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

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:

file	records
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

Risultati

Record Linkage

Gold Standard