

Ingegneria dei dati
Homework 5
(da svolgere in gruppo)

Paolo Merialdo

Homework 5: data integration

Obiettivo: integrare le sorgenti dati su Aziende presenti nel repository: [homework](#)

Analizzare le TUTTE sorgenti dati e individuare le principali eterogeneità

1. Definire uno schema mediato opportuno che abbia almeno 20 attributi. Allineare gli schemi delle sorgenti allo schema mediato. È possibile usare:
 - Una soluzione manuale
 - Una soluzione custom basata su chatGPT
 - FlexMatcher <https://flexmatcher.readthedocs.io/en/latest/>
 - Coma <https://sourceforge.net/projects/coma-ce/>
 - Uno dei tool del progetto Valentine <https://github.com/delftdata/valentine>
2. Popolare lo schema mediato con i dati delle sorgenti
3. Calcolare il Record linkage:
 - Creare una ground-truth con almeno 100 coppie in matching. Accertarsi che la ground-truth contenga anche casi "difficili"
 - Definire almeno due diverse strategie di blocking
 - Calcolare il pairwise matching con la libreria Python Record Linkage Toolkit <https://recordlinkage.readthedocs.io/en/latest/> (offre anche soluzioni per il blocking) a partire dalle diverse strategie di blocking scelte e confrontare accuratamente i risultati
 - Calcolare il pairwise matching con uno strumento alternativo tra i seguenti:
 - DeepMatcher (soluzione neural network) <https://github.com/anhaidgroup/deepmatcher>
 - Ditto (soluzione neural network) <https://github.com/megagonlabs/ditto>
 - EMT (soluzione neural network molto simile a Ditto) <https://github.com/brunnurs/entity-matching-transformer>
 - Confrontare i risultati che si ottengono usando diverse combinazioni di blocking e di pairwise matching

Termini di consegna

- Preparare un documento scritto di 4 pagine e una presentazione di 15' che descrivano:
 - Le caratteristiche salienti delle sorgenti
 - Lo schema mediato
 - Le soluzioni che avete scelto per integrare i dati
 - Le prestazioni (in termini di precision, recall, F-measure, tempi di calcolo, sforzo umano) confrontando le diverse soluzioni
- Il documento e la presentazione vanno consegnati il giorno prima dell'orale sul modulo all'indirizzo:

<https://forms.office.com/e/5nmvtKgY11>

Il progetto sarà quindi discusso il giorno dell'orale

Date Orale

- 27 gennaio 2025 ore 14.00
 - 05 febbraio 2025 ore 14.00
 - 13 febbraio 2025 ore 9.00
 - 20 febbraio 2025 ore 14.00
 - 27 febbraio 2025 ore 9.00
-
- Indicare le proprie preferenze compilando questo modulo (entro il 19 gen 2025):

<https://forms.office.com/e/vzrws4Up3i>