



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Informatica

## Tecniche di Record Linkage

Alberici Federico - 808058

Bettini Ivo Junior - 806878

Cocca Umberto - 807191

Traversa Silvia - 816435

**Anno Accademico 2019 - 2020**

---

# Indice

<b>Ricerca</b> . . . . .	<b>2</b>
Record Linkage - introduzione . . . . .	2
Metodologie (?) di Record Linkage . . . . .	2
Tools usati . . . . .	2
Python Record Linkage Toolkit . . . . .	2
<b>Esperimenti</b> . . . . .	<b>2</b>
Dataset . . . . .	2
Pulizia dei dati . . . . .	2
Risultati . . . . .	3
Record Linkage . . . . .	3
Gold Standard . . . . .	3

---

## Ricerca

### Record Linkage - introduzione

### Metodologie (?) di Record Linkage

### Tools usati

#### Python Record Linkage Toolkit

Python Record Linkage Toolkit è una libreria che permette di effettuare record linkage sia in una sola fonte di dati che in multiple. Il package contiene metodi di indexing, come blocking e sorted neighbourhood indexing, funzioni per il confronto con diverse misure di similarità possibili e diversi algoritmi di classificazione, sia supervisionati che non.

## Esperimenti

### Dataset

Il dataset scelto per effettuare i nostri esperimenti contiene elenchi di ristoranti di Manhattan provenienti da 12 siti web, presi settimanalmente da Gennaio a Marzo 2009. Sono riportati in tutti il nome del ristorante, l'indirizzo e la città.

### Pulizia dei dati

Prima di applicare le tecniche di record linkage è stata effettuata una pulizia dei dati. Avevamo a disposizione 7 file .txt, uno per ogni settimana considerata, contenente dati appartenenti a tutti i siti web. In totale erano presenti 215555 record, suddivisi come illustrato nella tabella:

---

<b>file</b>	<b>records</b>
restaurants_2009_1_22.txt	30401
restaurants_2009_1_29.txt	30775
restaurants_2009_2_05.txt	30805
restaurants_2009_2_12.txt	30863
restaurants_2009_2_19.txt	30876
restaurants_2009_2_26.txt	30898
restaurants_2009_3_12.txt	30937

Table 1: Record presenti nei file txt forniti

In particolare, per ogni ristorante sono presenti il seguente numero di record:

<b>restaurant</b>	<b>records</b>
ActiveDiner	6184
DiningGuide	814
FoodBuzz	2079
MenuPages	13143
NewYork	1774
NYPmag	5124
NYTimes	3095
OpenTable	1539
SavoryCities	4536
TasteSpace	3635
TimeOut	14007
VillageVoice	2684

Table 2: Record per ristorante

**Risultati**

**Record Linkage**

**Gold Standard**