Aktivnosti, projekti i potencijal za primjenu umjetne inteligencije i strojnog učenja u baštinskim ustanovama

12. festival hrvatskih digitalizacijskih projekata
4. i 5. svibnja 2023.

DRAGANA KOLJENIK

Umjetna inteligencija, strojno učenje i baštinske ustanove

ŠTO MORAMO ZNATI

- Osnovni pojmovi
- Utjecaj na knjižnice
- Primjena u digitalnim zbirkama
- Zbirke kao podaci
- Projekti i aktivnosti
- Preporuke i zaključak

Što je UI (AI)?

Umjetna inteligencija je široka grana računalne znanosti koja se bavi izgradnjom pametnih strojeva sposobnih za obavljanje zadataka koji obično zahtijevaju ljudsku inteligenciju, a u nekim zadacima i nadilaze ljudske sposobnosti.

Slaba tj. uska UI

Djeluje unutar ograničenog konteksta i predstavlja simulaciju ljudske inteligencije.

Izvodi objektivne funkcije pomoću modela treniranih na podacima.

Često spada u kategorije dubokog učenja ili strojnog učenja.

Jaka UI

Stroj sa punim setom kongnitivnih sposobnosti kakve imaju ljudi.

Jaka umjetna inteligencija daleko je iznad mogućnosti trenutnih AI tehnologija.

Naziva se i Generalnom umjetnom inteligencijom (AGI).

Evolucija UI

Reaktivni strojevi

Donosi odluke na temelju podatka, ali nema pamćenje pa ne može učiti.

Ograničena memorija

Pamti podatke i može učiti iz prošlih iskustava, ima sposobnost predviđanja.

Teorija uma

Razumije koncepet ljudskog uma i donosi odluke kroz samorefleksiju.

Samosvijest

Postaje svjestan sebe i razumije svoje postojanje u svijetu.

Što je strojno učenje (ML)?

Vrsta umjetne inteligencije koja podrazumijeva „obuku“ ili „treniranje“ strojeva tako da uče iz podataka, a kako bi mogli predviđati ili donositi odluke bez eksplicitnog programiranja za svaki pojedinačni scenarij.

Nadgledano

Oslanja se na označene podatke koji se koriste za treniranje modela, kako bi model donio odluku o neoznačenim podacima.

Čovjek „uči“ model kako da radi – treniranje i uporeba označenih ulaznih i izlaznih podataka prije nego što sustav klasificira novi set podataka.

Nenadgledano

Ne zahtijeva ljudski input jer se modeli sami treniraju s neobrađenim i neoznačenim podacima.

Često se koristi u ranoj fazi istraživanja kako bi se bolje razumio skup podataka.

Duboko učenje i neuronske mreže

Napredna vrsta strojnog učenja koja koristi neuronske mreže inspirirane strukturom mozga. Uče treniranjem na setu podataka da bi s vremenom mogli izvršavati sve kompleksnije zadatke.

Obrada prirodnog jezika (NLP)

Područje umjetne inteligencije koje uključuje korištenje algoritama za analizu i tumačenje ljudskog jezika, poput teksta i govora, kako bi računalo "razumjelo" značenje.

AI model

Algoritam koji je istreniran na podacima za obavljanje određenog zadatka.

Jezični modeli

Algoritmi koji mogu odrediti koliko je vjerojatno da je niz riječi valjana rečenica - pokušavaju predvidjeti sljedeću najprikladniju riječ koja će popuniti prazan prostor u rečenici ili izrazu, na temelju konteksta zadanog teksta.

Veliki jezični modeli (LLM)

Jezični modeli koji koriste algoritme dubokog učenja za obradu i razumijevanje prirodnog jezika. Treniraju se na ogromnim količinama tekstualnih podataka kako bi naučili obrasce i odnose entiteta u jeziku.

Transformer modeli

Model dubokog učenja s tzv. mehanizmom samopažnje koji različito vrednuje značaj svakog dijela ulaznih podataka i prati odnose u podacima, poput reda riječi u rečenici.
Koristi se prvenstveno u područjima računalnog vida (CV) i obrade prirodnog jezika (NLP) gdje su upravo zahvaljujući ovom modelu u posljednih nekoliko godina napravljeni veliki iskoraci.

Generativna UI

Kad god tehnologija umjetne inteligencije generira nešto sama, to se može nazvati "generativnom umjetnom inteligencijom".

Generativni pred-trenirani transformer model (GPT)

Model strojnog učenja koji koristi nenadzirane i nadzirane tehnike strojnog učenja za razumijevanje i generiranje jezika nalik ljudskom.

GPT-3 je treniran na oko 45 terabajta tekstualnih podataka (6,5 milijuna stranica dokumenata; četvrtina cijele Kongresne knjižnice) s više od 175 milijardi parametara.

AI u knjižnicama

PHOTO BY: DALL-E

Genrativni transformer model baziran na neuronskom mrežama i tehnikama dubokog učenja.Treniran je na kombinaciji slika i opisnog teksta (razumije jezik kako bi mogao generirati slike).

Potencijal

Povećati učinkovitost poslovanja

Poboljšati korisničke usluge

Povećati mogućnost otkrivanja znanja

Kreiranje novih uloga za ustanove

Podrška zajednicama

Prepreke

Ljudski resursi

Financijski resursi

Etička i pravna pitanja

Drugi prioriteti ustanova

Slabo umrežavanje i suradnja

Pitanja oko etičnosti i pravna pitanja u kontekstu digitalnih zbirki

Opća pitanja o etičnosti AI

Pristranost, moralne odluke, privatnost, ravnoteža snaga, vlasništvo, utjecaj na okoliš, redundantnost poslova itd.

S napretkom tehnologije sve je više pitanja oko etičnosti.

Pravna pitanja

Pitanje privatnosti i zaštite osobnih podataka.
Pitanje autorskih prava.
Pitanje licenciranja.
Pitanje ponovne upotrebe i distribucije podataka.

Porijeklo zbirki i povijesna pristranost

Pretpostavke, predrasude, isključenja i pristranosti bilo kojeg skupa podataka odrazit će u svim procesima ML-a koji koriste taj skup podataka - bitna je kontekstualizacija i dokumentacija etički problematičnih podataka.

AI, ML i digitalne zbirke

Omogućavanje bolje pronalažljivosti u zbirkama i između zbirki najčešće je spominjana upotreba umjetne inteligencije i strojnog učenja u kontekstu digitalnih zbirki baštinskih ustanova.

Bolja pretraživost

Nadopunjuje ljudsku katalogizaciju i opise; sve veća količina digitalizirane građe zahtijevat će ML tehnike kako bi zbirke bile pretražive.

Bolji rezultati

Omogućuje točnije i bogatije rezultate pretraživanja metodama poput grupiranja, klasifikacije i dr.

"Serendipity"

Omogućuje "slučajne" pronalske u zbirkama zbog novostvorenih puteva kroz zbirke i među zbirkama.

Moguće uporabe tehnika strojnog učenja za pronalažljivost

- Grupiranje i klasifikacija
- (pred) Obrada
- Optičko prepoznavanje znakova (OCR)
- Prepoznavanje rukopisa
- Prepoznavanje i dodavanje metapodataka
- Ekstrakcija povijesnih tabličnih podataka
- Anotacija vizualnih podataka
- Anotacija audio podataka
- Povezivanje zbirki

Važnost podataka

Collections as Data

The Santa Barbara Statement on Collections as Data

2017.

Institute of Museum and Library Services, Always Already Computational: Collections as Data projekt

1. Potaknuti računalno korištenje
2. Etičnost u fokusu
3. Smanjivanje prepreka za upotrebu
4. Specifične potrebe korisnika
5. Dokumentacija
6. Otvorenost podataka
7. Interoperabilnost
8. Transparentnost
9. Podaci koji opisuju podatke
10. Proces u tijeku.

Jesu li naše zbirke spremne za upotrebu u projektima strojnog učenja?

Obećanja o mogućnostima primjene AI i ML u digitalnim knjižnicama i zbirkama ozbiljno su ograničena nedostatkom i nedovoljnim pristupom strojno upotrebljivih podataka kao i dovoljno velikim korpusima podataka za treniranje u baštinskim ustanovama.

SAMA DIGITALIZACIJA NIJE DOVOLJNA ZA ISTRAŽIVANJE I IMPLEMENTACIJI ML

Pristup strojno upotrebljivim podacima

Baštinske ustanove moraju svoj fokus prebaciti na stavljanje strojno upotrebljivih podataka na raspolaganje istraživačima.

Interoprabilnost metapodatka!

Ustanove mogu ponuditi jednostavne opcije za preuzimanje podataka (tzv. download dumps) različitih veličina i formata.

Jedan skup strojno upotrebljivih podataka može potaknuti brojne eksperimente, vizualizacije, tumačenja i argumente, kako unutar baštinske ustanove tako i od strane vanjskih istraživača.

Izrada setova podataka za treniranje modela

Baštinske ustanove su zbog specifičnosti svojih zbirki do sad u ML projektima većinom morale izrađivati specifične setove podataka za treniranje modela (npr. treniranje modela za prepoznavanje rukopisa ili elemenata starih skeniranih novina.

To se često radi u crowdsourcing projektima kako bi se smanjilo potrebno vrijeme za taj ogroman posao.

Bibliografski zapisi i metapodaci najčešće nisu dovoljni kao korpus setova podataka za treniranje modela.

Vrste podataka u ML projektu

Podaci za treniranje modela

Testni podaci

Neoznačeni podaci

Novi (ML) podaci

Proces **ML projekta**

Prikupljanje materijala

(pred) obrada materijala

Izrada seta podataka za treniranje

Odabir i podešavanje modela za treniranje

Treniranje modela na setu podataka za treniranje

Testiranje

Upotreba modela na ostatku materijala

Davanje na korištenje / implementacija

Procjena, dokumentacija i upravljanje procesom strojnog učenja

Dokumentacija svih koraka u ML projektu - od opisa zbirke i seta podataka, odabira i procjene modela, načina treniranja modela, načina stavljanja podataka i modela u otvoreni pristup itd.

Jesu li naše ustanove spremne za integriranje ML podataka u svoje sustave i infrastrukturu?

Rezultati projekta ML još uvijek su rijetko integrirani u sustave i infrastrukturu baštinskih ustanova, a integracija bi donijela svoj set izazova. Problem predstavlja i zabrinutost stručnjaka oko pouzdanosti ML generiranih podataka.

SVIJET RUČNO IZRAĐIVANIH OPISA MORA SE POVEZATI SA SVIJETOM AUTOMATSKI IZRAĐIVANIH OPISA.

"Bilo bi mudro da počnemo razmišljati o strojevima i algoritmima kao o novoj vrsti korisnika — korisnika koji ne zamjenjuje ljudske korisnike, ali ima neke drugačije potrebe i može zahtijevati drugačiji skup vještina i drugačiji način razmišljanja o tome kako bi se naši resursi mogli koristiti."

CHRIS BOURG, MIT

Projekti i aktivnosti

Newspaper Navigator

2020.

Library of Congress,
Benjamin Charles Germain Lee

15 mjeseci

Najveći set podataka ekstrahiranih vizualnih sadržaja iz povijesnih novina ikada proizveden.

Set podataka: vizualni sadržaji iz 16,3 milijuna stranica skeniranih novina "Chronicling America".

Set podataka za treniranje modela: crowdsourcing projekt Beyond Words

Vizulani sadržaji: naslovi, fotografije, ilustracije, karte, stripovi, oglasi i natpisi ispod slika.

Aplikacija za pretraživanje: 1,56 milijuna slika.

Living With Machines

2018. -

British Library
Alan Turing Institute
i dr.

Najveći projekt digitalne humanistike u UK.

Set podataka: digitalizirani popisi stanovništva, novine, knjige i karte iz razdoblja 19. stoljeća iz fonda BL.

Rezultati:

- publikacije (članci, knjige)
- setovi podataka (digitalizirane zbirke, derivirani setovi podataka iz zbirki)
- alati za vizualizaciju podataka (Micromap, PressTracer)
- softverski paketi, istraživački alati i kod
- crowdsourcing
- radionice i tutorijali

Saint George on a Bike

2019. -

Barcelona Supercomputing Center
Europeana Foundation

Izrada AI sustava pomoću neuronskih mreža, jezičnih modela i semantičkog zaključivanja koji će moći razumjeti i opisati detalje vizualnog sadržaja u ikonografskim slikama od 12. do 18. stoljeća.

Treniranje podataka: crowdsourcing.

video

Humans.Machine.Culture.
- Artificial Intelligence for the Digital Cultural Heritage

2022. - 2025.

Berlin State Library (SBB)

Nastavak Qurator projekta

- Izrada inteligentnih metoda za generičku analizu dokumenata
- Izrada alata za analizu vizualnog materijala
- AI potpomognuta analiza sadržaja i predmetno označivanje
- Pristup i upravljanje podacima

National Neibourghs

Carnegie Mellon University
University of Pittsburgh

Robots Reading Vogue

Digital Humanities Lab, Yale University Library

PixPlot

Digital Humanities Lab, Yale University Library

Annif

National Library of Finland

Crowdsourcing

TRANSKRIBUS

HUMANS IN THE LOOP

BEYOND WORDS

BY THE PEOPLE

AI i ML: baštinske ustanove

ZAKLJUČCI I PREPORUKE

Edukacija

Uključivanje javnosti

Integracija ML u proces digitalizacijie

Dokumentiranje

Davanje pristupa podacima

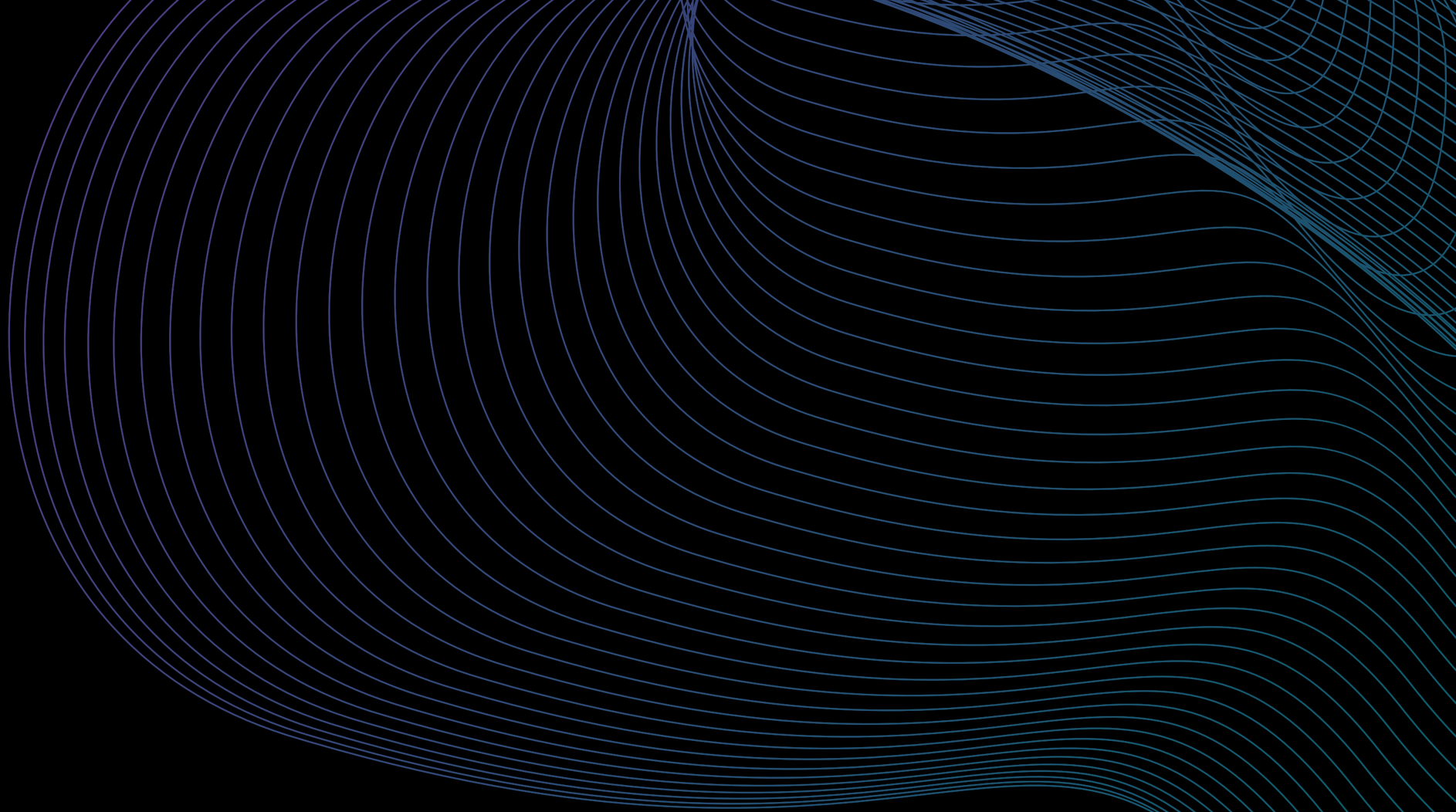
Briga o etičnosti i pravnim pitanjima

Intenzivna suradnja unutar i van sektora

Dijeljenje

"Naša je strategija postati svjesni umjetne inteligencije, a ne ju nužno implementirati posvuda."

MIA RIDGE, BRITISH LIBRARY



Hvala na pažnji!