ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Prezentare Solutie cu Inteligenta Artificiala pentru
Transcriere Voce si **Analiza Automata de Text**

IntelliDockers

www.zettacloud.ai

[Iulie, 2023]

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Cuprins

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Despre Zetta Cloud

Zetta Cloud România a fost înființată în 2013 la Cluj-Napoca, un habitat ideal pentru încurajarea talentului și inovării tehnologice. Scopul companiei noastre este de a oferi soluții software pentru o înțelegere profundă a conținutului folosind Inteligența Artificială.

Fondatorii Zetta Cloud au un background solid în industria software-ului, a Inteligenței Artificiale și a soluțiilor pentru companii. Echipa noastra e formata din **specialisti Big Data** (ingineri seniori calificați în domeniul tehnologiilor Big Data), **cercetători seniori AI** (Hybrid AI, Machine Learning, Teoria jocurilor, Teoria grafurilor, Algoritmi evolutivi) și **cercetători în ştiințe sociale și politice** (OSINT, Social Media, statisticieni, cercetători în științe politice și sociale).

Portofoliul nostru de soluții variază de la **Digital News Technologies** (soluții software bazate pe AI pentru sectorul Digital Media) şi **Cercetare în Inteligența Artificială** (dezvoltarea algoritmilor AI utilizând abordări de învățare deep learning și învățare automată), până la **servicii de analiză inteligentă de date** (servicii de analiză de date pentru date deschise și social media).

Zevo Technology, partner Zetta Cloud, este o companie specializată în machine learning și inteligenţă artificială aplicate în domenii precum procesarea automată a vorbirii şi semnalelor multimedia (audio, imagini, video). Zevo comercializează produse software proprii de dictare, transcriere automată a vorbirii în text, identificare automată a vorbitorului, sinteză de vorbire pornind de la text și altele. Toate aceste produse sunt disponibile în variantă “on premise” sau “cloud”. Suplimentar faţă de produsele software, Zevo oferă servicii de consultanță specializată și servicii de cercetare-dezvoltare personalizate în domenii precum procesarea automată a semnalului audio-video, vorbirii și imaginilor.

Echipa Zevo este formată din cercetători și ingineri cu experiență, majoritatea având un doctorat în domeniu, pasionați de evoluția tehnologică a domeniului în care lucrează, angrenați în permanenţă în proiecte de cercetare internaţionale și la curent cu ultimele descoperiri și invenții în materie de deep learning și inteligență artificială. Activitatea de peste 10 ani în domeniu, cu publicații științifice relevante în comunitatea academică reprezintă cea mai bună garanție a calității serviciilor de cercetare-dezvoltare ce pot fi prestate de echipa Zevo.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Solutiile Zetta Cloud

IntelliDockers (www.intellidockers.com) sunt motoare de analiză a textului și a vorbirii pentru procesarea sigură a datelor. Oferim tehnologie de procesare a limbajului natural de ultimă generație bazată pe Deep Learning și rețele neuronale ai ca motoare autonome. IntelliDockers sunt implementate pe infrastructura dvs. izolată, făcându-le potrivite pentru procesarea sigură a datelor.

Factory (www.Factory-nocode.ai) este o platformă de instruire ai fără cod pentru procesarea sigură a datelor, care oferă siguranță completă pentru datele clienților și care poate fi utilizată fără cunoștințe de codificare sau experiență anterioară. Cu Factory puteți instrui, evalua și implementa propriile modele ai pentru înțelegerea documentelor fără a scrie nicio linie de cod.

TrustServista (www.trustservista.com) este o platformă software unică care poate determina automat originea, calitatea conținutului și credibilitatea știrilor online. Acesta utilizează inteligența artificială pentru a înscrie automat articole de știri online sau conținut de știri dintr-o perspectivă de calitate și încredere.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

IntelliDockers Speech Transcription powered by Zevo Tech

Descriere generală

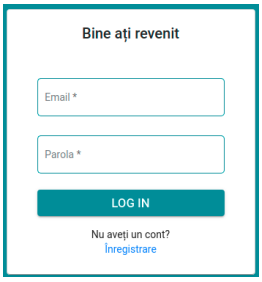
IntelliDockers Speech Transcription powered by Zevo Tech realizează transcrierea vorbirii dintr-un fișier sau flux audio în text. Soluția suportă diverse formate audio / video la intrare, lista de formate include formatele **.asf, .wmv, .mp4, .m4v, .ts, .mkv**.

Transcrierile sunt exportate în diverse formate, **inclusiv Microsoft Word**. Motorul AI poate transcrie vorbire în limba română și poate fi adaptat în funcție de necesități la alte domenii de vorbire sau caracteristici specifice anumitor vorbitori, inclusiv la alte limbi.

Platforma oferă o interfață Web prietenoasă prin care utilizatorul poate încărca, asculta și transcrie fișiere audio, respectiv poate urmări și corecta transcrierea rezultată. Suplimentar, aceste funcționalități pot fi oferite prin intermediul unui serviciu Web, disponibil unor potențiale aplicații third-party. Pentru o experiență cât mai bună, recomandăm utilizarea browserului Mozilla Firefox sau Google Chrome.

Fluxul principal

Pentru autentificarea în aplicație se vor folosi datele de acces (email și parolă) pentru fiecare utilizator înregistrat în parte.

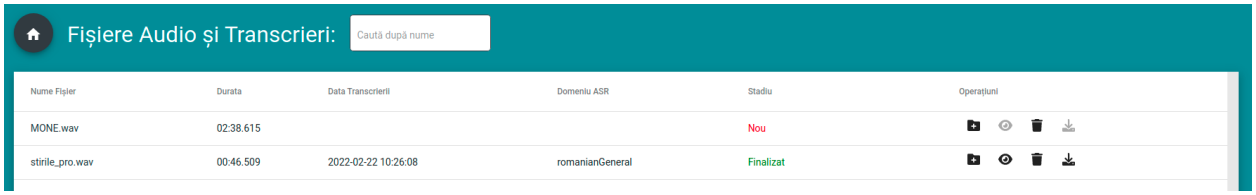


După autentificare, interfața Web prezintă pagina principală a aplicației, intitulată “Fișiere Audio și Transcrieri”, ce oferă următoarele funcționalități:

- Fișiere audio
  - Vizualizarea listei de fișiere audio încărcate în aplicație
  - Încărcarea unui nou fișier audio

- Ștergerea unui fișier audio
- Transcrieri
  - Inițierea unei noi transcrieri;
  - Vizualizarea listei de transcrieri realizate anterior;
  - Descărcarea unei transcrieri realizate anterior;
  - Vizualizarea unei transcrieri realizate anterior.

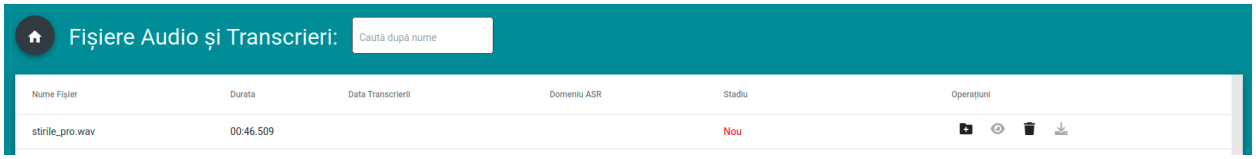
Vizualizarea listei de fișiere audio și a transcrierilor realizate anterior este ilustrată în figura de mai jos.



Încărcarea unui nou fișier audio se realizează din pagina Fișiere Audio și Transcrieri prin apăsarea butonului Încărcare () și selectarea unui fișier audio în format .mp3, .wav sau .flac.

Pentru o calitate optimă a transcrierii, fișierele audio ce conțin mai multe canale (fluxuri audio paralele) ar trebui separate în mai multe fișiere audio cu un singur canal, încărcate și transcrise separat.

După încărcarea unui nou fișier audio, acesta este afișat în lista de fișiere cu statusul Nou.



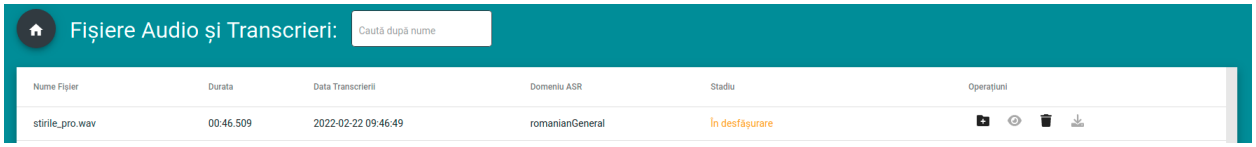
Ștergerea unui fișier audio se realizează din pagina Fișiere Audio și Transcrieri prin apăsarea butonului Șterge () asociat respectivului fișier audio. ATENȚIE: un fișier audio odată șters nu mai poate fi recuperat.

Inițierea unei noi transcrieri pentru un fișier audio se realizează din pagina Fișiere Audio și Transcrieri prin apăsarea butonului Inițiază Transcrierea () asociat respectivului fișier audio.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

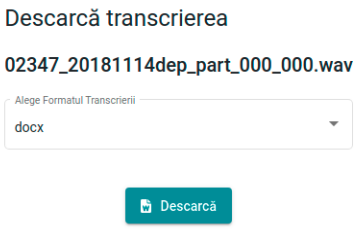
Aplicația va solicita detalii privind transcrierea ce se dorește a fi efectuată. Versiunea curentă a aplicației permite alegerea domeniului de transcriere. Versiunea curentă a aplicației are un singur domeniu de transcriere disponibil: **limba română, general**.

După apăsarea butonului Inițiază Transcrierea este afișată din nou pagina principală a aplicației, iar statusul transcrierii pentru respectivul fișier audio devine În desfășurare.



După ce statusul transcrierii devine Finalizat, aceasta poate fi descărcată prin apăsarea butonului Descarcă Transcrierea ( ) asociat respectivei linii din lista de fișiere audio și transcrieri.

Aplicația va solicita alegerea formatului în care este dorită transcrierea. Versiunea curentă a aplicației permite descărcarea transcrierii în format .docx și .json.



Prin apăsarea butonului Descarcă fișierul text asociat transcrierii poate fi descărcat pe calculatorul personal.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Vizualizarea unei transcrieri anterioare

După ce statusul transcrierii devine Finalizat, aceasta poate fi vizualizată prin apăsarea butonului Vezi Transcrierea ( ) asociat respectivei linii din lista de fișiere audio și transcrieri. Apăsarea acestui buton conduce la deschiderea paginii Detalii Transcriere. Această pagină permite:

- Vizualizarea transcrierii sub forma unei liste de segmente de text atribuite diverșilor vorbitori (așa cum ilustrează figura de mai jos)
- Ascultarea fișierului audio și urmărirea simultană a cuvintelor transcrise
- Navigarea prin fișierul audio și prin transcrierea aferentă
- Descărcarea transcrierii

Ascultarea fișierului audio și urmărirea simultană a cuvintelor transcrise se poate face din fereastra Detalii transcriere apăsând butonul Play sau apăsând pe bara de progres a fișierului audio. Aplicația va reda fișierul audio și va sublinia în transcriere cuvintele aferente respectivului moment de timp.

Navigarea prin fișierul audio și prin transcrierea aferentă se realizează apăsând pe cuvintele transcrise sau pe bara de progres a fișierului audio. Acest lucru conduce la redarea fișierului audio începând cu poziția selectată. Descărcarea transcrierii se realizează prin apăsarea butonului Descarcă .

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Integrarea Extractorului de Entități (NER)

Platforma IntelliDockers Speech Transcription powered by Zevo Tech poate integra și alte motoare cu Inteligență Artificială pentru procesarea de limbaj natural pentru analiza textului extras din vorbire în fluxul principal. Unul dintre motoarele de procesare de limbaj natural care ar adauga valoare în cazul dvs. ar fi Extractorului de Entități (NER).

Entitățile (cuvintele cheie) clasificate în funcție de tip (persoane, organizații, locații, produse, altele) constituie informația factuală principală a oricărui conținut. Extractorul de entități identifică automat aceste entități, este antrenat cu conținut generic și suportă un set de tipuri standard de entități, detaliate mai jos.

Spre exemplu, prin aplicarea Extractorului de Entități pe bucata de transcriere de mai jos, se pot vedea cuvintele cheie identificate și clasificate automat de către AI.

O fală din Arad LOC, aflată în spatele acestei porţi, unde Ioan Crişan PERS. Avea ferma de somn african, este posibil să fie locul unde dispozitivul exploziv, a fost amplasat în maşina omului de afaceri, înainte de atacul criminal 10 zile TIME autoturismul a fost parcat aici şi Crişana PERS folosit altul pentru deplasări locul are puţină pază, fiica lui Crişan PERS, avocat în Arad LOC, fosta soţie a deputatului Liberal TITLE, Sergiu Vâlcea PERS a confirmat pentru ştirile Pro Tv ORG, că a transmis informaţia anchetatorilor ipoteza potrivit căreia Bomba a fost plasată în autoturism, la un service din Arad este tot mai puţin luată în calcul a stat cel mult 20 şi 5 de minute TIME, cât să îi fie înlocuit la maşină, 2 anvelope.

Lista cu toate tipurile de entități pe care le identifica și clasifica automat motorul cu IA în acest moment se regaseste în tabelul de mai jos.

Entity Type (Long)	Short	Description
LOCATION	LOC	Un oraș, stat, țară, regiune, clădire, monument, apă, parc sau adresă
ORGANIZATION	ORG	O corporație, instituție, agenție sau altă structură organizațională
PERSON	PERS	O persoană identificată prin nume, pseudonim sau alias
PRODUCT	PROD	Produs sau brand
TITLE	TITLE	Titlul asociat cu o ocupație sau un statut
NATIONALITY	NAT	Naţionalitate sau referința la o țară sau

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

		regiune de origine
RELIGION	REL	Referire la o religie sau un grup religios, precum și referirile la adepții acesteia
IDENTIFIER_CREDIT_CARD_NUM	CC	Numerele cardului de credit
IDENTIFIER_EMAIL	@	Adrese de e-mail
IDENTIFIER_MONEY	M	Monede
IDENTIFIER_PERSONAL_ID_NUM	PID	Numere de identificare a persoanei
IDENTIFIER_PHONE_NUMBER	TEL	Numere de telefon
IDENTIFIER_URL	URL	Adrese Web
TEMPORAL_DATE	DATE	Date calendaristice
TEMPORAL_TIME	TIME	Ore
IDENTIFIER_DISTANCE	DIST	Distanțe
IDENTIFIER_LATITUDE_LONGITUDE	GPS	Coordonatele de latitudine și longitudine

Prin integrarea în interfața Web a rezultatelor extractorului de entități, se pot regăsi ușor cuvintele cheie în vorbirea transcrisă, cuvintele cheie extrase (împreuna cu tipurile lor) putand fi folosite în platforma pentru indexarea transcrierilor în vederea căutării rapide și facile după aceste cuvinte cheie.

IntelliDockers Text Analytics este o solutie completa de analiza automata de text si documente in format digital (scris).
Principalele functionalitati ale solutiei sunt:

- Peste 10 modele productivizate cu Deep Learning pentru diferite operatii de intelegere automata a textului.
- Procesare in peste 50 de limbi.
- Livrare on-premises in infrastructura clientului pentru securitate maxima a datelor.
- Livrare sub forma unor Docker Containers, cu timp minim de instalare si punere in productie.
- Integrare a solutiei cu ajutorul REST API.

Utilizare IntelliDockers se preteaza pentru urmatoarele scenarii:

- **Open Source Intelligence (OSINT)** - monitorizare si triaj surse deschise, website-uri, forumuri, site-uri de stiri, bloguri, etc, cu conditia ca datele sa fie deja colectate de un alt sistem.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

- **Social Media Intelligence (SOCMINT)** -monitorizare si triaj retele sociale (continut scris), cu conditia ca datele sa fie deja colectate de un alt sistem.

Functionalitati Text Analytics

Extractor Entitati

Entitățile numite, clasificate în funcție de tip (persoane, organizații, locații, produse etc.), constituie informațiile factuale de bază ale oricărui conținut. Entitățile extractor XLU extrage entitățile numite din textul dat.

Tipuri de entități

În prezent, susținem următoarea listă completă a tipurilor de entități:

Tip entitate (lung)	Scurt	Descriere
LOCAȚIA	LOC	Un oraș, stat, țară, regiune, clădire, monument, corp de apă, parc sau adresă.
ORGANIZARE	ORG	O corporație, instituție, agenție sau alt grup definit de o structură organizațională.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

PERSOANĂ	PERS	Un om identificat după nume, pseudonim sau alias.
PRODUS	PROD	Un produs de marcă produs de o organizație.
TITLU	TITLU	Denumire asociată cu o ocupație, un birou sau un statut.
NAȚIONALITATE	NAT	Trimitere la o țară sau regiune de origine.
RELIGIE	REL	Referirea la o religie sau teologie organizată, precum și la adepții ei.
IDENTIFICATOR_CREDIT_CARD_NUM	CC	Numere de card de credit.
IDENTIFICATOR_E-MAIL	@	Adrese de e-mail.
IDENTIFICATOR_BANI	M	Monede.
IDENTIFICATOR_PERSONAL_ID_NUM	PID	Numere de identificare personale.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

IDENTIFICATOR_NUMĂR_TELEFON	TEL	Numere de telefon.
IDENTIFICATOR_URL	URL-UL	Adrese web.
TEMPORAL_DATE	DATA	Data.
TIMP_TEMPORAL	E TIMPUL	E timpul.
IDENTIFICATOR_DISTANȚĂ	DIST	Distanța.
IDENTIFICATOR_LATITUDINE_LONGITUDINE	GPS	Locații geografice în coordonate latitudine și longitudine.

Limbi acceptate

Sprijinirea oficială a peste 40 de limbi, cum ar fi: Afrikaans, Arabă, Bengali, bulgară, Burmeză, Chineză, olandeză, engleză, estonă, Finlandeză, franceză, georgiană, germană, greacă, Ebraică, hind.ind, maghiară, italiană, japoneză, Javanez, kazah, Korean, Malay, Malayalam; Marathi, Persană, Poloneză, Poloneză, Română, Rusă, Sârbă, spaniolă, Swati, Tagalog, Tamil, Telugu, tailandez, turc, ucrainean, Urdu, vietnameză, Yoruba și multe altele.

Clasificator automat IPTC

Clasificarea automată a documentelor utilizând taxonomia standard IPTC (Consiliul Internațional de telecomunicații de presă - organismul global de standarde al mass-media de știri). Clasificarea bazată pe taxonomii personalizate (brevete, securitate cibernetică, informații militare sau altele) poate fi creată la cerere.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Clasificatorul IPTC în afara cutiei este instruit să clasifice documentele conform acestor clase IPTC de nivel superior.

ID-UL	Etichetă		Legătură IPTC
01000000	artă, cultură și divertisment	Probleme legate de avansarea și rafinarea minții umane, de interese, abilități, gusturi și emoții.	http://cv.iptc.org/newscodes/subjectcode/01000000
02000000	crimă, drept și justiție	Stabilirea și/sau declarația regulilor de comportament în societate, aplicarea acestor reguli, încălcarea regulilor și pedepsirea infractorilor. Organizațiile și organismele implicate în aceste activități.	http://cv.iptc.org/newscodes/subjectcode/02000000
03000000	dezastru și accident	Omul a făcut și evenimente naturale care au dus la pierderea de viață sau leziuni la creaturi vii și / sau daune la obiecte	http://cv.iptc.org/newscodes/subjectcode/03000000

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

		sau proprietăți neînsuflețite.	
04000000	economie, afaceri și finanțe	Toate aspectele legate de planificarea, producția și schimbul de avere.	http://cv.iptc.org/newscodes/subjectcode/04000000
05000000	educație	Toate aspectele de a promova cunoașterea indivizilor umani de la naștere până la moarte.	http://cv.iptc.org/newscodes/subjectcode/05000000
06000000	probleme de mediu	Toate aspectele legate de protecție, deteriorare și starea ecosistemului planetei Pământ și a împrejurimilor sale.	http://cv.iptc.org/newscodes/subjectcode/06000000
07000000	sănătate	Toate aspectele legate de bunăstarea fizică și mentală a ființelor umane.	http://cv.iptc.org/newscodes/subjectcode/07000000
08000000	interes uman	Obiecte mai ușoare despre indivizi, grupuri, animale sau obiecte.	http://cv.iptc.org/newscodes/subjectcode/08000000

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

09000000	muncă	Aspecte sociale, organizații, norme și condiții care afectează angajarea efortului uman pentru generarea de bogăție sau furnizarea de servicii și sprijinul economic al șomerilor.	http://cv.iptc.org/newscodes/subjectcode/09000000
10000000	stilul de viață și timpul liber	Activități întreprinse pentru plăcere, relaxare sau recreere în afara locurilor de muncă plătite, inclusiv pentru alimentație și călătorii.	http://cv.iptc.org/newscodes/subjectcode/10000000
11000000	politica	Exercitarea puterii la nivel local, regional, național și internațional sau lupta pentru putere și relațiile dintre organele de conducere și state.	http://cv.iptc.org/newscodes/subjectcode/11000000
12000000	religie și credință	Toate aspectele existenței umane care implică teologie, filozofie, etică și spiritualitate.	http://cv.iptc.org/newscodes/subjectcode/12000000

13000000	știință și tehnologie	Toate aspectele legate de înțelegerea umană a naturii și a lumii fizice, precum și dezvoltarea și aplicarea acestei cunoașteri	http://cv.iptc.org/newscodes/subjectcode/13000000
14000000	problemă socială	Aspecte ale comportamentului uman care afectează calitatea vieții.	http://cv.iptc.org/newscodes/subjectcode/14000000
15000000	sport	Exercițiu competitiv care implică efort fizic. Organizațiile și organismele implicate în aceste activități.	http://cv.iptc.org/newscodes/subjectcode/15000000
16000000	tulburări, conflicte și război	Acte de protest și/sau violență motivate social sau politic.	http://cv.iptc.org/newscodes/subjectcode/16000000
17000000	vremea	Studiul, raportarea și predicția fenomenelor meteorologice.	http://cv.iptc.org/newscodes/subjectcode/17000000

Limbi acceptate

Sprijinirea în prezent a următoarelor limbi (14): Arabă, chineză, engleză, farsi, franceză, Germană, ebraică, maghiară, italiană, poloneză, Portugheză, română, rusă, spaniolă.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Analiza de sentimente

Determină polaritatea oricărui conținut text: Negativ (neg), neutru (neu), pozitiv (poz).

Limbi acceptate

Sprijinirea în prezent a următoarelor limbi (14): Arabă, chineză, engleză, farsi, franceză, Germană, maghiară, italiană, japoneză, poloneză, Portugheză, română, rusă, spaniolă.

Similaritate Semantica

Motorul de similaritate semantică găsește un înțeles identic conținut de piesele de conținut analizate, ignorând sintaxa sau gramatica. Puteți compara chiar și piese de conținut scrise în diferite limbi. Acesta poate fi utilizat pentru gruparea documentelor pe baza informațiilor pe care le conțin.

Există două tipuri ale motorului de similaritate semantică:

- Una care este dependentă de limbă (capabilă să evalueze similitudinea dintre texte de aceeași limbă) și una care este agnostică limbă (puteți compara bucăți de conținut scrise în diferite limbi) care utilizează reprezentări de propoziții cu sursă deschisă
- BIBLIOTECĂ LASER pentru a calcula embeddings de propoziții multilingve. Aroma agnostică a limbajului (laser semantic similitudine motor) funcționează bine pe texte scurte - în jur de 100-200 de jetoane. Aroma dependentă de limbă nu are această limitare.

Limbi acceptate

Pentru aroma dependentă de limbă susținem în prezent următoarele limbi (9): Arabă, engleză, farsi, germană, maghiară, Italiană, poloneză, română, rusă.

Motorul de similitudine semantică CU LASER susține următoarele limbi (92): Afrikaans, albanez, Amharic, Arabă, armeană, Aymara, Azerbaidjan, bască, bielorusă, bengaleză; Limba berberă, bosniacă, Breton, bulgară, birmaneză, Catalan, Central/Kadazan Dusun, Central Khmer, Chavano, chinez; Coasta Kadazan, Cornish, Croată, Cehă, daneză, Olandeză, Est mari, engleză, Esperanto, estonă, Finlandeză, franceză, galiciană, georgiană, germană, Greacă, Hausa, ebraică, hindi, maghiară, Islandeză, Ido, indoneziană, Interlingua, Interlingue; Irlandeză,

italiană, japoneză, Kabyle, kazahă, Coreeană, kurdă, letonă, latină, Lingua Franca Nova; Lituaniană, germană/saxonă, macedoneană, malgașă, malaeză, Malaivalam, Maldivian (Divehi), Marathi, norvegian (Bokmål), Occitan; Persană (farsi), poloneză, portugheză, română, rusă, Sârbă, Sindhi, Sinhala, slovacă, slovenă, Somaleză, spaniolă, Swahili, suedeză, Tagalog, Tajik, Tamil, tătar, Telugu, tailandez, Turcă, uighur, ucraineană, urdu, uzbecă; Vietnamez, Wu chinez și Yue chinez.

Extractor de conținut

Având în vedere un URL, extrage cel mai relevant text din acesta (textul principal al articolului de știri).

Extrage numai textul relevant din articolele online, eliminând anunțurile, textul irelevant și comentariile. Acesta a fost instruit să determine și să extragă numai conținut relevant din paginile HTML, care pot fi apoi procesate folosind alte motoare NLP. Modelul utilizat pentru a determina și extrage conținutul relevant este independent de limba conținutului.

Extractor de metadate

Extrage toate metadatele publice pentru un articol specificat de URL.

Având în vedere o adresă URL a articolului, extractorul de metadate va extrage următoarele metadate:

- Titlu
- Autor
- Miniatură
- Data publicării
- Editorul
- Logo-ul editorului

Acest motor este independent de limbă.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Scor de Calitate Stiri

Calculează scorul de calitate al unui articol de știri având în vedere adresa sa URL.

Motorul trebuie configurat cu referințe la un extractor de context, un extractor de entități, un analizor de sentiment și un extractor de metadate, cu utilizarea următoarelor variabile de mediu în fișierul de licență:

- METADATE_EXTRACTOR_HOST=HTTP://[IP]:[PORT]

- CONTENT_EXTRACTOR_HOST=HTTP://[IP]:[PORT]

- ENTITATE_EXTRACTOR_HOST=HTTP://[IP]:[PORT]

- SENTIMENT_EXTRACTOR_HOST=HTTP://[IP]:[PORT]

Summarizer extins (XT)

Înțelege sensul unui text citind doar propozițiile cheie.

Această versiune utilizează statistici numerice TF și TF-IDF (frecvență-frecvență inversă a documentului) care arată cât de important este un cuvânt sau o combinație de cuvinte pentru un document. Folosind această abordare, fiecare propoziție și frază dintr-un document primește o greutate, iar algoritmul de sumarizare ia în considerare acestea.

Această abordare permite Summarizer XT IntelliDocker să gestioneze o multitudine de tipuri de documente. Setul de limbi acceptate este extins, de asemenea.

Limbi acceptate

Sprijinirea în prezent a următoarelor limbi (15): Arabă (ara), daneză (dan), olandeză (nld), engleză (eng), finlandeză (fin), Franceză (fra), germană (deu), maghiară (hun), italiană (ita), norvegiană (nor), Portugheză (por), română (ron), rusă (rus), spaniolă (spa), suedeză (swe).

NOTĂ: Când utilizați motorul, asigurați-vă că furnizați limba corectă a textului care urmează să fie procesat, deoarece aceasta este o informație crucială pentru sumarizator, astfel încât să poată împărți corect propoziții / fraze și astfel încât să poată gestiona cuvintele și frecvențele cuvintelor în mod corect.

ZA Cloud SRL, str. Govora 16A, 400664
Cluj-Napoca, România, Tel: +40 723 175 635,
Email: office@zettacloud.ai

Clusterer

Clusterer IntelliDocker îndeplinește sarcina de a grupa documente similare în grupuri (partiții), unde documentele din cadrul aceluiași cluster (partiție) prezintă un grad mai mare de similaritate între ele decât cu orice alt document din orice altă partiție (cluster).

Clusterer IntelliDocker implementează un **algoritm de grupare ierarhică aglomerativă** pe documentele de intrare, utilizând un prag de distanță semantică pentru a decide unde să pună limitele clusterelor. Începe cu punctele ca grupuri individuale și, la fiecare pas, fuzionează cea mai apropiată pereche de clustere.

Gruparea ierarhică este adesea descrisă ca o abordare mai bună a clusteringului de calitate, dar este limitată din cauza complexității timpului său pătratic. De fiecare dată când începe un proces de clustering, motorul calculează o matrice de distanță între fiecare document din lista de intrare. Asta înseamnă că motorul Clusterer trebuie să efectueze n * (n-1) / 2 calcule de distanță înainte de a putea începe procesul de clustering.

Implementarea curentă a motorului Clusterer folosește fire pentru a paraleliza cât mai mult posibil cele mai grele calcule. Nivelul paralelismului este configurabil utilizând un parametru de nivel concurabil. Fișierul de licență (YAML) este locul în care setați nivelul de paralelism prin definirea variabilei CONCURRENCY_LEVEL Environment astfel:

(...)

mediu: # nivelul implicit de concure este setat la 1 dacă această variabilă lipsește

- CONCURRENCY_LEVEL=6

Motorul nu include modalități de calcul al distanțelor dintre documentele de intrare, așa că deleagă acest calcul unei instanțe externe de similitudine semantică IntelliDocker. Desigur, orice sistem capabil să calculeze o distanță semantică relevantă între două documente ar putea fi folosit în schimb, Dar trebuie să arate ca un IntelliDocker, ceea ce înseamnă că trebuie să aveți un proxy pentru sistemul dvs. De similitudine semantică care oferă / proces și / proces-fișier endpoints cu date de intrare fiind prezentate așa cum este descris mai jos.

Fișierul de licență (YAML) este locul în care îi spui Clustererului unde poate găsi motorul de similitudine semantică. Faceți acest lucru prin definirea variabilei de mediu SIMILARITY_HOST astfel:

(...)

mediu:

- SIMILITUDINE_HOST=[host]:[port]

NOTĂ: Asigurați-vă că gazda este accesibilă din interiorul containerului Clusterer.

Fig. Diagrama motorului Clusterer

Extractor de text (ai/OCR)

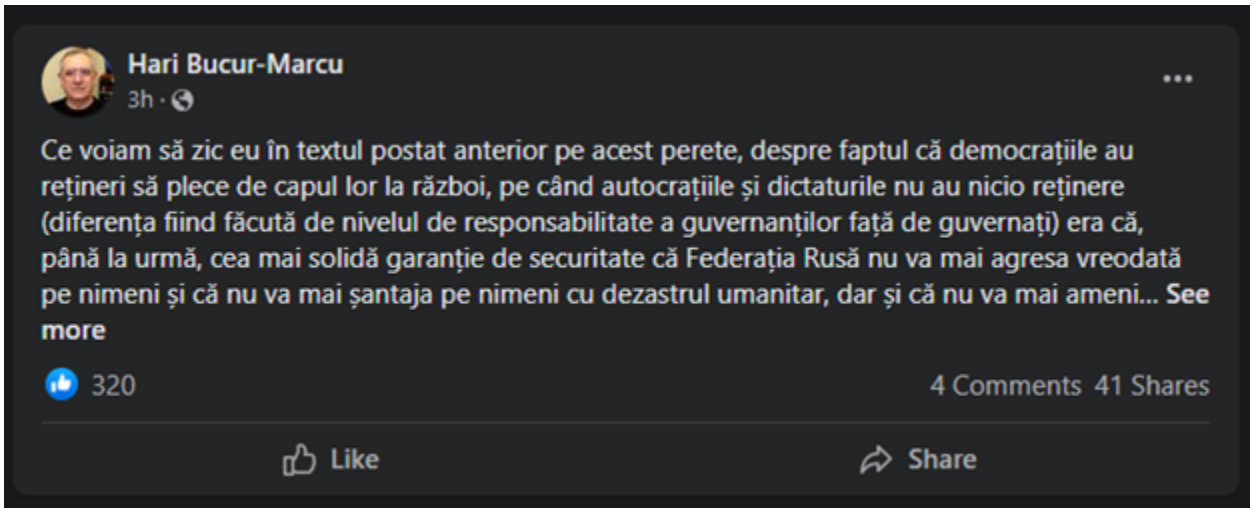
Transcribes text dintr-o imagine dată, scanare (.jpg, .png, .tif) sau fișier PDF folosind un model de recunoaștere text (computer Vision).

Limbi acceptate

În prezent susținem următoarele limbi (5): Arabă, engleză, germană, italiană, română.

Exemplu de utilizare

Rularea extractor de text pentru limba română pe imaginea de mai jos (captura de ecran a unei postări pe Facebook) va extrage tot textul afișat cu posibilitatea de a recunoaște limba specifică diacritice.



curl -X POST "http://localhost:8989/rest/process-file" -H "accept: application/json" -H "Content-Type: multipart/form-data" -F "content=@facebook-post-ron.png;type=image/png"

Rezultatul este:

{

"text": "[en] Hari Bucur-Marcu E\ncae EI)\n\nCe voiam să zic eu în textul postat anterior pe acest perete, despre faptul că democrațiile au\nreţineri să plece de capul lor la război, pe când autocrațiile și dictaturile nu au nicio reținere\n(diferenţa fiind făcută de nivelul de responsabilitate a guvernanților față de guvernați) era că,\n\npână la urmă, cea mai solidă garanţie de securitate că Federaţia Rusă nu va mai agresa vreodată\npe nimeni și că nu va mai şantaja pe nimeni cu dezastrul umanitar, dar și că nu va mai ameni... See\nmore\n\n \n\n& 320 4 Comments 41 Shares\n\nbi că PI)\n\f"

}

