Ingegneria dei dati 2025/2026

Homework 6

(da svolgere in gruppo)

Paolo Merialdo

Homework 6

L'obiettivo del progetto è integrare i dati su automobili disponibili da diverse sorgenti. Considerare i dati disponibili in queste sorgenti:

- https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data *
- https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset *

1. Per ciascuna sorgente analizzare la percentuale di valori nulli e di valori unici di ciascun attributo
2. Definire uno schema mediato
3. Allineare le sorgenti allo schema mediato
4. Implementare e valutare una strategia di record linkage (vedi prox slide)

Homework 6

4. Record linkage

4.A Generare un ground-truth per valutare diversi soluzioni e addestrare modelli di ML. Per questo passaggio sfruttiamo la presenza dell'attributo VIN, che rappresenta il numero di telaio. Tuttavia, dobbiamo prestare attenzione al fatto che i dati sono rumorosi; a tal fine, definire una strategia per fare verifiche ad hoc e pulire i dati (senza perderne la rappresentatività). Se si decide di produrre una ground-truth curata a mano, valutare l'uso di Label Studio

4.B Eliminare gli attributi VIN dai due dataset e dalla ground-truth .

4.C Dalla ground-truth creare tre dataset (training, validation, test)

4.D Definire due strategie di blocking B1 e B2

4.E Definire regole di record linkage con la libreria Python Record Linkage

4.F Addestrare un modello usando la libreria Python Dedupe

4.G Addestrare un modello con la versione di Ditto disponibile a questo indirizzo: https://github.com/MarcoNapoleone/FAIR-DA4ER

4.H Valutare le prestazioni di diverse pipeline (B1-dedupe, B2-dedupe, B1-RecordLikage, B2-RecordLikage, B1-ditto, B2-ditto) in termini di precision, recall, F1-measure, tempi di training, tempi di inferenza

Homework 6

- Preparare una relazione di circa 10 pagine che descrive le principali sfide affrontate nel progetto, la caratterizzazione delle fonti e, in dettaglio, la valutazione sperimentale.
- Preparare una presentazione di 20' che descrive architettura e valutazione sperimentale della soluzione.

Termini di consegna: il giorno prima dell'esame.

Caricare la relazione e la presentazione attraverso il seguente modulo:

https://forms.office.com/e/EbqbR7dvK4