Architetture Dati

# Comparazione di misure di similarità tra stringhe nell'ambito del record linkage

Lorenzo Pirola - matr. 816418

# Indice

# 1 Introduzione

Lo scopo di questo progetto è quello di confrontare l'efficienza di diversi metodi per il confronto approssimativo di stringhe applicate nell'ambito del record linkage. Il progetto è basato sull'articolo *"A Comparison of Personal Name Matching: Techniques and Practical Issues"*[1] di P.Christen, nel quale le diverse tecniche sono utilizzate con l'obiettivo specifico di confrontare i nomi di persona. In questo progetto le misure verranno utilizzate per confrontare nomi commerciali di società. Più precisamente, le misure verranno utilizzate per identificare quali tra le 2000 aziende più grandi e influenti al mondo secondo Forbes[1] sono anche presenti nell'indice di borsa statunitense Standard & Poor 500[2].
I vari records sono stati uniti attraverso un record linkage basato su soglia, mentre le prestazioni dei singoli metodi sono state valutate grazie a un terzo dataset, il quale è stato costruito per essere etichettato come ground truth.
Il progetto è stato realizzato in Python (versione 3.8). Sia il codice sia i datasets utilizzati sono stati caricati su in una repository pubblica su GitHub che può essere consultata al seguente link: `https://github.com/lpirola13/ArchitetturaDati/`.

# 2 Datasets

## 2.1 Introduzione

I due dataset principali impiegati nella realizzazione del progetto sono entrambi pubblici e reperibili sulla piattaforma Kaggle:

1. Il primo dataset[3] è relativo alle aziende pubbliche americane quotate in borsa e appartenenti all'indice S&P 500. Nelle sezione successive verrà indicato brevemente con il nome *SP500*.

2. Il secondo dataset[4] è relativo alla classifica delle 2000 aziende più influenti al mondo stilata da Forbes. Nelle sezioni successive verrà indicato brevemente con il nome *FORBES*.

Il terzo dataset, utilizzato per la valutazione delle tecniche di confronto, è stato costruito appositamente e indica semplicemente quali aziende appartengono sia al dataset *FORBES* sia al dataset *SP500*.

## 2.2 SP500

L'indice di borsa S&P 500 segue l'andamento di un paniere azionario costituito dalle 500 aziende statunitensi di maggiore capitalizzazione, le quali sono quotate alla borsa di New York (NYSE), all' American Stock Exchange (AMEX) o al NASDAQ[5]. Il dataset *SP500* contiene le informazioni basilari di queste aziende ed è costituito da 505 records e 8 colonne. Il motivo per il quale i records sono 505 e non 500 è perchè 5 aziende possiedono azioni di due classi differenti (A e B), utili a differenziare i diritti degli azionisti.
Nella tabella 1 sono riportate i nomi delle colonne del dataset e una loro breve descrizione.

| Colonna | Descrizione |
|---|---|
| Ticker symbol | Simbolo azionario |
| Security | Nome dell'azienda |
| SEC filings | Tipologia di rendiconto finanziario da depositare alla SEC |
| GICS Sector | Settore in rispetto al Global Industry Classification Standard (GICS[6]) |
| GICS Sub Industry | Sotto-Industria in rispetto al Global Industry Classification Standard (GICS) |
| Address of Headquarters | Indirizzo della sede centrale |
| Date first added | Anno in cui l'azienda è stata aggiunta all'indice |
| CIK | Central Index Key, usato dalla SEC per identificare le aziende |

Tabella 1: Campi del dataset SP500

[1] `https://www.forbes.com/global2000`
[2] `https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview`
[3] `https://www.kaggle.com/dgawlik/nyse?select=securities.csv`
[4] `https://www.kaggle.com/ash316/forbes-top-2000-companies`
[5] `https://it.wikipedia.org/wiki/S%26P_500`

## 2.3 FORBES

Ogni anno la rivista statunitense di economia Forbes pubblica una lista delle prime 2000 aziende pubbliche al mondo. La posizione di ogni azienda nella classifica è definita sulla base di quattro metriche: fatturato, utile, attivo e capitalizzazione di mercato[6]. Il dataset *FORBES* contiene tutte queste informazioni ed è costituito da 2000 records e 9 colonne.
Nella tabella 2 sono riportate i nomi delle colonne del dataset e una loro breve descrizione.

| Colonna | Descrizione |
|---|---|
| Rank | Posizionamento nella classifica |
| Company | Nome dell'azienda |
| Country | Nazione |
| Sales | Fatturato (in miliardi di dollari) |
| Profit | Utile (in miliardi di dollari) |
| Assets | Attivo (in miliardi di dollari) |
| Market Value | Capitalizzazione di mercato (in miliardi di dollari) |
| Sector | Settore di riferimento |
| Industry | Industria di riferimento |

Tabella 2: Campi del dataset FORBES

## 2.4 Valutazione della Data Quality

Prima di proseguire con la fase di preprocessing è stata valutata rapidamente la qualità complessiva dei dati contenuti nei due datasets.
Innanzitutto, entrambi i datasets fanno riferimento all'anno 2017. Grazie alle informazioni contenute nel campo *Date first added* del dataset *SP500* si apprende che l'ultima azienda è stata aggiunta nel 1 maggio 2017. I dati contenuti nel dataset *FORBES*, invece, sono relativi al 7 aprile 2017, come indicato dalla stessa rivista in un articolo[6]. Perciò, anche se i dati non sono aggiornati o gli ultimi disponibili, tutte le società presenti in *FORBES* sono potenzialmente contenute anche in *SP500*.
Per quanto riguarda l'accuratezza semantica dei dati, bisogna notare che non tutti nomi commerciali delle aziende nel campo *Company* del dataset *FORBES* sono completamente accurati: sono presenti alcune abbreviazioni o mancano alcuni identificativi relativi alla forma societaria. Sono proprio questo tipo di errori che potrebbero influire sul processo di record linkage. Nei campi restanti non sono presenti errori gravi di altro tipo.
Considerando l'accuratezza sintattica, l'attenzione si focalizza sui campi *GICS Sector*, *GICS Industry* di *SP500* e *Sector*, *Industry* di FORBES. Sebbene entrambi facciano riferimento al Global Industry Classification Standard[7] del 2017, alcuni settori o industrie presenti nei datasets non rispecchiano quelli indicati nello standard.
In merito alla completezza, come è possibile notare nella figura 1, attraverso una breve analisi sui Missing Values sono stati individuati 198 valori mancanti nella colonna *Date first added* nel dataset *SP500*. Questi valori mancanti non rappresentano un problema, dato che la relativa colonna non è stata ritenuta utile durante il processo di record linkage. Osservando la figura 2, si osserva che nel dataset *FORBES* sono presenti 197 e 491 valori mancanti nelle rispettive colonne *Sector* e *industry*. Le azioni prese circa questi valori mancanti saranno approfondite nelle prossime sezioni.
Fatte queste considerazioni sulla qualità dei dati, è possibile continuare con il processo di record linkage.

[6] https://www.forbes.com/sites/andreamurphy/2017/05/24/2017-global-2000-methodology-how-we-crunch-the-numbers/#4e86fafe61d4
[7] https://www.msci.com/gics

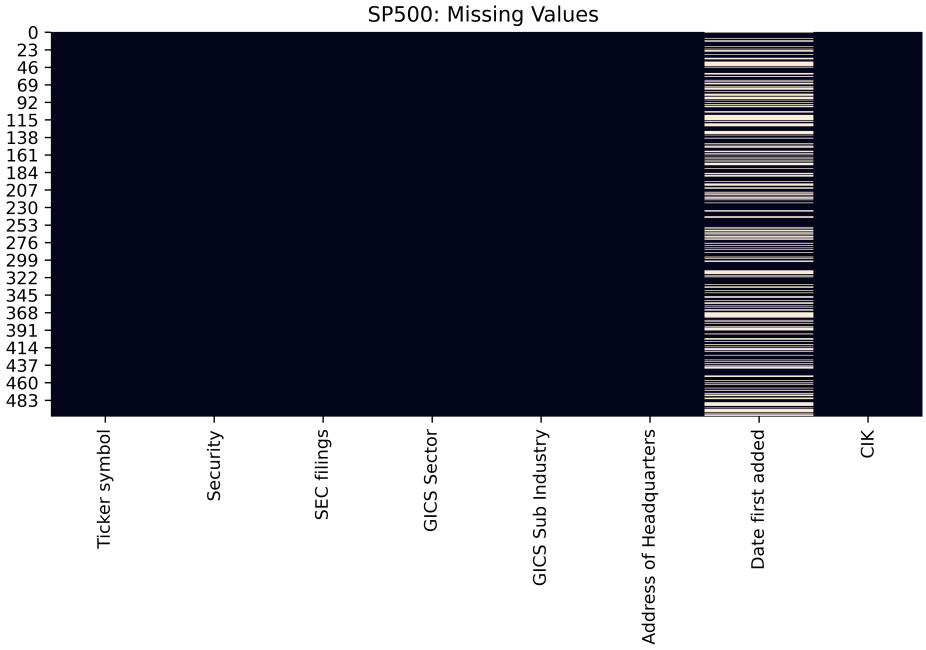



Figura 1: Occorrenze dei Missing Values in SP500


Figura 2: Occorrenze dei Missing Values in FORBES

# 3 Record Linkage

## 3.1 Introduzione

L'obiettivo del progetto è quello di valutare le performance di diversi metodi per il confronto approssimativo di stringhe nell'ambito del record linkage. La tipologia di record linkage impiegata è quella basata su soglia, mentre le prestazioni delle varie misure sono state valutate in base al numero di matches previsti correttamente e al tempo impiegato per il confronto.
Tutte le funzioni utilizzate e necessarie allo scopo fanno parte della libreria Python Record Linkage Toolkit[8]. Come metodi per il confronto di stringhe sono stati scelti gli stessi messi a disposizione dalla libreria.

[8] `https://recordlinkage.readthedocs.io/en/latest/about.html`

## 3.2 Preprocessing

Una fase molto importante per il record linkage è il pre-processing dei dati. In questo caso si è prestata molta attenzione alla standardizzazione e alla pulizia dei dati.
Innanzitutto, grazie alla funzione *clean* messa a disposizione dalla libreria, tutti i caratteri sono stati convertiti in minuscolo, mentre tutti i caratteri non necessari (Es. &, %, $) e i simboli di punteggiatura sono stati rimossi.
Successivamente, visto che nei datasets sono presenti nomi di società americane, sono stati rimossi tutti gli identificativi relativi alla forma societaria che spesso appaiono nei loro nomi. Più precisamente sono stati rimossi gli identificativi *Corp., Inc., Co., Ltd., Plc., Lpc., Hldg., The.*
Per quanto riguarda i valori mancanti nelle colonne *Sector* e *Industry* di *FORBES* non è stata presa nessuna particolare decisione.

## 3.3 Indexing

Per ridurre i potenziali confronti e quindi diminuire la complessità computazionale è stata impiegata la tecnica di indexing.
L'idea iniziale era quella di applicare la tecnica di blocking tradizionale, andando ad utilizzare come chiavi una delle coppie di campi *GICS Sector* e *Sector* oppure *GICS Industry* e *Industry*. Però, considerando l'elevato numero di valori mancanti nelle due colonne, questa idea è stata scartata. Una soluzione poteva essere quella di non considerare i records in cui è presente un valore mancante in una delle due colonne, ma così facendo non verrebbero esaminati 491 records, ovvero circa il 25% del totale. Una soluzione alternativa comprendeva la sostituzione del valore mancante con il più frequente, ma anch'essa è stata scartata poiché ogni scelta sbagliata avrebbe escluso un record dal confronto.
Rimossa questa possibilità, l'approccio rimasto è stato il sorted neighborhood. La chiave di ordinamento scelta, sulla quale sono stati costituiti i vari blocchi, è il nome della società, ovvero le colonne *Security* di *SP500* e *Company* di *FORBES*. Per evitare che degli errori di battitura influissero sull'ordinamento dei nomi è stata utilizzata anche una funzione di codifica fonetica. In particolare, è stata impiegata la funzione Soundex, mentre la finestra di potenziali candidati è stata fissata a 31.
Grazie all'impiego dell'indexing si è passati da un numero totale di 1010000 confronti a 24907 confronti.

## 3.4 Confronto

Il confronto delle stringhe è avvenuto utilizzando le apposite funzioni messe a disposizione dalla libreria. Esse sono descritte brevemente nell'articolo [1]:

- La misura di distanza Levenshtein, meglio conosciuta come la classica distanza di edit.
- La misura di distanza Demerau-Levenshtein, che aggiunge alla distanza di edit l'operazione di trasposizione.
- La misura di distanza Jaro, comunemente usata per il matching di nomi nei sistemi di data linkage. Considera gli inserimenti, le cancellazioni e le trasposizioni.
- La misura di distanza Jaro-Winkler che estende la distanza di Jaro aumentando la similarità tra due stringhe se esse condividono la stessa parte iniziale.
- La misura di distanza Smith-Waterman, simile alla distanza di edit, ma considera anche intervalli di caratteri e pesi per caratteri specifici. Per questo motivo la misura risulta appropriata per nomi composti in cui compaiono iniziali o abbreviazioni.
- La misura di similarità basata sull'algoritmo Longest Common Sub-string (LCS). Anche questa misura risulta appropriata per nomi composti, ma nel caso in cui le parole al loro interno sono scambiate di posizione.

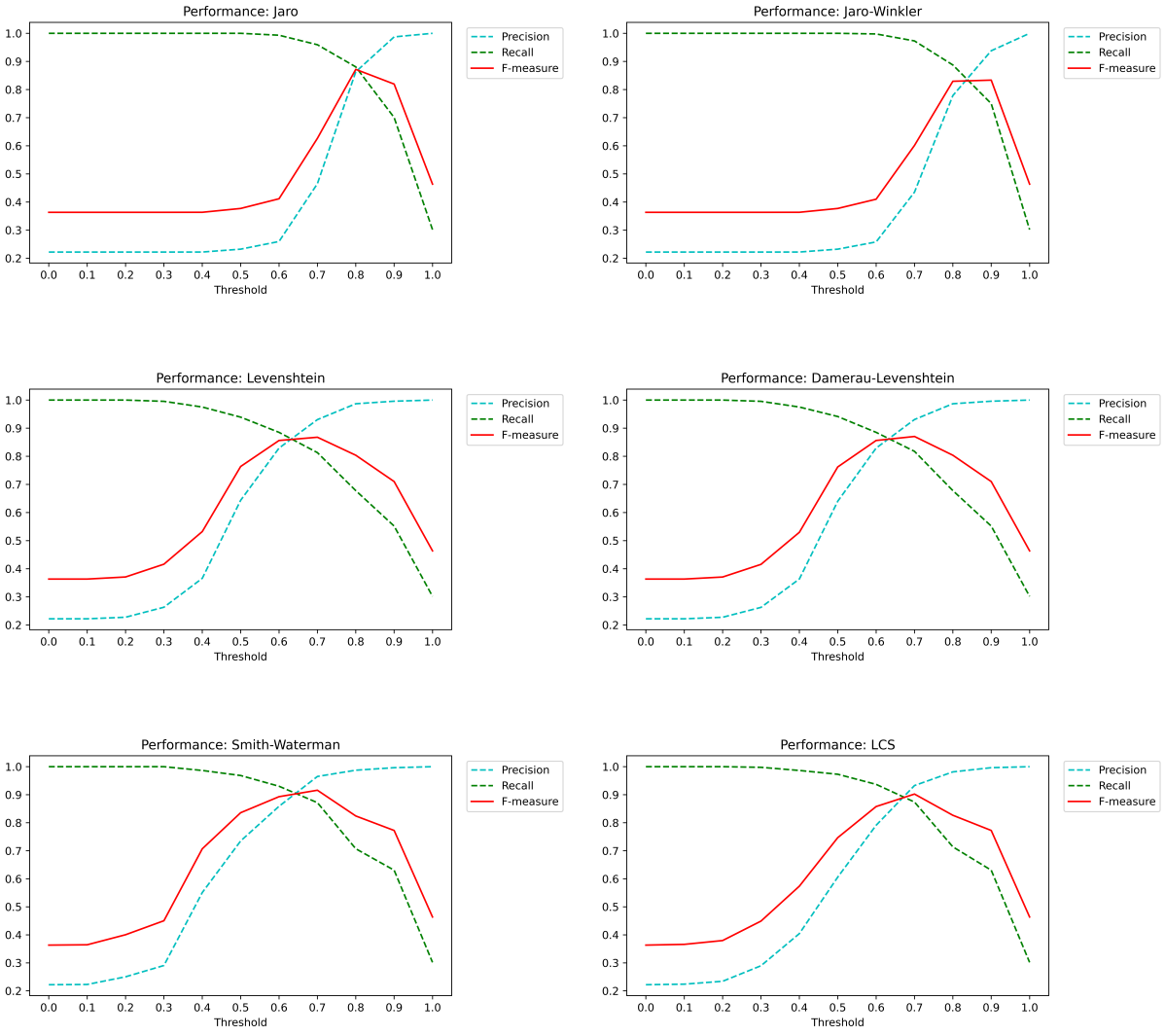
Il confronto è avvenuto sulle coppie di campi *Security* di *SP500* con *Company* di *FORBES* e *GICS Sector* di *SP500* con *Sector* di *FORBES*. Quindi, per ogni candidato identificato durante la fase di indexing sono stati calcolati due valori di similarità, entrambi compresi tra 0 e 1, applicando ognuno dei metodi elencati precedentemente.

### 3.5 Classificazione

La classificazione dei record in match e non-match è avvenuta utilizzando un approccio basato su soglia. Il valore di similarità finale di ogni record è stato calcolato attraverso una funzione che attribuisce un peso maggiore alla similarità tra i nomi commerciali delle aziende rispetto ai settori. Questa decisione è stata presa considerando che la colonna settore relativa al dataset *FORBES* contiene diversi valori mancanti, i quali avrebbero potuto impattare sul risultato. La funzione utilizzata associa al nome della società un peso pari al 75% e al settore il restante 25% e restituisce un valore compreso tra 0 e 1.
Le prestazioni delle varie misure di similarità sono state valutate al variare del valore soglia. Più precisamente, i valori soglia utilizzati sono $0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$. Quindi ogni coppia di records con un valore di similarità maggiore o uguale al valore di soglia predefinito è stato considerato come un match, i restanti sono stati considerati come non-match.
Infine, i risultati della classificazione ottenuti variando il valore di soglia sono stati confrontati con i valori reali contenuti nel terzo dataset. Sono stati dunque calcolati i valori di *precision*, *recall* e *f-measure*. Inoltre, è stato calcolato un valore medio di f-measure e un valore medio relativo al tempo impiegato.

## 4 Risultati

I risultati sono stati riassunti in grafici e tabelle per un migliore confronto. Ogni grafico è relativo ad una misura di similarità e su ognuno di essi sono riportati i valori di precision (azzurro), di recall (verde) e di f-measure (rosso) al variare del valore di soglia.



| Metodo | F-measure |
|---|---|
| Jaro | 0.49 |
| Jaro-Winkler | 0.485 |
| Levenshtein | 0.592 |
| Damearau-Levenshtein | 0.592 |
| Smith-Waterman | 0.635 |
| LCS | 0.609 |

Tabella 3: Valore di F-measure medio

Nella tabella 3 sono riportati, per ogni misura di similarità, i valori medi di f-measure ottenuti durante i test.
La complessità computazionale è stata misurata considerando il tempo di esecuzione dei confronti. Per ogni misura di similarità è stata calcolata una media di tempi di esecuzione al variare del valore di soglia. Questi tempi medi sono riportati nella figura 4 e nella tabella 3.

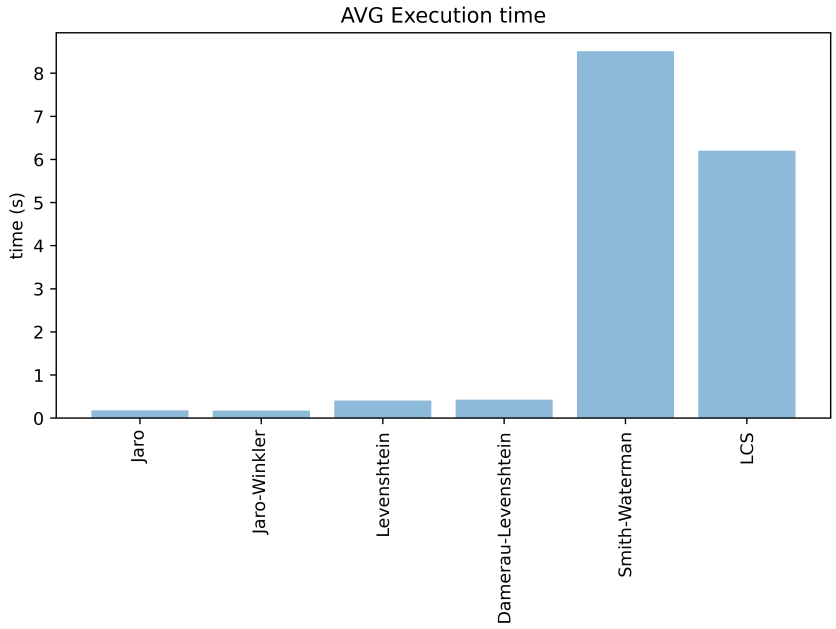



Figura 3: Tempo di esecuzione medio

| Metodo | Tempo |
|---|---|
| Jaro | 0.199 |
| Jaro-Winkler | 0.178 |
| Levenshtein | 0.404 |
| Damearau-Levenshtein | 0.43 |
| Smith-Waterman | 8.438 |
| LCS | 6.09 |

Tabella 4: Tempo di esecuzione medio

### 4.1 Analisi dei risultati

Dai risultati in tabella 3 si evince che le prestazioni migliori le hanno ottenute le misure Smith-Waterman e LCS. Questo risultato è giustificato dalla presenza di molti nomi composti all'interno dei due datasets e, come già anticipato nella sezione 3.4, le due misure erano le più adatte a identificare questo tipo di stringhe. A differenza di quanto espresso nell'articolo, i metodi Jaro e Jaro-Winkler sono risultati i meno adatti. Ciò potrebbe essere giustificato dal fatto che nei due datasets gli errori di battitura sono pressoché assenti, a differenza dei nomi di persona contenuti nei datasets utilizzati nell'articolo.
Per quanto riguarda i tempi di esecuzione, il metodo più veloce è risultato essere lo Jaro-Winkler mentre il più lento lo Smith-Waterman. Anche questi risultati sembrano essere in linea con quanto indicato nell'articolo.

## 5 Conclusioni

Nel progetto sono state confrontate diverse tecniche per il confronto approssimativo di stringhe nell'ambito del record linkage. Queste tecniche sono state valutate tenendo in considerazione due aspetti: la qualità dei risultati prodotti e la complessità computazionale richiesta. I risultati, come indicato anche nell'articolo, indicano che non esiste una tecnica migliore delle altre, ma è possibile scegliere la tecnica più adatta in base tipologia di dati su cui bisogna lavorare. Nella scelta della tecnica bisogna anche considerare anche la complessità computazionale, specialmente nel caso in cui il volume dei dati è grande o quando sono richiesti dei risultati in breve tempo.

## Riferimenti bibliografici

[1] P. Christen. A comparison of personal name matching: Techniques and practical issues. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*. IEEE, 2006.