AKTIVITETI 03
GJUETARI I IDENTITETEVE TE DYFISHTA

Logjika e Deduplikimit me AI

Workshop: AI dhe Dokumentet Biometrike

Dita 1 | Aktiviteti 3 nga 6 | Kohezgjatja: 60 minuta

Objektivat e te Nxenit

1. Kuptoni problemin e rekordeve duplikate ne databaza qeveritare
2. Mesoni teknikat e fuzzy matching per identifikimin e duplikateve
3. Perdorni AI per te zbuluar cifte duplikate me variacione
4. Vleresoni besueshmmerine e perputhjeve (HIGH/MEDIUM/LOW)
5. Dokumentoni dhe raportoni duplikatet e gjetura

Dataset: 200 rekorde me 15 cifte duplikate te fshehura

Problemi: Duplikatet ne Databaza

Pse ndodhin duplikatet?

* Futja manuale nga operatore te ndryshem
* Migrimi i te dhenave nga sisteme te vjetra
* Mungesa e validimit ne kohe reale
* Ndryshimi i emrave (martese, korrigjime)
* Gabime shkrimi dhe transkriptimi

Pasoja: Statistika te gabuara, sherbime te dyfishta, kosto te shtuara

Tipet e Duplikateve

Tipi	Shembull Origjinal	Shembull Duplikat
Kapitalizim	Agron Hoxha	AGRON hoxha
Gabim shkrimi	Gentiana	Genntiana
Variacion emri	Gezim	Gëzim
Format date	1987-03-15	15/03/1987
Shkurtim adrese	Rruga Skenderbeu	Rr. Skenderbeu
Format telefoni	+355681234567	068-123-4567
Email variant	agron.hoxha@	agron.hoxha1@

Fuzzy Matching - Koncepti

Cfare eshte Fuzzy Matching?

Teknika per gjetjen e stringave qe jane 'te ngjashme' por jo identike. Perdor algoritme si Levenshtein distance, Soundex, Jaro-Winkler.

Shembull Levenshtein Distance:

String 1	String 2	Distanca	Ngjashmeria
Agron	Agron	0	100%
Agron	agron	1	80%
Agron	Agroon	1	83%
Agron	Petrit	6	0%

Nivelat e Besueshmmerise

Niveli	Pershkrimi	Veprimi
HIGH	Perputhje e forte ne shume fusha (emri, datelindja, ID)	Shqyrtim i shpejte, bashkim i mundshem
MEDIUM	Perputhje ne disa fusha, ndryshime te vogla	Kerkon verifikim manual
LOW	Ngjashmeri e dobet, mund te jene persona te ndryshem	Investigim i plote para vendimit

Gjithmone verifikoni manualisht perpara bashkimit te rekordeve!

Qasja me AI

Pse AI per deduplikim?

* Kupton kontekstin (Tirana = Tirane)
* Trajton variacione te shumta njekohesisht
* Shpjegon arsyen e perputhjes
* Vlereson besueshmmerine automatikisht
* Procedon qindra rekorde ne minuta

Shablon Prompt-i

```
Kam nje dataset me 200 rekorde qytetaresh shqiptare.
Identifiko ciftet e mundshme duplikate duke krahasuar:
- Emri dhe mbiemri (konsidero variacione, gabime, kapitalizim)
- Datelindja (formate te ndryshme)
- Adresa (shkurtime, variante qytetesh)
- Telefoni (formate te ndryshme)
- Email (variante te vogla)
- Kodi ID (gabime te mundshme)

Per cdo cift te dyshuar, raporto:
1. ID e rekordeve (REC-XXXX, REC-YYYY)
2. Fushat qe perputhjen
3. Fushat qe ndryshojne
4. Niveli i besueshmmerise (HIGH/MEDIUM/LOW)
5. Rekomandimi (BASHKO/VERIFIKO/INJORO)
```

Detyra Juaj (40 Minuta)

1. Hapni A03_rekordet_qytetareve.csv (200 rekorde)

2. Ngarkoni ne Claude/Gemini

3. Aplikoni prompt-in e deduplikimit

4. Per cdo cift te gjetur, dokumentoni:

- ID-te e rekordeve

- Tipin e variacionit

- Nivelin e besueshmmerise

5. Krahasoni rezultatet me koleget

SFIDE: Gjeni te 15 ciftet duplikate!

Shembull i Raportit te Deduplikimit

Fusha	Vlera
Cifti	REC-0023 <-> REC-0187
Emri Rek.1	Agron Hoxha
Emri Rek.2	AGRON hoxha
Datelindja	Identike: 1987-03-15
Adresa	Rek.1: Rruga Skenderbeu / Rek.2: Rr. Skenderbeu
Variacion	NAME_CASE + ADDRESS_ABBREV
Besueshmeria	HIGH
Rekomandimi	BASHKO (pas verifikimit)

Praktikat me te Mira

[Verifikim] Asnjehere mos bashkoni rekorde automatikisht pa verifikim

[Dokumentim] Regjistroni arsyen e cdo bashkimi per auditim

[Prioritet] Filloni me perputhjet HIGH, pastaj MEDIUM

[Kontekst] Konsideroni qe grate mund te kene ndryshuar mbiemrin

[Backup] Ruani kopje te datasetit origjinal para bashkimit

Pikat Kyçe

[Problem] Duplikatet jane problem i zakonshem ne databaza qeveritare

[Fuzzy] Fuzzy matching gjen perputhje te peraferta, jo vetem identike

[AI] AI kupton kontekstin dhe variacuonet me mire se rregullat e thjeshta

[Besueshmeri] Klasifikoni perputhjet ne HIGH/MEDIUM/LOW per prioritizim

Ne vazhdim: Aktiviteti 04 - Simulimi i Auditit te Paragjykimit