

AKTIVITETI 08

LABORATORI I IDENTITETEVE SINTETIKE

Krijimi i te Dhenave Testuese per Sistemet e Zbulimit

Workshop: AI dhe Dokumentet Biometrike - DITA 2

Dita 2 Mengjes | Aktiviteti 8 nga 12 | Kohezgjatja: 75 minuta

Objektivat e te Nxenit

1. Kuptoni pse te dhenat sintetike jane esenciale per trajnim
2. Mesoni si te gjeneroni identitete realiste shqiptare
3. Njihni 12 llojet e gabimeve te zakonshme
4. Praktikoni zbulimin e gabimeve ne dataset sintetik
5. Krijoni dataset tuajin sintetik per perdonim te ardhshem

Dataset: 50 identitete (14 te sakta, 36 me gabime)

Pse te Dhena Sintetike?

Te Dhena Reale	Te Dhena Sintetike
Privatesi - risk i larte	Privatesi - zero risk
Sasi e kufizuar	Sasi e pakufizuar
Mund te mungojne gabime	Gabimet kontrollohen
Veshtire per t'u ndare	Lehte per t'u ndare
Ligjore komplekse	Pa kufizime ligjore

Per trajnim dhe testim, te dhenat sintetike jane zgjidhja me e sigurt dhe me e efektshme

Komponentet e nje Identiteti Sintetik

Komponenti	Shembull	Burime Referimi
Emri + Mbiemri	Agron Hoxha	Lista emrash shqiptare
Gjinia	M / F	Bazuar ne emer
Datelindja	15/03/1987	Random 18-70 vjec
Vendlindja	Tirane	Lista qytetesh shqiptare
NID	1234567890	10 shifra random
Telefoni	+355691234567	Prefiks 66/67/68/69
Email	agron.hoxha@email.com	I gjeneruar nga emri
Adresa	Rr. Durresit, Nr. 50	Rruge + numer random

12 Llojet e Gabimeve (Pjesa 1)

Lloji	Serioziteti	Shembull
Gabim Shkrimi (Typo)	LOW	Agorn -> Agron
Format Date i Gabuar	MEDIUM	1987/03/15 -> 15/03/1987
Date e Pavlefshme	HIGH	31/02/1990
NID Gjatesi Gabim	HIGH	12345678 (8 shifra)
NID Format Gabim	HIGH	I234567890 (shkronje)
Telefon Format Gabim	MEDIUM	069-12-345

12 Llojet e Gabimeve (Pjesa 2)

Lloji	Serioziteti	Shembull
Email i Pavlefshem	MEDIUM	emri@ (mungon domain)
Mosperputhje Moshe	HIGH	Lindje 1990, Mosha 45
Date ne te Ardhmen	CRITICAL	Datelindja 2026
Mosperputhje Gjini-Emer	MEDIUM	Agron, Gjinia: F
Shifra Kontrolluese	CRITICAL	MRZ check digit gabim
NID i Dyfishte	CRITICAL	Dy persona, nje NID

CRITICAL = Tregues i forte mashtrimi | HIGH = Kerkon verifikim | MEDIUM/LOW = Mund te jete typo

Strategjia e Zbulimit

Niveli 1 - Kontrolle Automatike:

- Gjatesia e fushave (NID=10, tel=12+ etj.)
- Formati (vetem numra ne NID, @ ne email)
- Validiteti datave (jo ne te ardhmen, jo pamundur)

Niveli 2 - Kontrolle Logjike:

- Konsistenza (gjinia-emer, moshe-datelindja)
- Unikiteti (NID i dyfishte)
- Shifrat kontrolluese (MRZ)

Niveli 3 - Kontrolle Kontekstuale:

- Vendndodhjet reale (qytete qe ekzistojne)
- Rruge qe ekzistojne ne qytetin e dhene

Perdorimi i AI per Zbulimin

Je ekspert i zbulimit te gabimeve ne te dhena identiteti.

Analizo kete rekord dhe identifiko problemet:

KONTROLLET:

1. NID: 10 shifra, vetem numra
2. Telefoni: +355 + 2 shifra prefix + 7 shifra
3. Email: format valid (x@y.z)
4. Datelindja: format DD/MM/YYYY, jo ne te ardhmen
5. Konsistenza gjini-emer
6. Moshe logjike (18-120)

RAPORTO:

- GABIMET: lista e detajuar
- SERIOZITETI: LOW/MEDIUM/HIGH/CRITICAL
- VEPRIMI: Korrigjo / Verifiko / Refuzo

Detyra Juaj (50 Minuta)

Faza 1 (10 min): Studioni llojet e gabimeve

- Lexoni referencen e 12 llojeve

Faza 2 (25 min): Analizoni datasetin

- Kontrolloni 20 rekorde nga dataseti
- Identifikoni gabimet dhe seriozitetin
- Dokumentoni gjetjet

Faza 3 (15 min): Krijoni 5 identitete tuajat

- 3 te sakta, 2 me gabime te qellimshme
- Ndani me koleget per verifikim

Statistikat e Datasetit

Totali: 50 identitete

Te sakta: 14 (28%)

Me gabime: 36 (72%)

Sipas Seriozitetit:

Serioziteti	Numri	Pershkrimi
NONE	14	Pa gabim
LOW	4	Typo e thjeshte
MEDIUM	14	Kerkon korrigjim
HIGH	13	Kerkon verifikim
CRITICAL	5	Tregues mashtrimi

Mjetet dhe Burimet

Mjeti/Burimi	Perdorimi	URL/Referencia
FakeNameGenerator	Emra realiste	fakenamegenerator.com
Mockaroo	Dataset i personalizuar	mockaroo.com
GenerateData	Te dhena te strukturuara	generatedata.com
Python Faker	Librari programatike	pip install faker
AI (Claude/GPT)	Gjenerim me kontekst	Prompt engineering

Per kontekst shqiptar, kombinoni mjetet me lista vendore emrash/vendesh

Praktikat me te Mira

[Realiste] Perdorni emra dhe vendndodhje reale shqiptare

[Proporcionale] 30% rekorde te sakta per benchmark

[Te shumellojshme] Perfshini te gjitha llojet e gabimeve

[Dokumentuar] Shenoni gabimet e qellimshme per verifikim

[Riperditesuar] Perditesoni datasetin rregullisht

Konsiderata Etike

- **KURRE mos perdorni te dhena reale pa leje**
- Te dhenat sintetike NUK duhet te ngathen me te dhena reale
- Shenojini qartesisht si 'SINTETIKE - VETEM PER TRAJNIM'
- Fshini pas perdomit nese nuk nevojiten me
- Kujdes: Kombinime te dhena te shkeputura mund te krijojne identitet real

Te dhenat sintetike duhet te perdoren vetem per qellime trajnimi dhe testimi

Pikat Kyce

[Sintetike] Te dhenat sintetike jane esenciale per trajnim te sigurt

[12 Lloje] Njihni llojet e gabimeve per t'i zbuluar

[AI Ndhimon] Perdorni AI per te kontrolluar sistematikisht

[Etike] Kurre mos i ngaterni me te dhena reale

Ne vazhdim: Aktiviteti 09 - Sfida e Validimit Kaskade