UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA
Dipartimento di ingegneria Enzo Ferrari
Corso di laurea Ingegneria Informatica

Documentazione e aggiunte alla libreria per l'entity resolution SparkER

Relatori:
Prof.ssa Sonia Bergamaschi
PhD Luca Gagliardelli

Candidato:
Arturo Bianchi
Matricola n° 132615

ANNO ACCADEMICO 2020/2021

SOMMARIO

- Entity resolution applicato a una situazione reale
- SparkER, cos'è e quali limitazioni presenta
- L'aggiunta di nuovi formati
- Semplificazione del workflow
- Documentazione dei metodi

ENTITY RESOLUTION

- L'entity resolution (ER) è un insieme di tecniche per identificare tutte le istanze di un oggetto, indipendentemente dal formato o dalla fonte.
- Presenta il problema della complessità quadratica.
- Non scala bene nell'ambito dei big data.

SCENARIO DI UTILIZZO REALE

- Creare un record unico in cui i profili corrispondenti allo stesso oggetto abbiano uno stesso identificatore;
- Utilizzare uno strumento di deduplication o linkage per sfruttare le descrizioni delle entità.

Catalog A

PID	Title	Description
P123X	Syringe 10x10	Syringe 10 ml 10 pack
P123Y	Syringe 10x100	Syringe 10 ml 100 pack
P456A	Insulin needle 4x10	Hypodermic insulin needle 4 mm 10 pack
...	...	...

Catalog B

ID	Name	Description
1	Syringes 10 ml	Syringe 10 ml x 10 pieces
2	Syringes 10 ml big pack	Syringe 10 ml x 100 pieces
3	Small needle	Needle for insulin 4 mm 10 pieces
...	...	...

SPARKER

- Una libreria open-source per python sviluppata in UNIMORE dal DBGROUP, già presente su GitHub
- Utilizza come ambiente il framework di Apache Spark, che consente di lavorare in modo efficiente nell'ambito dei big data, permettendo la scalabilità dell'ER
- SparkER implementa le tecniche di entity resolution, non presenti di default su Spark.

COSA SI PUÒ MIGLIORARE

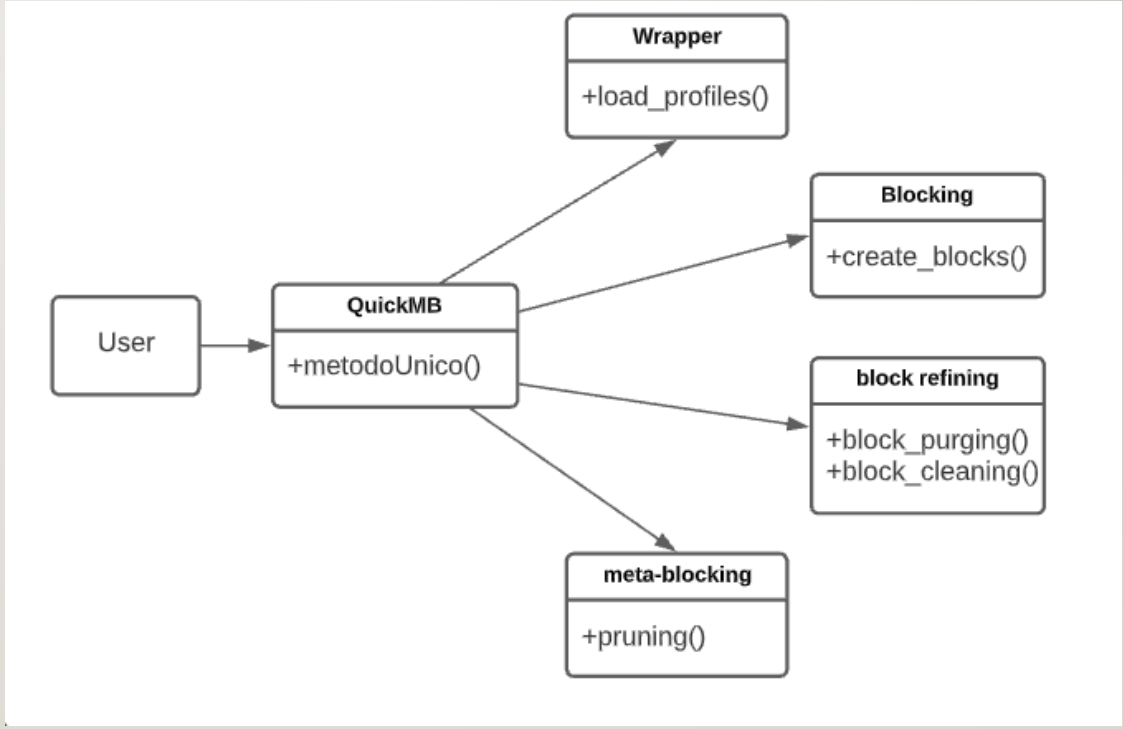
- Numero limitato di formati per caricare i profili;
- Workflow molto lungo con metodi che richiedono molti parametri;
- Mancanza di documentazione dei metodi;

AGGIUNTA DI NUOVI FORMATI

- Utilizzo del framework di pandas;
- Utilizzo del wrapper per i dataframe di pandas;
- Creazione di dataframe da dataset xlsx(excel) utilizzando l'engine openpyxl;
- Ottenimento di dataframe da query sql usando l'engine sqlalchemy.

FACADE PATTERN

- Utilizzo del design pattern facade (facciata) per implementare un'interfaccia più semplice;
- Implementazione per data linkage (due dataset), data deduplication (dataset singolo) e BLAST.

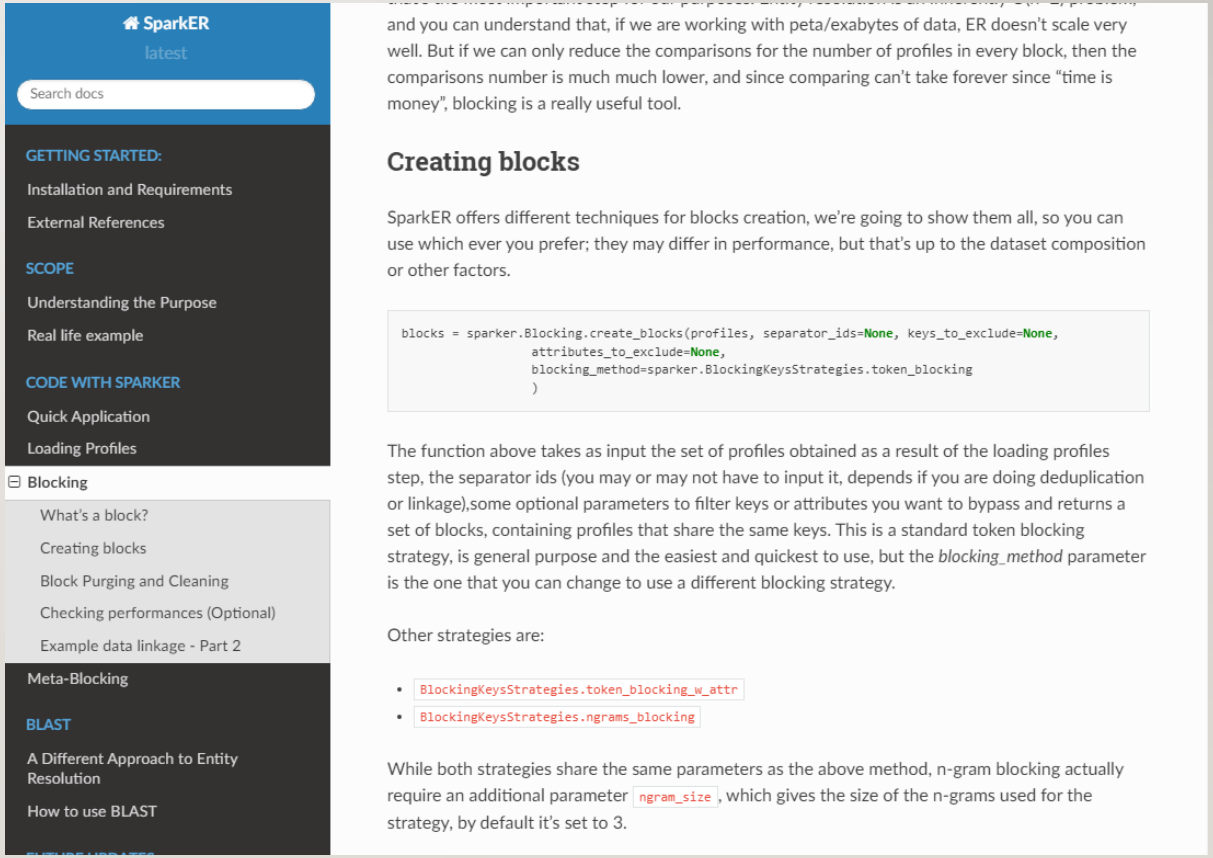


STRUMENTI PER LA DOCUMENTAZIONE

- Servizio di hosting con Read The Docs
- Sphinx e reStructuredText come strumento e linguaggio di markup per la realizzazione delle pagine.

STRUTTURA DELLA DOCUMENTAZIONE

- La documentazione spiega i metodi per l'implementazione del blocking/meta-blocking/BLAST (in inglese).
- Documentazione più pratica che teorica.
- Suddivisa in 5 sezioni, partendo dall'installazione della libreria e arrivando all'analisi dei risultati dei processi.



CONCLUSIONI

Lavorare con SparkER ha permesso:

- Un primo approccio al mondo dei big data, in particolare alla teoria dietro l'entity resolution;
- Studio e utilizzo di framework per l'analisi dei dati come Spark e librerie software come pandas;
- Scoperta e studio di alcuni degli strumenti nell'ambito dell'open-source.