Comparazione di misure di similarità tra stringhe nell’ambito del record linkage

Architetture dati

Lorenzo Pirola - Matricola 816418

Introduzione

- Il progetto è basato sull'articolo "*A Comparison of Personal Name Matching: Techniques and Practical Issues*" di Peter Christen.
- **SCOPO**: valutare le prestazioni di alcune misure di similarità nell'ambito del record linkage. Le prestazioni sono state valutate in termini di qualità dei matches e di tempo di esecuzione.
- **AMBITO**: nomi commerciali di società
- **STRUMENTI**: Python Record Linkage Toolkit

Datasets

SP500

- Contiene la lista delle 500 società statunitensi di maggiore capitalizzazione inserite all'interno dell'indice di borsa Standard & Poor 500
- 505 records e 8 colonne

Colonna	Descrizione
Ticker Symbol	Simbolo azionario
Security	Nome della società
SEC Filings	Tipo di rendiconto finanziario
GICS Sector	Settore in rispetto al GICS
GICS Sub Industry	Sotto-Industria in rispetto al GICS
Address of Headquarters	Indirizzo della sede centrale
Date first added	Anno di inserimento nell'indice
CIK	Central Index Key

Datasets

FORBES

- Contiene la classifica delle prime 2000 società pubbliche al mondo stilata dalla rivista di economia Forbes
- 2000 records e 9 colonne

Colonna	Descrizione
Rank	Posizionamento
Company	Nome della società
Country	Nazione
Sales	Fatturato (in miliardi di $)
Profit	Utile (in miliardi di $)
Assets	Attivo (in miliardi di $)
Market Value	Capitalizzazione di mercato (in miliardi di $)
Sector	Settore di riferimento
Industry	Industria di riferimento

Datasets

Data quality

	SP500	FORBES
Accuratezza Temporale	1 Maggio 2017	7 Aprile 2017
Accuratezza Semantica	✔	Nomi delle società
Accuratezza Sintattica	Settore e Industria	Settore e Industria
Completezza	Data di primo inserimento (198)	Settore e Industria (197 e 491)

Record Linkage

Pre-Processing

- Conversione di tutti i caratteri in minuscolo
- Rimozione dei caratteri non necessari (*&*, *%*, *$*, ecc.) e della punteggiatura
- Rimozione degli identificativi della forma societaria (*Inc*., *Co*., *Ltd*., ecc.)

Arthur J. Gallagher & Co. → arthur j gallagher

Packaged Foods & Meats → packaged foods meats

Record Linkage

Indexing

- **APPROCCIO**: Sorted Neighborhood
- **CHIAVE**: Nomi delle società
- **FUNZIONE DI CODIFICA FONETICA**: Soundex
- Numero di confronti ridotto da 1010000 a 25907

Record Linkage

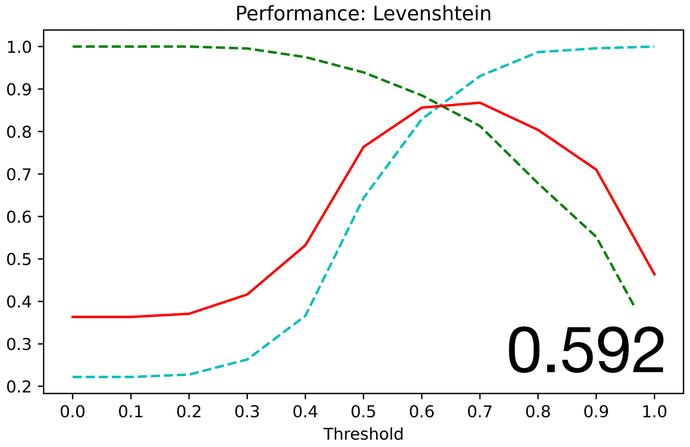
Confronto e Classificazione

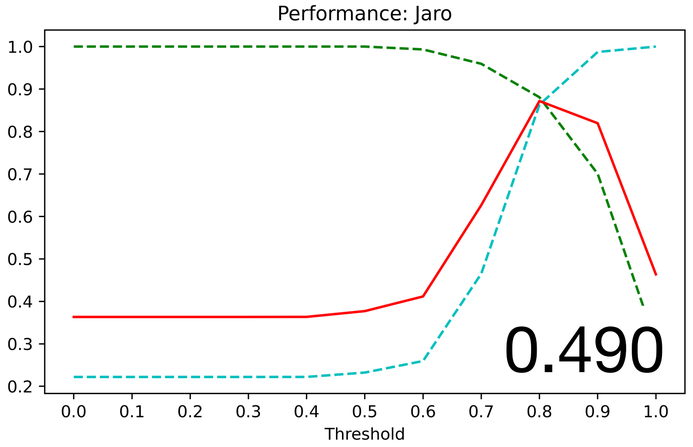
- **MISURE DI SIMILARITA’**: Jaro, Jaro-Winkler, Levenshtein, Damerau-Levenshtein, Smith-Waterman, LCS
- **CONFRONTO**: *Security* e *Company* e *GICS Sector* e *Sector*
- Impiego di un record linkage basato su soglia con funzione peso (75% Nome della società e 25% settore)
- Utilizzo di diversi valori soglia: *0*, *0.1*, *0.2*, *0.3*, *0.4*, *0.5*, *0.6*, *0.7*, *0.8*, *0.9*, *1*

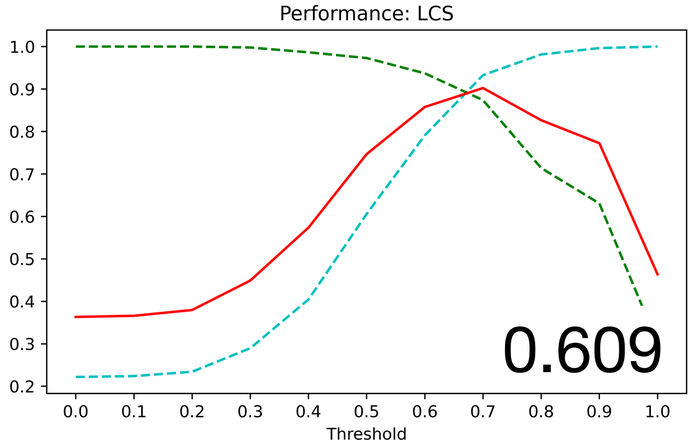
Risultati

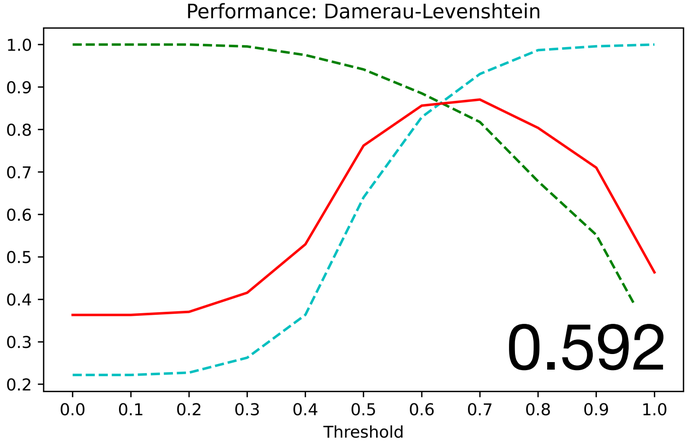
Performance

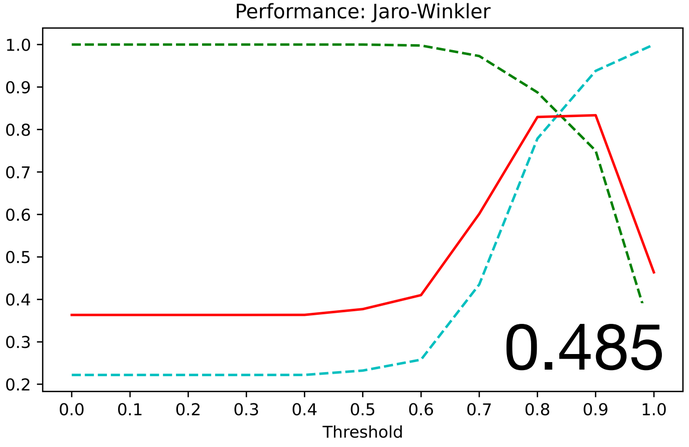
Precision Recall F-measure

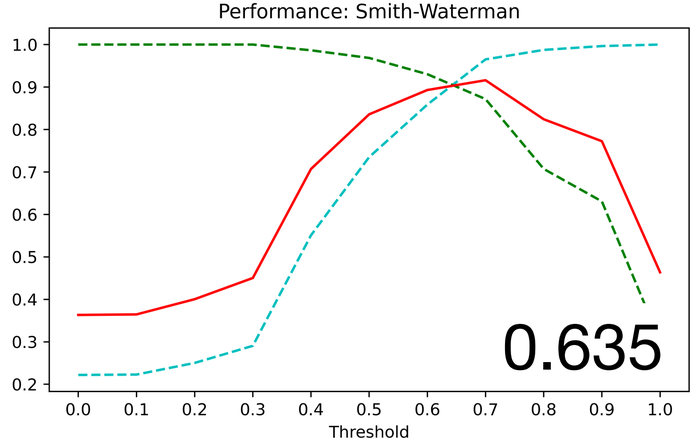






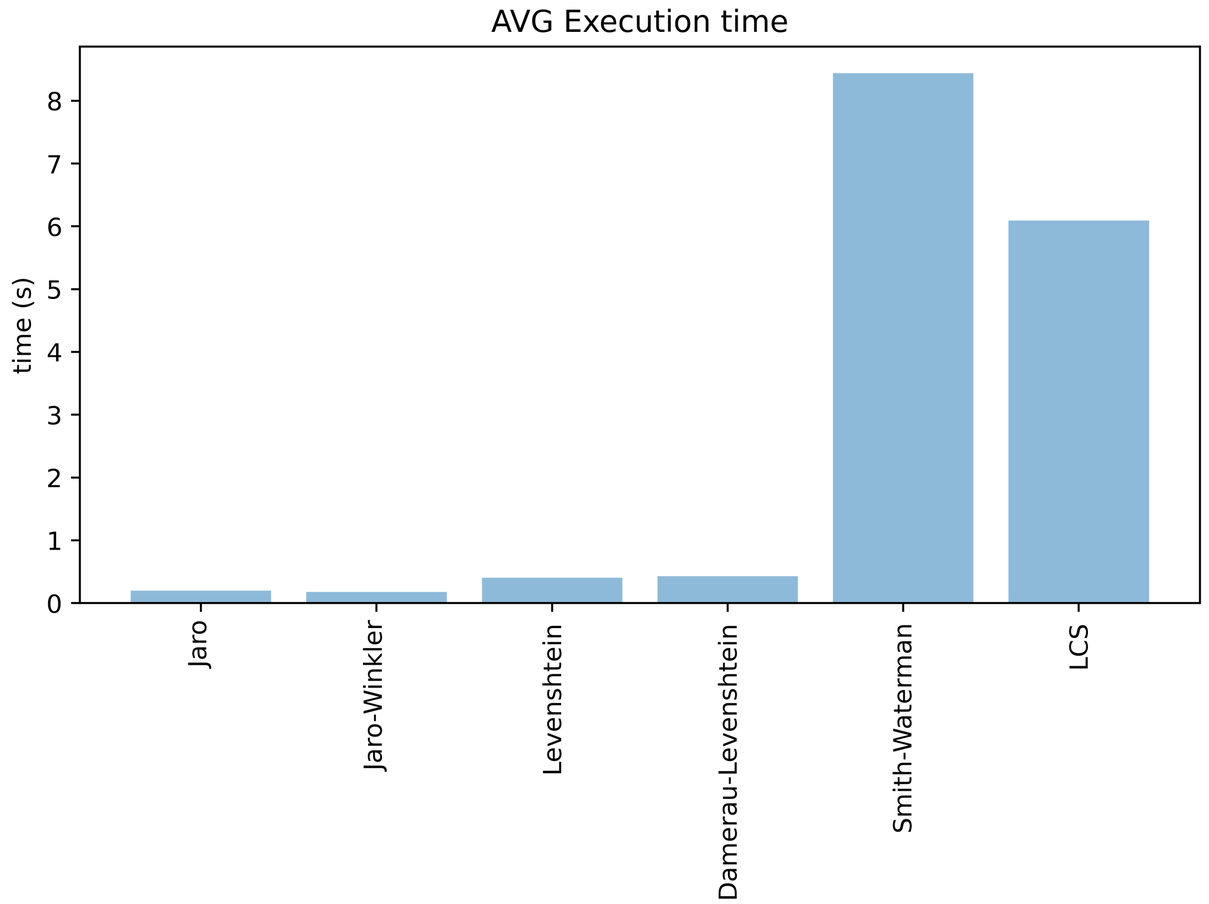






Risultati

Complessità computazionale



Metodo	Tempo (s)
Jaro	0,199
Jaro-Winkler	0,178
Levenshtein	0,404
Damerau-Levenshtein	0,430
Smith-Waterman	8,438
LCS	6,090

Conclusioni

- Non esiste una tecnica migliore delle altre, occorre scegliere la più adatta al formato dei dati.
- Occorre prestare attenzione alla complessità computazionale delle varie tecniche, considerando il volume dei dati da processare.