



LEGAT

Sistem informatic avansat, bazat pe inteligență artificială (IA) pentru identificarea și extragerea entităților din colecții de date nestructurate

CONSORȚIUL



UNIVERSITATEA BUCUREȘTI
Contractor/ Coordonator



ACADEMIA TEHNICĂ MILITARĂ „FERDINAND I”
Partener 1

ne**xt**gen

NEXTGEN SOFTWARE
Partener 2

Proiect

- Coordonator: Paul Irofti
- Proiect: PN-IV-P6-6.3-SOL-2024-0090
- Consorțiu:
 1. UB (coordonator)
 2. ATM
 3. Nextgen Software SRL.
- Echipa: 30 positions
- Finanțator: UEFISCDI
- Buget: 2.122.787 lei (~ 424.557 euro)
- Perioadă: 6 iunie 2024 - 5 iunie 2026

Echipa: Universitatea din București

- Paul Irofti -- Principal Investigator
- Radu Ionescu -- Senior Researcher
- Marius Popescu -- Senior Researcher
- Iulia Timofte -- Researcher
- Roxana Voicu -- Researcher
- Adriana Costache -- Assistant Researcher
- Eduard Poesina -- Assistant Researcher
- Ana Cristina Rogoz -- Assistant Researcher
- Silviu Gheorghe -- Master Student

Echipa: Academia Tehnică Militară

- Luciana Morogan -- Principal Investigator
- Ion Bica -- Senior Researcher
- Ștefan Adrian Toma -- Senior Researcher
- Mihai Coca -- Researcher
- Iulian Tiță -- Assistant Researcher
- Mirabela Medvei -- Assistant Researcher
- George Hariga -- Assistant Researcher
- Alexandra Buzățoiu -- Master Student
- Paul-Florinel Căsândroi -- Master Student
- Ilie-Cosmin Bilțan -- Master Student
- Florina Conchințoiu -- L1 Technician
- Andrei Brînzea -- L1 Technician

Echipa: Nextgen Software SRL

- Bogdan Legănaru -- Principal Investigator (June 2024 - March 2025)
- Vlad Gladin -- Senior Researcher
- Daniel Tache -- Researcher
- Alexandru Cocosila -- Researcher
- Viorel Tigănescu -- L2 Technology Engineer
- Bonciu Emilian Cristian -- L2 Technology Engineer
- Adrian Bogdan Sandu -- L2 Technician
- Cristian-Ștefan Ene -- L2 Technology Engineer
- Alin Ungureanu -- Researcher

Arhitectura LEGAT



LEGAT

Proiect realizat de consorțiul format din



Universitatea
București



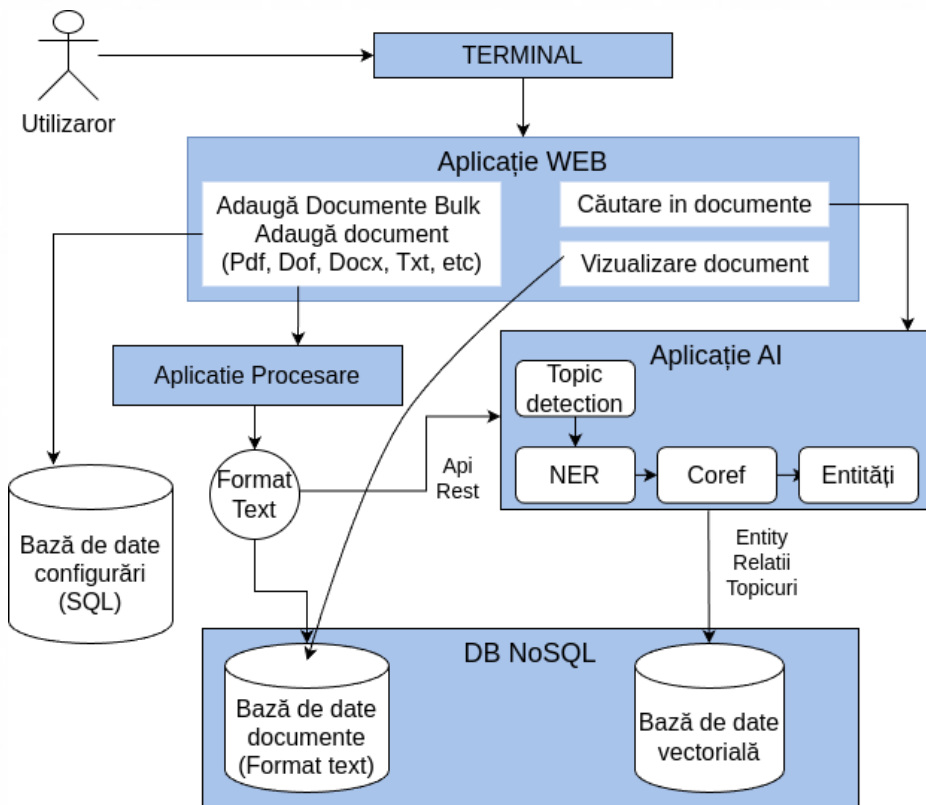
Academia Tehnică Militară
„Ferdinand I”

nextgen

Nextgen
Software

Arhitectură

- modul procesare
- modulul de caracterizare a entităților și realizare a legăturilor
- modulul baze de date
- modulul interfață cu utilizatorul



Echipament Hardware

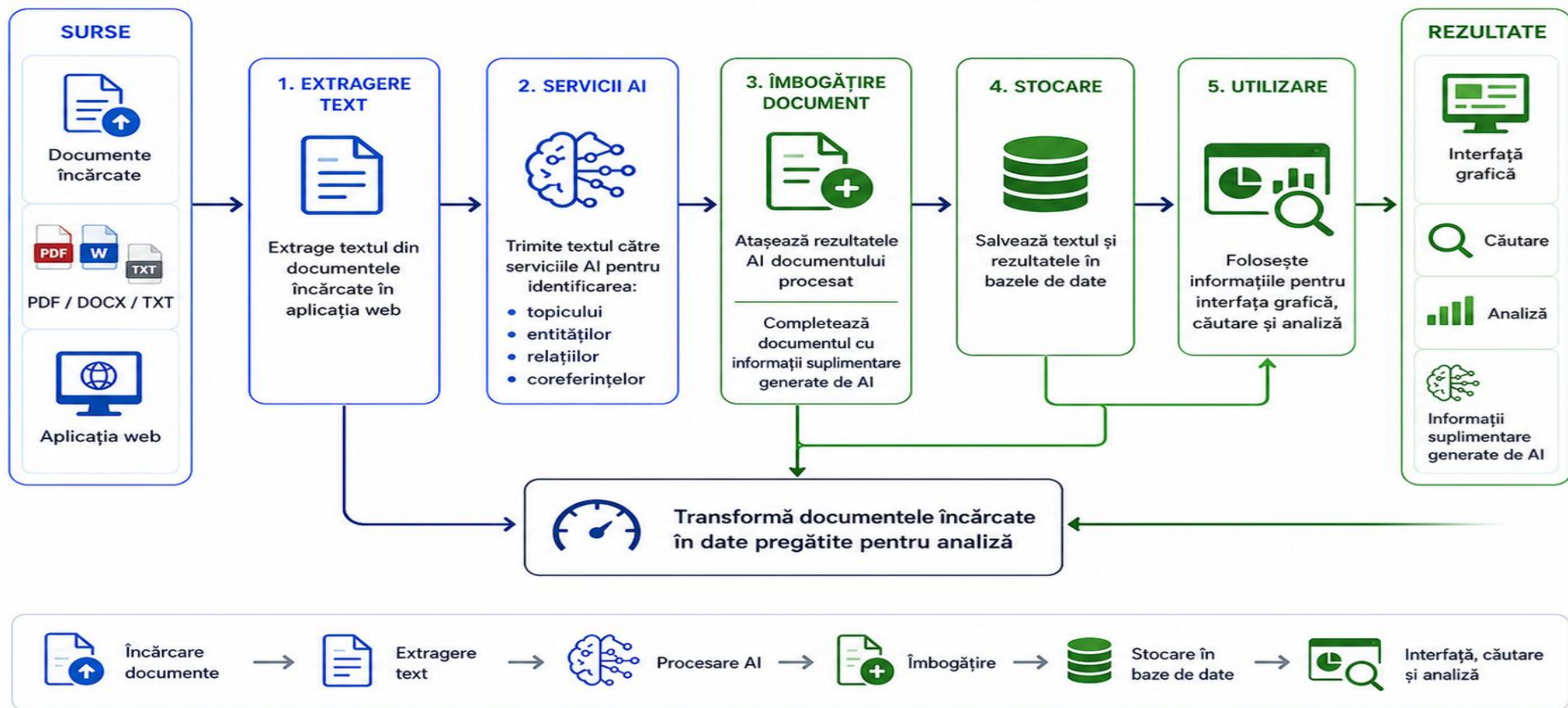
Server Sistem Informatic

- Aivres / KA YTUS KR2x80
- **CPU** AMD EPYC Gen4 9334 32C 64T
- **GPU** 4 x NVIDIA L40s 48GB
- **HDD** 6x 7.68TB NVMe Gen4
- **RAM** Memorie 1536 GB
- 3 porturi 1 Gbps

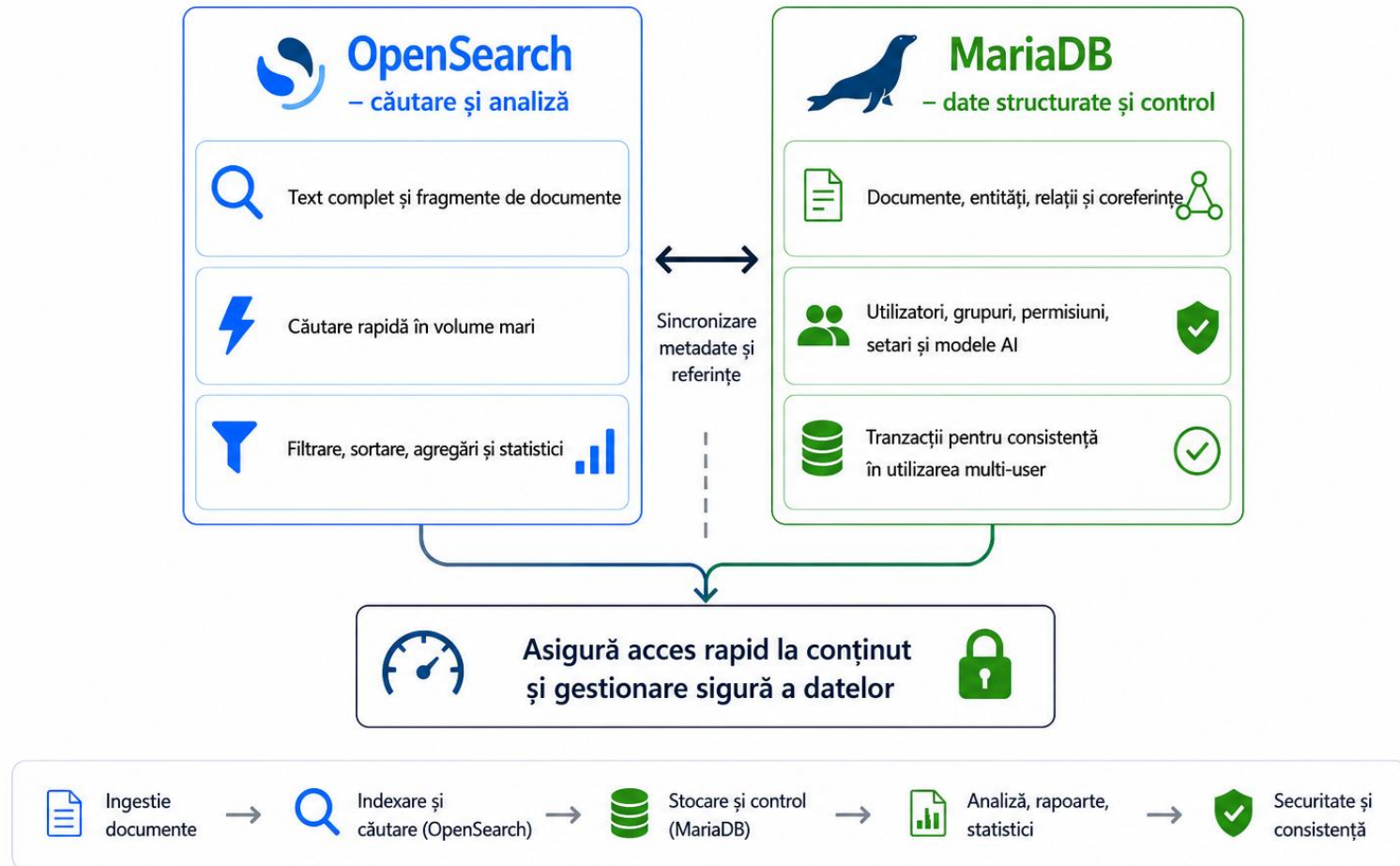
Terminal

- Lenovo Thinkbook 16 07IML
- **CPU** Intel Core Ultra 7155H
- **Conectivitate** 1 porturi 1Gbps
- **HDD** 8TB SSD (2x 4TB)
- **RAM** 64 GB

Arhitectura: Procesarea documentelor



Arhitectura căutare gestionare date



Componente AI



LEGAT

Proiect realizat de consorțiu format din



Universitatea
București



Academia Tehnică Militară
„Ferdinand I”

nextgen

Nextgen
Software

Modulul de caracterizare a entităților și realizare a legăturilor

Modulul este împărțit în 4 componente specializate:

- Named Entity Recognition (NER)
- CoReference Resolution (CoRef)
- Relationship Extraction
- Topic Detection

Named Entity Recognition (NER)

INFERENȚĂ

- Extrage entitățile relevante din text;
- Asociază un atribut fiecărei entitate EVENT, GPE, PERSON, FACILITY, ORGANIZATION, etc;
- Este setat și un scor de confidență;
- Pipeline hibrid;
- Regex pentru entități cu pattern fix;
- Model spaCy CNN-based Tok2Vec + Transition-Based NER:
 - Embeddings statice;
 - Contextualizare locală prin MaxoutWindowEncoder;
 - Fără mecanism Transformer sau self-attention.

ANTRENARE

- Entitățile etichetate din baza de date sunt transformate în exemple de antrenare;
- Per document, extragem span-urile de text și etichetele asociate entităților persoană, organizație, locație, identificatori
- Modelul utilizează un pipeline spaCy bazat pe Tok2Vec și Transition-Based NER, iar stratul final este adaptat la setul complet de tipuri de entități disponibile.
- Modelul reantrenat va identifica și clasifica automat entitățile etichetate manual în baza de date.

MULTILINGV

- engleză
- română
- maghiară
- rusă
- franceză;

Named Entity Recognition (NER)

România a participat la Campionatul Mondial de Fotbal din 1998.

Lionel Messi este participant al Campionatului Mondial de Fotbal din 2022.

Lionel Messi 100%
Entity Attributes
PERSON

România 96%
Entity Attributes
GPE

Campionatul Mondial de Fotbal 95%
Entity Attributes
EVENT

CoReference Resolution (CoRef)

2. OBIECTUL SI PRETUL CONTRACTULUI

2.1. - **Executantul** se obligă să execute activitățile pentru întreținerea, repararea, mentenanța,

dezvoltarea și modernizarea dispozitivului de semnalizare rutieră în municipiul Cluj-Napoca —
componenta pentru dezvoltarea și modernizarea dispozitivului de semnalizare rutieră, în conformitate
cu obligațiile asumate prin prezentul contract.

2.2. - **Achizitorul** se obligă să plătească executantului prețul convenit pentru îndeplinirea contractului.
Prețul plătit executantului de către achizitor este de 69.493,71 lei la care se adaugă TVA 13.203,80 lei,

conform prețurilor unitare/tip de activitate aprobate prin HCL nr. 189/2023. /
(VV

Coreference Clusters

Cluster 1 (5 mentions)

"Executantul" 96%

"Achizitorul" 76%

"Executantul" 96%

"Executantul" 98%

"Achizitorul" 90%

Cluster 2 (1 mention)

"Intersecție complexă BUC 22.020" 81%

Cluster 3 (3 mentions)

"Romania" 83%

"Romania" 83%

"Romania" 93%

Cluster 4 (25 mentions)

"Executantul" 97%

"Executantul" 99%

"Executantul" 98%

"Executantul" 99%

"Achizitorul" 98%

CoReference Resolution (CoRef)

ANTRENARE

- Antrenare pe texte adnotate manual pentru identificarea mențiunilor și gruparea lor în lanțuri de coreferință, folosind arhitectura multilingvă Maverick.
- ex: („Andrei”, „el”, „băiatul” → aceeași persoană).
- Modelul învață mai întâi să detecteze mențiuni, apoi să le lege în lanțuri de coreferință.

INFERENȚĂ

- Identifică toate mențiunile relevante din text.
- Compară mențiunile pentru a stabili dacă indică aceeași entitate.
- Grupează mențiunile în lanțuri de coreferință.
- Atribue fiecărui lanț un scor de încredere.
- Procesează textele lungi pe fragmente și unește rezultatele finale.

MULTILINGV

Modelul folosește un encoder multilingv (mmBERT), care poate înțelege simultan texte în mai multe limbi.

Alte limbi: model multilingv

- română
- engleză
- franceză
- rusă
- maghiară

Relationship Extraction

INFERENȚĂ

- Entitățile extrase din text sunt combinate în perechi;
- Fiecare pereche este analizată împreună cu propoziția/contextul în care apare;
- Clasificatorul prezice
 - relația dintre entități
 - *no-relation*
 - scor de încredere
- Model bazat pe **LUKE**
 - transformer cu mecanism **entity-aware self-attention**

ANTRENARE

- Pregătire date:
 - Relațiile etichetate din baza de date → transformate în exemple de antrenare;
 - Pentru fiecare relație extragem entitățile + spanul de text relevant;
- Adaptare model:
 - ultimul strat al modelului adaptat la setul complet de relații disponibile;
- Rezultat: model reantrenat
 - prezice relațiile etichetate manual în baza de date

SUPORT MULTILINGV

- Română
 - model dedicat, antrenat pe date românești (Wikipedia);
- Engleză
 - model dedicat, preantrenat;
- Alte limbi: model multilingv
 - franceză
 - rusă
 - maghiară

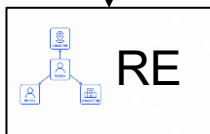


Relationship Extraction

Alexandru Ioan Cuza a fost ofițer în Armata Moldovei.



Alexandru Ioan Cuza a fost ofițer în Armata Moldovei.



Alexandru Ioan Cuza *military_branch* Armata Moldovei 51%

Compania Dacia este deținută de Grupul Renault.

Compania Dacia *owned_by* Grupul Renault 56%

Elena este soția lui Andrei din 2015.

Elena *relative:spouse* Andrei 100%

Mihai Eminescu a locuit în București o perioadă importantă din viața sa.

Mihai Eminescu *location:residence* București 46%



Relationship Extraction

RO

132

etichete Wikidata

EN

42

etichete TACRED

FR / RU / HU

19

etichete KBP37

Persoane & roluri

- occupation
- position_held
- employer
- military_rank
- relative

Locuri & administrație

- capital, country
- headquarters_location

Creații & cultură

- creator
- performer
- publisher

Știință, natură, sport...

Organizații

- founded_by
- members
- subsidiaries
- headquarters
- website

Persoane

- birth/death
- residence
- family
- religion, title, employee_of

Locuri

- city, country, state-or-province

Notă: toate seturile includ *no_relation*

Organizații

- founded / founded_by
- members
- subsidiaries
- top_members / employees
- alternate_names

Persoane

- employee_of
- title
- spouse
- origin
- alternate_names

Locuri

- city, country, state-or-province

Relationship Extraction - Limitări

DATE

- **Lipsa datelor în limba română în preantrenare**
 - nu este cunoscută proporția de conținut românesc utilizată pentru antrenarea modelului de bază
- **Posibile erori în relațiile generate automat**
 - relațiile pentru limba română au fost obținute prin **distant supervision** pe baza datelor Wikidata
- **Lipsa seturilor de date native** pentru limbi cu resurse reduse

MODEL

- **Limitările specifice modelelor neuronale**
- **Sensibilitatea la bias-uri prezente în date**, de exemplu asocierea eronată a unor locații cu reședința unei persoane, chiar dacă această informație nu este explicit prezentă în context;
- **Dificultatea interpretării deciziilor modelului**
- **Generalizarea redusă pe texte diferite de cele utilizate la antrenare**, de exemplu dificultăți în procesarea contractelor atunci când modelul a fost antrenat predominant pe texte enciclopedice.

Topic Detection

INFERENȚĂ

- calculează o reprezentare a textului
- pondere mai mare cuvintelor din lista de cuvinte relevante pentru toate topicurile
- calculează o reprezentare a textului
- determină cea mai similară reprezentare a unei clase

Topic **Romanian**

propagandă dezinformare 13%

ANTRENARE

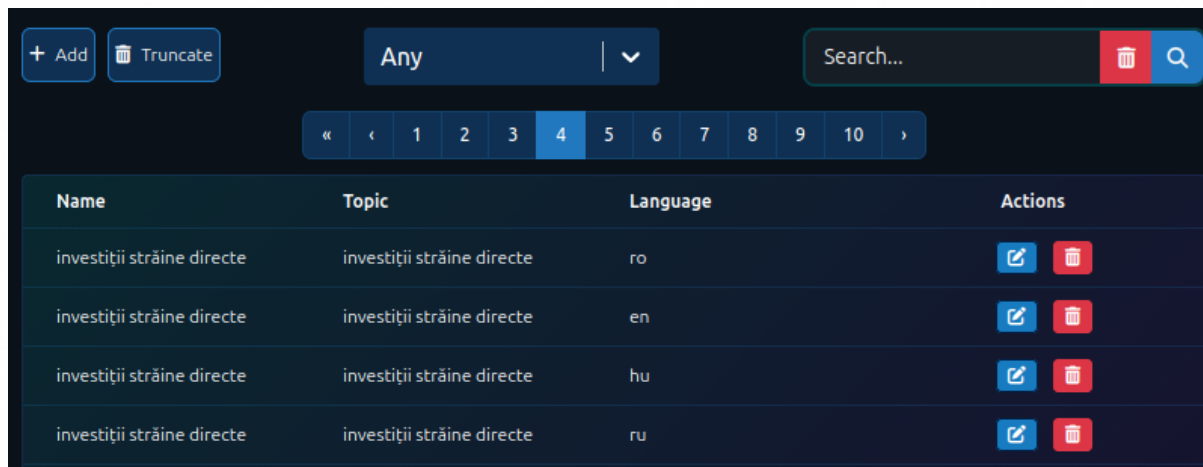
- per topic, se pleacă de la cuvintele relevante din lista de topicuri.
- bazat pe folosirea în documente, se caută alți termeni similari
- se caută termeni similari până la epuizare sau până se atinge un prag

- Se calculează câte o reprezentare pentru fiecare topic, bazat pe toți acești termeni (în context).

SUPORT MULTILINGV

- română - antrenat pe documentele în limba română.
- celelalte: antrenat pe toate documentele, inclusiv română
- bert-base-romanian-cased-v1, bert-base-multilingual-uncased

Topic detection: Lista de topicuri



Name	Topic	Language	Actions
investiții străine directe	investiții străine directe	ro	 
investiții străine directe	investiții străine directe	en	 
investiții străine directe	investiții străine directe	hu	 
investiții străine directe	investiții străine directe	ru	 

- lista conține numele topicurilor și cuvintele cheie în care documentele vor fi încadrate.
- topicurile nou create vor putea fi identificate la următoarea antrenare
- acuratețea crește cu numărul de documente

Topic Detection: Inferența

INFERENȚĂ:

- calculează o reprezentare a textului de clasificat
- calculează cea mai similară reprezentare a unei clase

MULTILINGV:

- Un model pentru română - antrenat pe documentele în limba română
- Un model pentru celelalte limbi - antrenat pe toate documentele, inclusiv română, în ideea în care documentele în română sunt mai numeroase și mai relevante

Interfața cu utilizatorul



LEGAT

Proiect realizat de consorțiul format din



Universitatea
București



Academia Tehnică Militară
„Ferdinand I”

nextgen

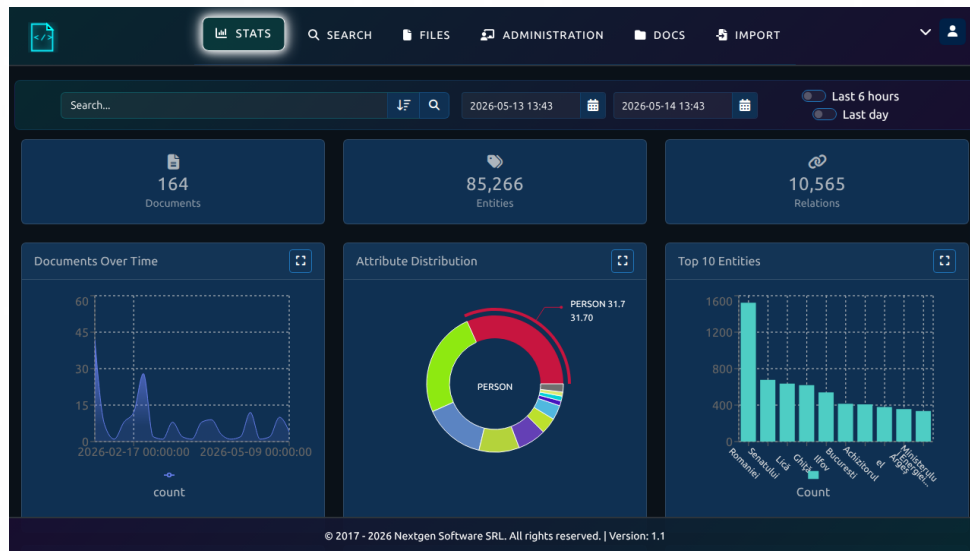
Nextgen
Software

Interfață de vizualizare grafica și complexă a datelor

Oferă o imagine de ansamblu asupra informațiilor extrase din documentele filtrate, prin afișări grafice ușor de interpretat

- Afișează metrici despre procesarea documentelor: număr de documente, entități și relații identificate
- Prezintă rezultatele sub formă de grafice intuitive: tendințe, distribuții și zone de interes

Interfața ajută utilizatorii să înțeleagă rapid datele extrase și să decidă unde este necesară o analiză mai detaliată





Interfață de căutare cu criterii de interogare multiple

Permite filtrarea rapidă a documentelor și identificarea informațiilor relevante în volume mari de date

- Filtrarea se poate realiza după textul din document sau după elementele identificate: entități, relații și topicuri
- Criteriile de căutare pot fi combinate prin reguli de tip AND/OR
- Rezultatele sunt afișate într-un tabel extensibil, ușor de explorat
- Fiecare rezultat poate fi deschis pentru vizualizarea fragmentului de text identificat și a poziției acestuia în document

The screenshot shows a web application interface for document search. At the top, there's a navigation bar with 'STATS', 'SEARCH', 'FILES', 'ADMINISTRATION', 'DOCS', and 'IMPORT'. Below this is a search bar with a dropdown menu showing 'Local Time' and 'Last day'. The main content area displays a table of search results with columns: View, Local Time (LocalTime), Name (Name), Type (Type), Processed At (ProcessedAt), and Language (Language). The table contains 10 rows of results, each with a document name, type, and processing time. At the bottom, there's a footer with '© 2017 - 2026 Nextgen Software SRL. All rights reserved. | Version: 1.1'.

View	Local Time (LocalTime)	Name (Name)	Type (Type)	Processed At (ProcessedAt)	Language (Language)
	2026-05-09T11:28:01.699	PopaTanda.pdf	pdf	2026-05-09T11:27:41.452	ro
	2026-05-09T11:25:39.552	PopaTanda.pdf	pdf	2026-05-09T11:25:21.150	ro
	2026-05-09T10:22:31.589	testrelationextraction.pdf	pdf	2026-05-09T10:22:15.041	ro
	2026-05-09T10:22:15.487	testrelationextraction.pdf	pdf	2026-05-09T10:21:44.317	ro
	2026-04-18T11:25:57.572	Contract487402.pdf	pdf	2026-04-18T11:19:22.878	ro
	2026-04-18T11:24:59.744	Contractachiziitiecombustibil.pdf	pdf	2026-04-18T11:19:22.809	ro
	2026-04-18T11:23:37.970	contractrefacereinstalatieelectrice.pdf	pdf	2026-04-18T11:19:22.676	ro
	2026-04-18T11:22:39.745	Contractachiziitiecombustibil1.docx	docx	2026-04-18T11:19:22.608	ro
	2026-04-18T11:22:17.644	Contract4060681.pdf	pdf	2026-04-18T11:19:22.539	ro
	2026-04-18T11:21:47.819	Contractnr.58351.pdf	pdf	2026-04-18T11:19:22.472	ro

Exemplu de căutare cu operator AND



The screenshot displays the Nextgen Software interface with a search results table. The search query is "entity:Popa Tanda and attribute:PERSON". The table has columns for View, Local Time, Name, Type, Processed At, and Language. There are three results listed, all for the file "PopaTanda.pdf".

View	Local Time (LocalTime)	Name (Name)	Type (Type)	Processed At (ProcessedAt)	Language (Language)
  	2026-05-09T11:28:01.699	PopaTanda.pdf	pdf	2026-05-09T11:27:41.452	ro
  	2026-05-09T11:25:39.552	PopaTanda.pdf	pdf	2026-05-09T11:25:21.150	ro
  	2026-03-09T13:20:53.400	PopaTanda.pdf	pdf	2026-03-09T13:20:40.223	ro

Total Results: 3

© 2017 - 2026 Nextgen Software SRL. All rights reserved. | Version: 1.1

Căutare “full text” și expandarea rezultate



The screenshot displays the Nextgen Software interface with a search results page. The top navigation bar includes tabs for STATS, SEARCH, FILES, ADMINISTRATION, DOCS, and IMPORT. The search bar at the top contains the text "Popa tando" and filters for "Local Time", "Name", "Type", "Processed At", and "Language". The results table shows four matches for the search phrase "Popa Tando". The first match is expanded, showing a list of text snippets from a PDF file named "PopaTanda.pdf". The snippets include dates like "1874" and "1875", and text about Wikisource and a poem. The bottom of the interface shows the total results count and an "Export CSV" button.

View	Local Time (LocalTime)	Name (Name)	Type (Type)	Processed At (ProcessedAt)	Language (Language)
	2026-05-09T11:28:01.699	PopaTanda.pdf	pdf	2026-05-09T11:27:41.452	ro
Q 4 matches for "Popa tando" (phrase)					
1 Popa Tando.					
2 1874					
3 Exportat de la Wikisource pe 14 noiembrie 2025					
273 într-atât, încât oamenii cale de-o poștă se feresc din drumul					
274 popii. A ajuns ca și ciuma. Dar mai rău decât toate este una:					
275 după atâta tîndălitură, oamenii i-au pus numele "Popa					
276 Tando". Apoi Popa Tando a și rămas.					
277 Vorbind drept, sătenilor numai într-un chip nu le plăcea					
278 felul popii. Fieștecare râdea bucuros cu popa de alții; nici					
274 popii. A ajuns ca și ciuma. Dar mai rău decât toate este una:					
275 după atâta tîndălitură, oamenii i-au pus numele "Popa					
276 Tando". Apoi Popa Tando a și rămas.					
277 Vorbind drept, sătenilor numai într-un chip nu le plăcea					
278 felul popii. Fieștecare râdea bucuros cu popa de alții; nici					
310 făcuse milă de bietul om și i-a dat dreptate, scăldînd ne					
	2026-05-09T11:25:39.552	PopaTanda.pdf	pdf	2026-05-09T11:25:21.150	ro
	2026-03-09T20:26:53.204	PopaTanda.pdf	pdf	2026-03-09T20:26:48.606	ro
	2026-03-09T20:26:53.204	PopaTanda.pdf	pdf	2026-03-09T20:26:48.606	ro
Total Results: 4					
Export CSV					

© 2017 - 2026 Nextgen Software SRL. All rights reserved. | Version: 1.1

Spațiu de lucru document

Oferă un spațiu dedicat pentru analiza detaliată a unui document

- Afișează textul documentului împreună cu topicul, attributele, entitățile, relațiile și coreferințele identificate (AI/user)
- Permite revizuirea umană a rezultatelor AI prin validare, corectare, completare sau eliminare de elemente
- Transformă documentul într-un spațiu de analiză verificabil și controlat
- Oferă legături dinamice către alte documente entități, coreferințe asociate

The screenshot displays a workspace for analyzing a document titled "PopaTanda.pdf". The main area shows the document text with various annotations. On the right, a sidebar provides a detailed analysis of the document's content.

Document Text and Annotations:

Popa Tanda Exportat de la Wikisource pe 14 noiembrie 2025

Capitolul I

Ierte-l Dumnezeu pe dascălul Pintilie Era cântăreț vestit. Și murăturile foarte mult îi plăceau. Mai ales dacă era cam răgușit, le bea cu gâlbenuș de ou și i se dregea organul, încât răsunau ferestrele când cânta Mântuiește Doamne norodul tău! Era dascăl în Buteucani bun sat și mare, oameni cu stare și socoteală, pomeni și ospețe de bogat. Iară copii n-avea dascălul Pintilie decât doi: o fată, pe care a măritat-o după Petrea Tapului, și pe Trandafir, părintele Trandafir, popa din Sărăceni. Pe părintele Trandafir să-l țină Dumnezeu! Este om bun; a învățat multă carte și cântă mai frumos decât chiar și răposatul său tată, Dumnezeu să-l ierte! și totdeauna vorbește drept și cumpănit, ca și când ar citi din carte. Și harnic și grăitor om este părintele Trandafir. Adună din multe și face din nimica ceva. Strănge, drege și culege, ca să aibă pentru sine și pentru alții.

Mult s-a ostenit părintele Trandafir în tinerețea lui. Școlile cele mari nu se fac numai iac-așa, mergând și venind. Omul sărac și mai are, și mai rabdă. Iară cu capul se lucrează mai greu decât cu sapa și cu furca. Dar toate s-au făcut și nici nu rămas lucru zadarnic. Trandafirică a ajuns popă în satul tătâne-său, în Buteucani, bun sat și mare, oameni cu stare și cu socoteală, dar la pomeni și la ospețe părintele Trandafir nu mergea bucuros.

Minunat om ar fi părintele Trandafir dacă nu l-ar strica un lucru. Este cam greu la vorbă, cam aspru la judecată: prea de-a dreptul, prea verde-fâșie. El nu mai sucește vorba, ci spune drept în față, dacă i s-a pus ceva pe inimă. Nu e bine să fie omul așa. Oamenii se prea supără când le luăm

Sidebar Analysis:

Search entities...

Topic: Romanian

propagandă dezinformare 13%

Attributes: New attribute +

DEFAULTS:

- ☒ 7 EVENT
- ☒ 66 FACILITY
- ☒ 34 GPE
- ☒ 14 LOCATION
- ☒ 22 ORGANIZATION
- ☒ 121 PERSON

OTHER:

- ☒ CAR_PLATE_RO
- ☒ CNP

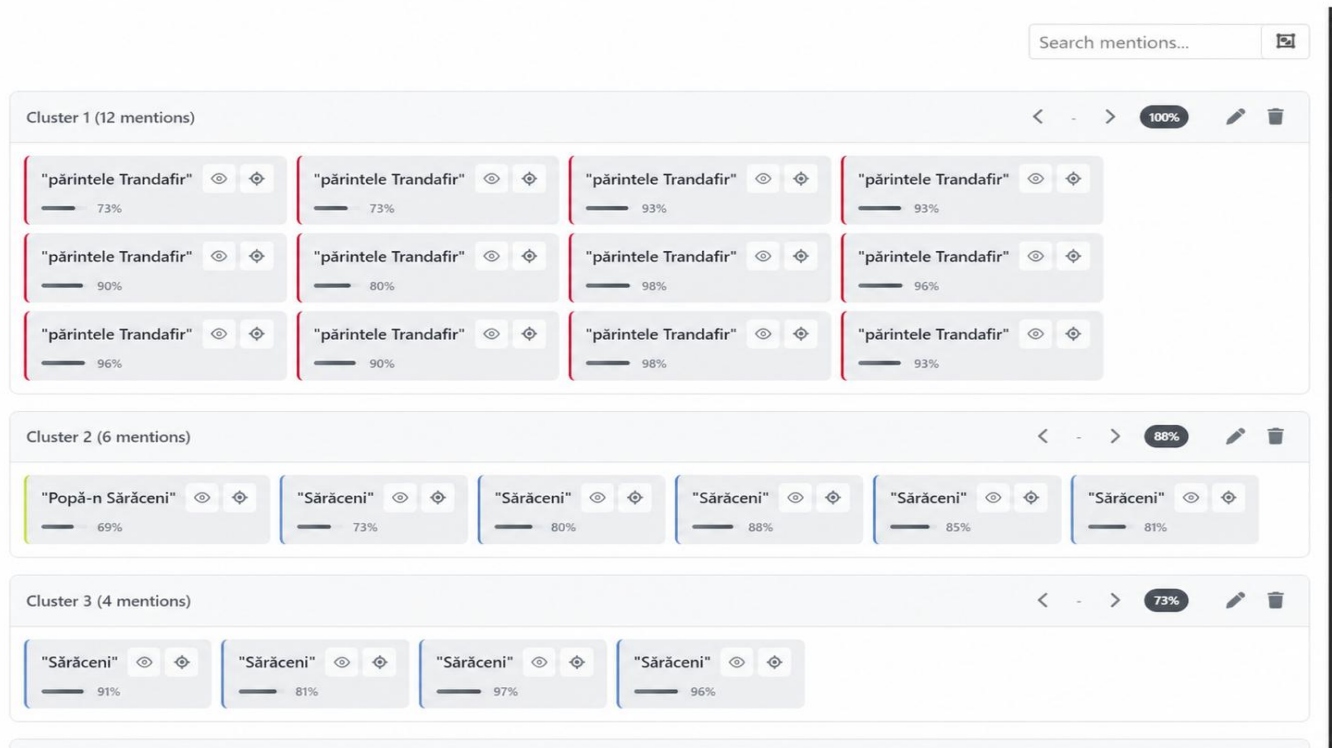
Entities: Search entities...

- Popa Tanda 100%
- Wikisource 92%
- Pintilie 71%
- Doamne 92%
- Buteucani 85%

Relations: Search relations...

- Popa Tanda ↔ relativșchi ↔ Pintilie 32%
- Wikisource ↔ characters ↔ Pintilie 61%

Coreference Clusters

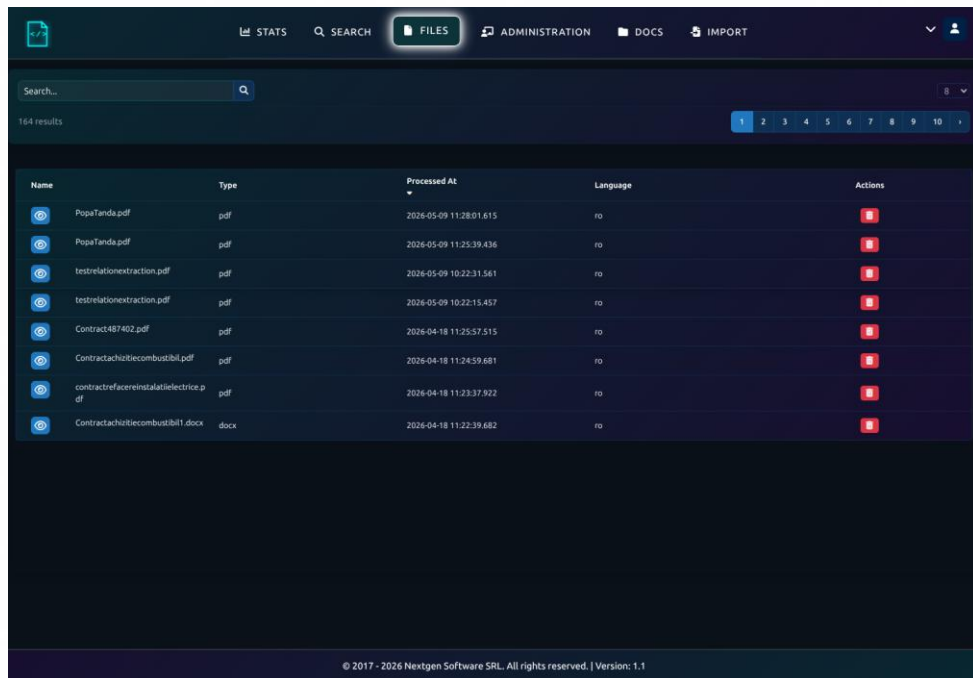


Istoricul documentelor

Oferă o vedere de ansamblu asupra documentelor încărcate și procesate în platformă

- Permite filtrarea rapidă după nume, tip de fișier sau limbă
- Ajută utilizatorii să verifice rapid documentele procesate
- Oferă acces direct către vizualizarea documentului

Funcționează ca un istoric simplificat al fișierelor încărcate, facilitând urmărirea și accesarea documentelor disponibile în platformă.



Name	Type	Processed At	Language	Actions
PopaTanda.pdf	pdf	2026-05-09 11:28:01.615	ro	[Icon]
PopaTanda.pdf	pdf	2026-05-09 11:25:39.436	ro	[Icon]
testrelationextraction.pdf	pdf	2026-05-09 10:22:31.561	ro	[Icon]
testrelationextraction.pdf	pdf	2026-05-09 10:22:15.457	ro	[Icon]
Contract487402.pdf	pdf	2026-04-18 11:25:57.515	ro	[Icon]
Contractachiziitecombustibil.pdf	pdf	2026-04-18 11:24:59.681	ro	[Icon]
contractrefacereinstalatielectrice.pdf	pdf	2026-04-18 11:23:37.922	ro	[Icon]
Contractachiziitecombustibil1.docx	docx	2026-04-18 11:22:39.682	ro	[Icon]

Administrare modele AI

Oferă administratorilor control asupra modelelor AI utilizate în procesul de analiză a documentelor

- Permite gestionarea modelelor AI folosite pentru topicuri, entități, relații și coreferințe
- Oferă vizibilitate asupra stării și versiunilor disponibile pentru fiecare model
- Permite pornirea sau oprirea procesului de antrenare pentru limba selectată
- Oferă posibilitatea de a schimba versiunea activă a unui model
- Permite păstrarea și gestionarea versiunilor anterioare ale modelelor AI

Asigură control operațional asupra modelelor active și asupra procesului de îmbunătățire

© 2017 - 2026 Nextgen Software SRL. All rights reserved. | Version: 1.1

Concluzii



LEGAT

Proiect realizat de consorțiul format din



Universitatea
București



Academia Tehnică Militară
„Ferdinand I”

nextgen

Nextgen
Software

Limitări și dificultăți

- **decalajul de performanță** al modelelor în limba română comparativ cu cele în limba engleză
- **existența și calitatea** seturilor de date pentru limba română
- **dificultatea** evaluării performanței pe date reale
- vizualizarea completă a textului unui document, menținând integritatea elementelor contextuale detectate în procesul de analiză semantică
- limitările privind extragerea precisă a textului din documente digitalizate la rezoluții reduse (OCR)
- specializarea pe documente cu formatare informală (ex. contracte, procese verbale)

Direcții viitoare

- îmbunătățirea performanțelor **modelelor de bază** pentru limba română
- crearea unor **seturi de date** publice mai cuprinzătoare
- **modele care urmează instrucțiuni** pentru topic detection
- adaugarea mai multor limbi de circulație
- îmbunătățirea cu informații tip OSINT
- optimizare modul vizualizare
- conectivitate pentru preluare din baze de date existente de documente

DEMO



LEGAT

Proiect realizat de consorțiu format din



Universitatea
București



Academia Tehnică Militară
„Ferdinand I”

nextgen

Nextgen
Software

Întrebări?



LEGAT

Proiect realizat de consorțiul format din



Universitatea
București



Academia Tehnică Militară
„Ferdinand I”

nextgen

Nextgen
Software