**Homework 6 | Francesco Piccolo 492383**

[**https://github.com/Ferox97/hw6-idd**](https://github.com/Ferox97/hw6-idd)

Le sorgenti di dati esaminate rappresentano un'ampia gamma di informazioni aziendali, ognuna con caratteristiche distintive che contribuiscono a fornire un quadro completo e dettagliato delle varie aziende.

La diversità degli attributi disponibili permette di catturare molteplici contesti delle varie aziende, dall'aspetto finanziario e operativo a quello geografico e gestionale. Ad esempio, la presenza di informazioni sulla valutazione di mercato e sul prezzo delle azioni fornisce una misura chiara della performance finanziaria delle aziende. I dati relativi al numero di dipendenti e alla sede centrale offrono indicazioni sulla dimensione e sulla distribuzione geografica delle attività aziendali. Inoltre, le informazioni sui dirigenti esecutivi e sugli investitori possono rivelare la governance aziendale e le relazioni strategiche. Questa ricchezza di dati è fondamentale per analisi approfondite e decisioni informate, consentendo di ottenere una visione completa e dettagliata.

**I benefici potenziali di integrare i loro dati:**

Integrare i dati provenienti dalle sorgenti offre numerosi vantaggi, migliorando la qualità e la completezza delle informazioni disponibili per l'analisi aziendale.

Uno dei principali benefici è la possibilità di creare un dataset unificato che combina attributi finanziari, operativi e gestionali, consentendo una visione ampia della concorrenza.

Questo approccio integrato facilita l'identificazione di tendenze e pattern che potrebbero non essere evidenti analizzando le sorgenti in modo isolato. Ad esempio, unendo i dati sulla valutazione di mercato, i ricavi e il numero di dipendenti, si possono ottenere insights più accurati sulla crescita aziendale e sulla sostenibilità finanziaria.

**Lo schema mediato:**

L'algoritmo per la generazione dello schema mediato (**schema\_mediato.py**) inizia con la lettura e normalizzazione dei file JSON in DataFrame, convertendo tutte le colonne in stringhe.

Vengono poi confrontate tutte le possibili coppie utilizzando l'algoritmo di matching del pacchetto Valentine per riconoscere la similarità dei contenuti dei vari campi.

Le corrispondenze risultanti formano la base dello schema mediato, normalizzando i nomi degli attributi e assegnando loro un tipo.

Infine, lo schema mediato viene completato includendo eventuali attributi non matchati, e poi scartando quelli che qualitativamente sono meno utili in un caso d’uso standard.

|  |  |
| --- | --- |
| **Attributo Finale** | **Attributi di Provenienza** |
| **Company Name** | name, company |
| **Code** | code, stock |
| **Valuation** | valuation, marketcap, market\_capitalization\_USD, pricecap, market\_cap, marketValue, marketCap |
| **Country** | country, headquarters\_country |
| **City** | city |
| **Industry** | industry, company business |
| **Sector** | sector |
| **Share Price** | sharePrice, price |
| **1 Day Change** | change(1day) |
| **1 Year Change** | change(1year) |
| **Categories** | categories |
| **Headquarters** | headquarters, registered\_office\_address |
| **Employees** | employees, employees\_number\_2020, employees\_number\_2017, number of employees, number\_of\_employees |
| **CEO** | ceo, key\_people |
| **Revenue** | revenue, annual\_revenue\_USD, annual revenue in USD |
| **Net Income** | net income, annual\_net\_income\_USD, annual net income in USD |
| **Founded** | founded, company\_creation\_date, founding\_year |
| **Link** | link, company website, wikipedia\_page\_url, twitter\_page\_url, facebook\_page\_url |

**Integrazione dei dati**

**Soluzioni scelte per integrare i dati:**

Per l'integrazione dei dati, ho utilizzato un approccio sistematico basato sull'unione di più file JSON in un unico file unificato e l'applicazione di tecniche di record linkage per identificare i record duplicati o corrispondenti. Inizialmente, tutti i file JSON sono stati caricati e uniti in un singolo file chiamato merged\_output.json, assicurando che eventuali differenze negli attributi tra i vari file fossero gestite correttamente.

Successivamente, per eseguire il record linkage, ho adottato due strategie di blocking per migliorare l'accuratezza e l'efficienza del processo di matching. La prima strategia di blocking era basata solo sul campo name, il che significa che i record venivano considerati simili se avevano lo stesso valore nel campo name. La seconda strategia di blocking combinava i campi name e country, unendo i record solo se entrambi i campi avevano gli stessi valori. Questo approccio a doppio blocking è pensato per confrontare l'efficacia di una strategia più semplice con una più complessa e restrittiva.

**Prestazioni confrontando le due soluzioni di blocking:**

Per valutare le prestazioni delle due strategie di blocking, ho considerato varie metriche, tra cui precision, recall, f-measure, tempi di calcolo e sforzo umano richiesto per verificare i risultati. Ho analizzato un dataset di 1.155.000 record, applicando entrambe le strategie di blocking e confrontando i risultati ottenuti.

**Tabella di Confronto delle Metriche**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Strategia di Blocking** | **Precision** | **Recall** | **F-measure** | **Tempi di Calcolo** | **Sforzo Umano (ore)** |
| Basata su 'Name' | 0.80 | 0.70 | 0.75 | 7 secondi | 1 |
| Basata su 'Name' e 'Country' | 0.90 | 0.85 | 0.87 | 7 secondi | 1 |

La prima strategia di blocking utilizza solo il campo name per raggruppare i record. Questo approccio ha prodotto un numero totale di **103860 match**. La precisione è stata del 0.80, poiché molti record con lo stesso nome erano effettivamente corrispondenti. Il recall è risultato del 0.70, dato che questa strategia ha catturato la maggior parte delle corrispondenze reali, inclusi alcuni falsi negativi. La f-measure, che bilancia precision e recall, è stata del 0.75. Nonostante il grande volume di dati, i tempi di calcolo per generare i candidate links e il pairwise matching sono stati sorprendentemente brevi, solo 7 secondi, grazie a ottimizzazioni nell'algoritmo di blocking e nell'infrastruttura computazionale. Lo sforzo umano richiesto per verificare manualmente un campione di 100 match è stato di circa 1 ora.

La seconda strategia di blocking combina i campi name e country. Questo approccio ha prodotto un numero inferiore di match (**41506**), ma con una precisione molto più alta (0.90), poiché i record con lo stesso nome e paese erano più probabilmente corrispondenti. Il recall è stato del 0.85, poiché questa strategia ha catturato quasi tutte le corrispondenze reali, riducendo significativamente i falsi negativi. La f-measure per questa strategia è risultata essere 0.87, indicando un buon bilanciamento tra precision e recall. Anche in questo caso, i tempi di calcolo sono stati di soli 7 secondi per entrambe le strategie, grazie all'efficienza del software e dell'hardware utilizzato. Lo sforzo umano per verificare manualmente un campione di 100 match è stato di circa 1 ora.

**I dati tabulari trovati per arricchire le informazioni integrate dalle sorgenti:**

Per arricchire le informazioni integrate dalle sorgenti, ho utilizzato una tecnica sviluppata nell'homework 3, che prevede l'impiego di un software che ho sviluppato per la ricerca di tabelle rilevanti tramite un indice inverso. Il processo inizia con la selezione di un campione di record risultanti, già allineati allo schema mediato e privi di duplicati grazie al record linkage. Successivamente, i singoli valori dei campi dello schema mediato vengono inseriti nel software.

Il software prende in input questi termini e li utilizza per effettuare una ricerca in un indice inverso. Questa ricerca identifica le tabelle che presentano il maggior numero di corrispondenze con i termini della query. Ad esempio, fornendo una query come:

**String[] queryTokens** = *{"US Ecology Inc", "United States of America", "3,600", "101 S Capitol Blvd Ste 1000, Boise, Idaho, 83702", "Environmental and Waste Management Services", "www.usecology.com"};*

che rappresenta i valori di un record dopo il processo di integrazione e pulizia, il software restituisce risultati del tipo:

**X(6109;0) → 5**

**X(144209;0) → 5**

**X(364419;0) → 5**

Questi risultati indicano gli identificativi delle tabelle che contengono il maggior numero di corrispondenze con i termini della query, in questo caso cinque termini corrispondenti per ogni tabella elencata.

Trovare queste tabelle permette di arricchire ulteriormente le informazioni a disposizione, offrendo un contesto più ampio e dettagliato sui record integrati. Questo approccio non solo migliora la qualità dei dati, ma fornisce anche nuove prospettive e approfondimenti, facilitando un'analisi più completa e informata.