yellow commemorative plaque titled 'UPORNIKI Z RAZLOGOM'. To the right of the image is a list of tags under the heading 'Oznake'. The tags are: 'UPORNIKI Z RAZLOGOM*', 'TE) STAVBI JE BIL OB RAZOROŽEVANJU SLOVENSKE', 'TERITORIALNE OBRAMBE', 'MAJA 1990 PREPOZNAN', 'RESNICNI NAMEN JUGOSLOVANSKEGA VOJAŠKEGA VRHA', 'AGRESIJA NA REPUBLIKO SLOVENIJO. IZ KOČEVJA JE ZATO', 'ODJEKNI PRVI JAVNI PROTEST ZOPER RAVNANJE', 'E ZVEZNE', 'ARMADE, KOT ODGOVOR NA', 'ODVZEM OROŽJA PA JE BILA', 'NA TEM MESTU', 'IDEJNO', 'IN', 'OPERATIVNO ZASNOVANA', 'MANEVRSKA STRUKTURA', 'NARODNE ZASCITE, TAJNA OBRAMBNA', 'ORGANIZACIJA', 'KI', 'JE', 'POMENILA', 'ZAMETEK VOJSKE', 'NASTAJAJOCE', 'SLOVENSKE', and 'DRZAVE'.

MEDIA

Datoteke

sword.png

Auto Refresh: OFF

ocr-test2 1.png

Auto Refresh: OFF

ocrocr.png

unnamed.jpg

Isči

Slika **Oznake**

Oznake

UPORNIKI Z RAZLOGOM*

TE) STAVBI JE BIL OB RAZOROŽEVANJU SLOVENSKE

TERITORIALNE OBRAMBE

MAJA 1990 PREPOZNAN

RESNICNI NAMEN JUGOSLOVANSKEGA VOJAŠKEGA VRHA

AGRESIJA NA REPUBLIKO SLOVENIJO. IZ KOČEVJA JE ZATO

ODJEKNI PRVI JAVNI PROTEST ZOPER RAVNANJE

E ZVEZNE

ARMADE, KOT ODGOVOR NA

ODVZEM OROŽJA PA JE BILA

NA TEM MESTU

IDEJNO

IN

OPERATIVNO ZASNOVANA

MANEVRSKA STRUKTURA

NARODNE ZASCITE, TAJNA OBRAMBNA

ORGANIZACIJA,

KI

JE

POMENILA

ZAMETEK VOJSKE

NASTAJAJOCE

SLOVENSKE

DRZAVE.