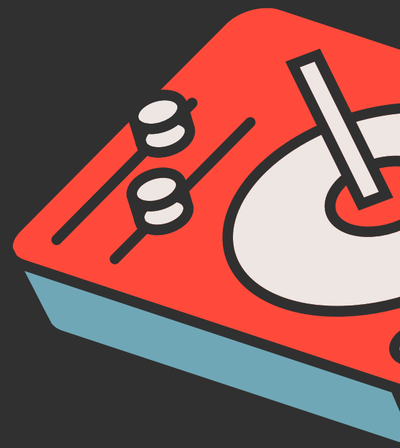
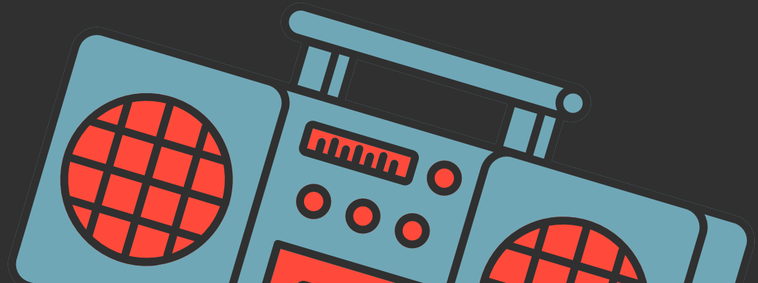
UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA
CORSO DI LAUREA MAGISTRALE IN DATA SCIENCE



Progetto di Data Management
Grafo di brani e artisti delle ultime dieci edizioni del Festival di Sanremo.

Andrea Maver - 828775
Simone Radaelli - 845065
Oscar Zanotti - 873763

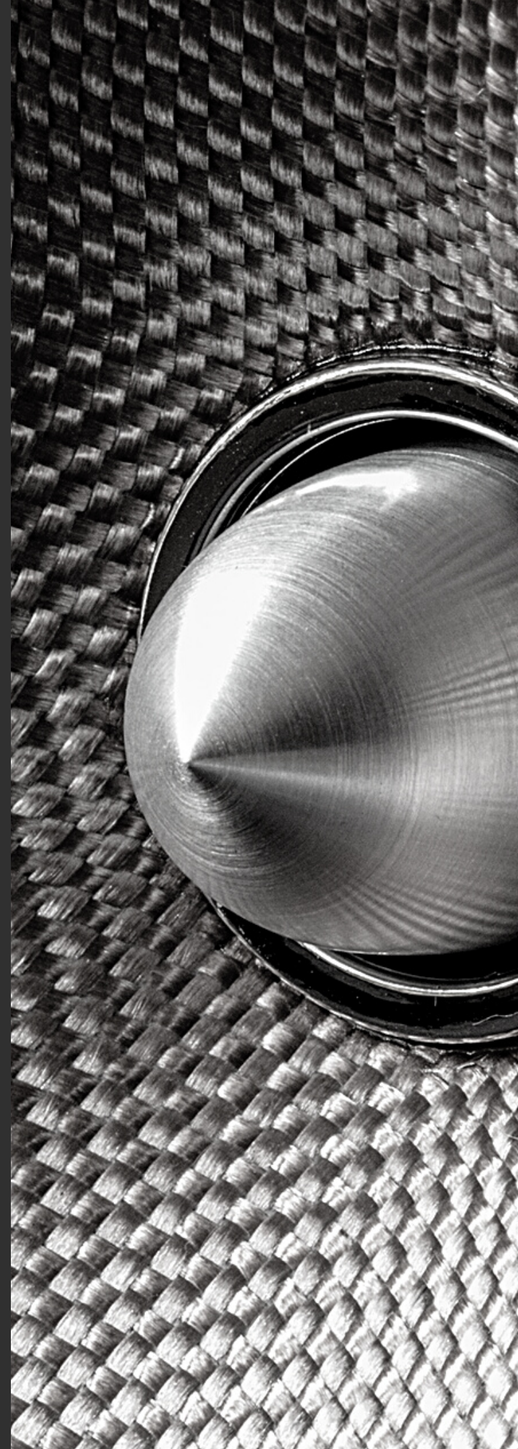


INTRODUZIONE

Lo sviluppo del seguente progetto si basa sulla creazione di un database a grafo riguardante i brani e i protagonisti che hanno partecipato alle ultime dieci edizioni del Festival di Sanremo, dal 2013 al 2022.

PROCESSO DI SVILUPPO

1. OTTENIMENTO DATI
2. ARRICCHIMENTO DATI
3. ANALISI ESPLORATIVE
4. MODELLAZIONE DEL GRAFO
5. DATA QUALITY



1.OTTENIMENTO DATI

Tramite l'utilizzo di API e di tecniche di Scraping sono stati ottenuti i dati.

API

Tramite R sono stati ottenuti tre dataset, riguardanti:
1. i brani
2. gli artisti
3. i generi musicali

Tramite Python il dataset viene arricchito aggiungendo il testo delle canzoni.

SCRAPING

Tramite Python sono state effettuate delle query Google per ottenere il periodo di attività.

Tramite la libreria 'wikipedia' è stato possibile ottenere:
1. i membri dei gruppi
2. la posizione in classifica

Tramite l'utilizzo di espressioni regolari’ è stato possibile ottenere i nomi degli autori

2.ARRICCHIMENTO DATI

Le tabelle ottenute tramite R vengono arricchite con le diverse sorgenti, in particolare:

- Il testo dei brani ottenuto con Genius alle canzoni
- L'anno di inizio carriera ottenuto tramite query Google agli artisti
- Posizione in classifica dei brani tramite Wikipedia
- La lista degli autori ottenuta attraverso recensiamomusica.com

Per effettuare l'arricchimento vengono eseguiti degli step di pre-processing, in quanto i brani vengono definiti diversamente nelle diverse fonti, tramite costrutti come "festival di sanremo" o "featuring"

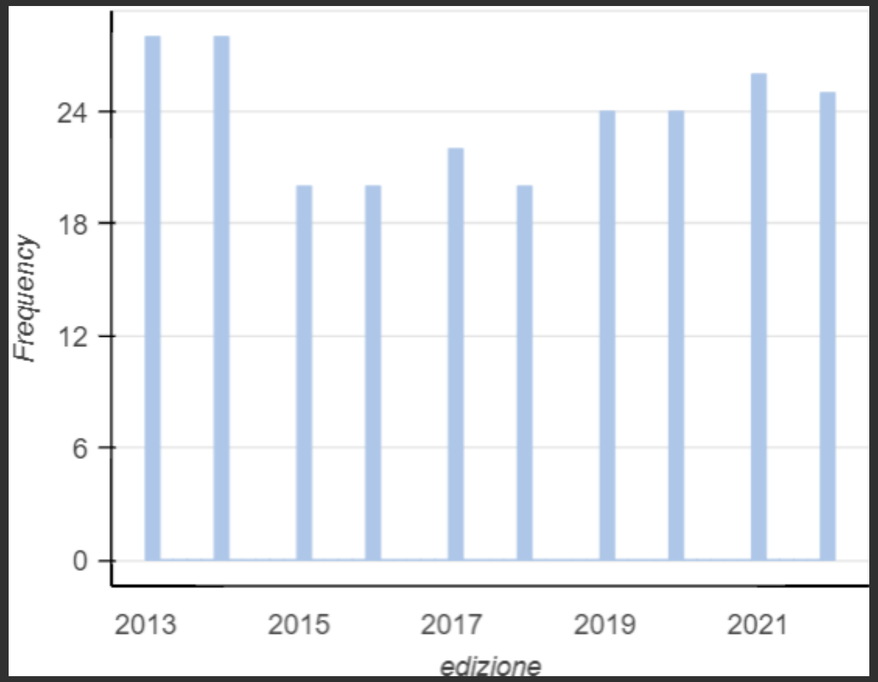
Alla fine di raccolta e arricchimento si ottengono diversi dataset per modellare le diverse entità, cioè Brani, Artisti, Generi, Autori, Membri dei gruppi, Edizione.

3.ANALISI ESPLORATIVE

BRANI

In seguito alle operazioni di arricchimento è stato ottenuto un dataset di 237 righe e 16 colonne (posizione, titolo, edizione, esplicita, id_track, ballabilità, energia, chiave, rumorosità, tonalità, parlato, positività, tempo, durata, ritmo, testo).
Delle 237 righe che compongono il dataset, 11 di queste presentano valori nulli.

Numero di canzoni per edizione

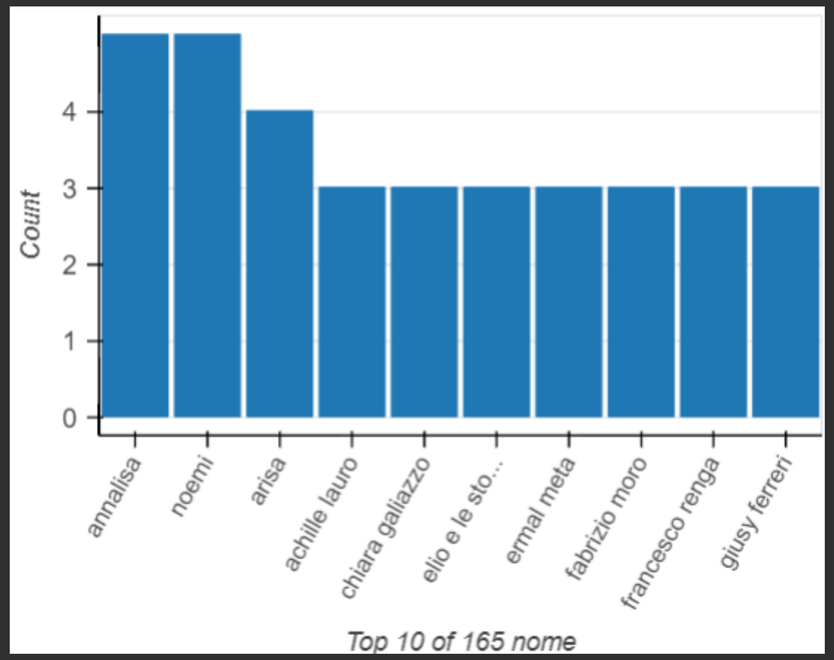


3.ANALISI ESPLORATIVE

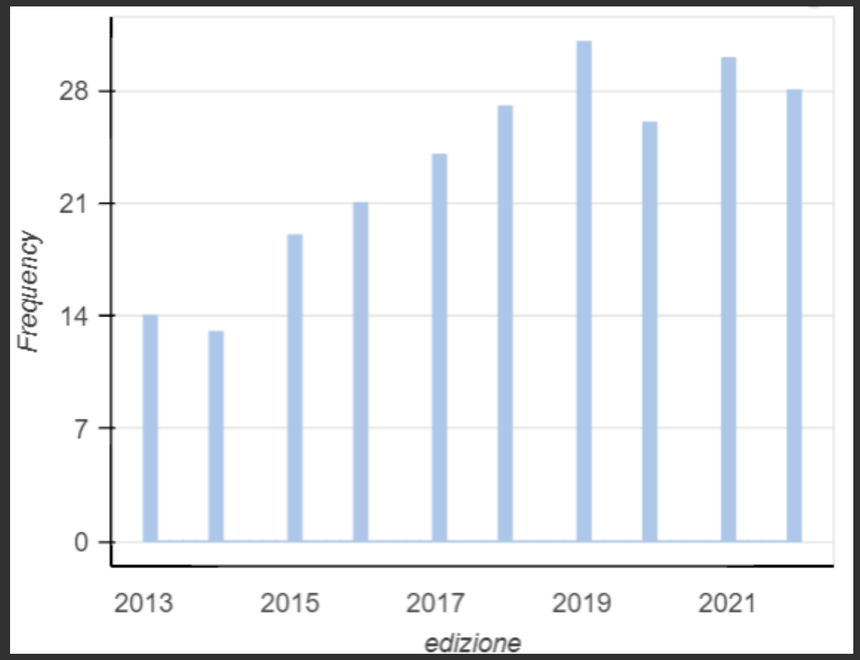
ARTISTI

In seguito all'arricchimento è stato ottenuto un dataset di 233 righe e 4 colonne (id_artist, nome, edizione, periodo attività).
Delle 233 righe che compongono il dataset, 7 di queste presentano valori nulli.

Artisti con più partecipazioni



Numero di artistiper edizione

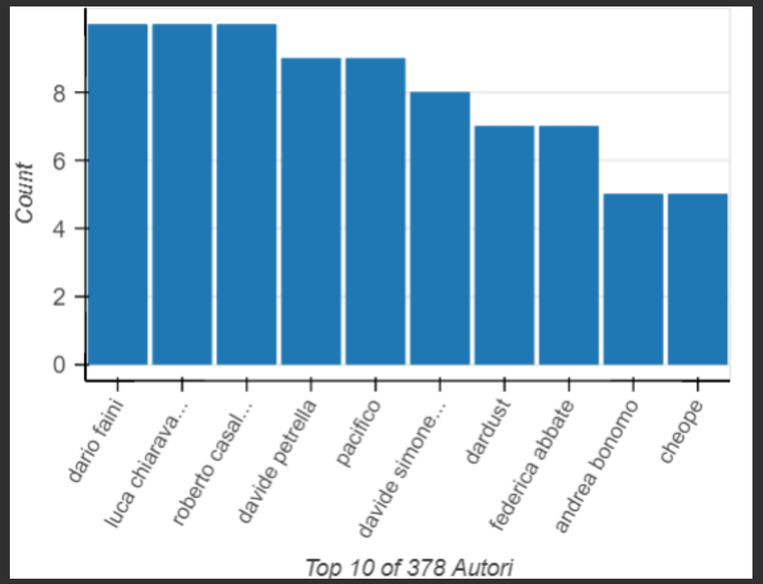


3.ANALISI ESPLORATIVE

AUTORI

Il dataset riguardante gli autori si compone di 378 righe e 2 colonne, cioè l'ID della traccia e il nome degli autori che l'hanno composta

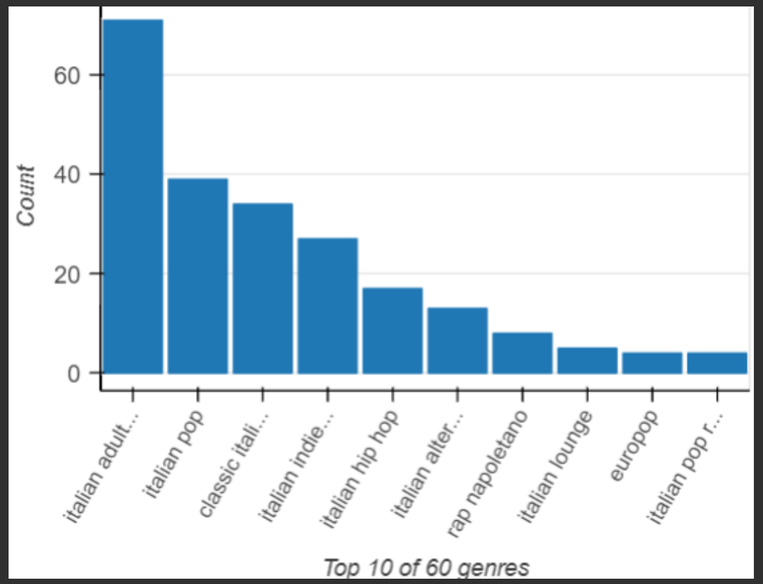
Autori che hanno scritto più brani



GENERI

Il dataset riguardante i generi si compone di 288 righe e 2 colonne, cioè l'ID dell'artista e il genere a cui viene associato

Generi pù popolari



4.MODELLAZIONE DEL GRAFO

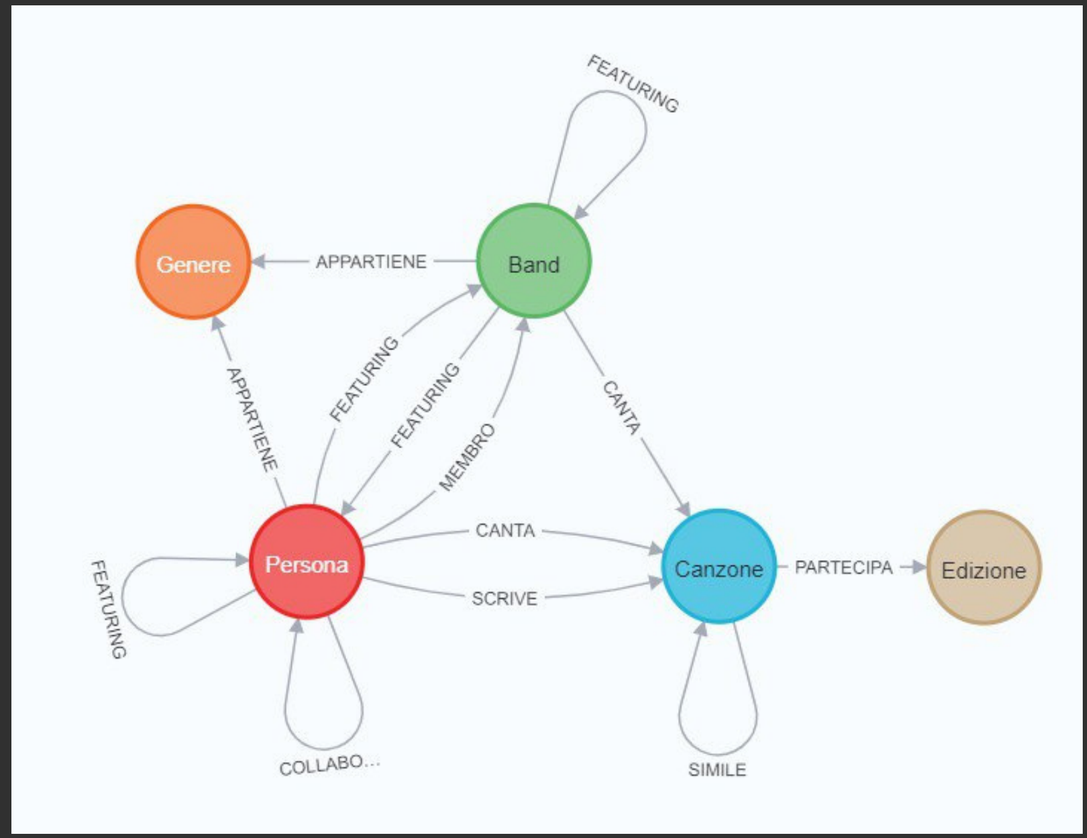
Dai dataset ottenuti in precedenza, tramite l'utilizzo di Neo4j è stato possibile creare un grafo composto da 806 nodi e 3176 relazioni.

NODI

- CANZONE: 237
- EDIZIONE: 10
- GENERE: 60
- PERSONA: 473
- BAND: 26

RELAZIONI

PARTECIPA: 237
APPARTIENE: 288
CANTA: 259
SCRIVE: 632
FEATURING: 35
MEMBRO: 104
COLLABORA: 734
SIMILE: 887



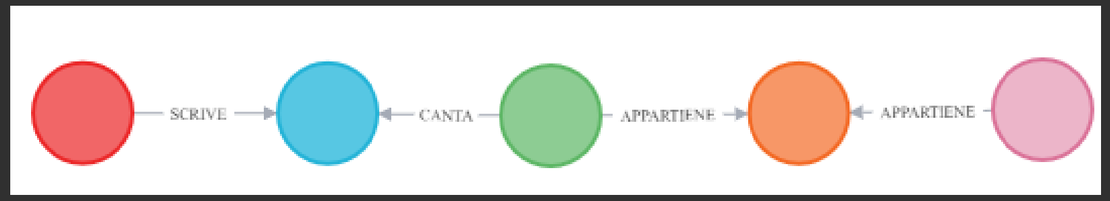
ESEMPIO 1

```
match (a:Artista)-[c:COLLABORA]-(b:Persona)
with a.nome as artista,
collect(b.nome) as collaboratori,
collect(distinct(c.anno)) as anni,
count(distinct(c.anno)) as num_anni,
count(distinct(b)) as num_persone,
count(c) as num_collaborazioni,
(count(distinct(b))*1.0/count(c)) as rapp_pers_coll
where count(c) > 3
return artista, collaboratori, anni, num_anni, num_persone,
num_collaborazioni,rapp_pers_coll
order by num_anni desc, rapp_pers_coll desc, num_persone desc
```

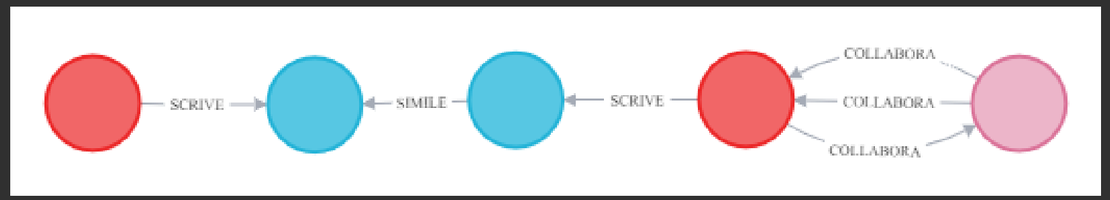
"artista"	"collaboratori"	"anni"	"num_anni"	"num_persone"	"num_collaborazioni"	"rapp_pers_coll"
"pacifico"	["marco buccelli","malika ayane","cesare chiodo","francesco gabbani","davide tagliapietra","roby facchinetti","niccolò contessa","bungaro","enzo avitabile","rocco rampino","andrea fresa","alessandra flora","alessandra flora","giovanni pallotti","malika ayane","giovanni truppi","gianna nannini","elisabetta sgarbi","giovanni caccamo","mirco mariani"]	["2022","2021","2018","2020","2013","2015"]	6	18	20	0.9
"mahmood"	["charlie charles","dardust","fedez","michelangelo","dargen d'amico","alessandro la cava","francesca michielin","dardust","alessandro raina","dario faini","davide simonetta","blanco"]	["2019","2021","2022","2020"]	4	11	12	0.9166666666666666

ESEMPIO 2

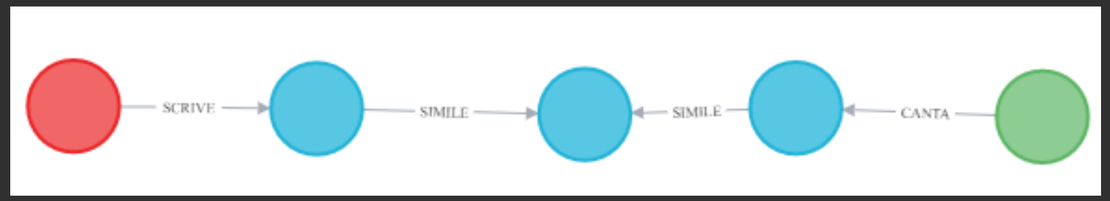
```
1  MATCH p=shortestPath(
2  (elio:Persona {nome:"stefano belisari"})-[*..5]-(:Persona
   {nome:""})) RETURN p
```



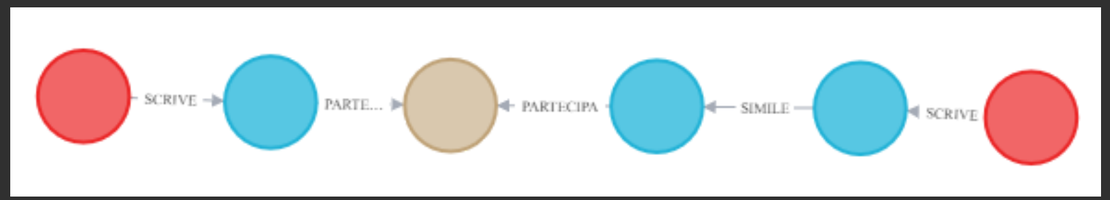
IVA ZANICCHI - ARTISTA



ACHILLE LAURO - ARTISTA



MANESKIN - BAND



STEFANO MAIUOLO - AUTORE

5.DATA QUALITY

COERENZA

Questa misura di qualità è stata necessaria appena conclusa la fase di ottenimento dati, per rendere coerenti tutte le informazioni riferite alla stessa persona ed evitare che ci riferisse alla stessa persona in modi diversi.

METODO AUTOMATICO

Record Linkage tra le colonne 'nome'
dei dataset artisti, autori e membri

ESEMPIO

Dargen D'Amico - Dargen D’amico
Edwyn Roberts - Edwin Roberts
Lodo Guenzi - Lodovico Guenzi

METODO MANUALE

Ricerca manuale

ESEMPIO

Annalisa - Annalisa Scarrone
Elio - Stefano Belisari

5.DATA QUALITY

COMPLETEZZA

Questa misura di qualità è stata utilizzata per osservare quanto le tecniche di download dei dati abbiano prodotto dei risultati esaustivi.

BRANI

Table completeness:

140 celle nulle -> 3.70%

Attribute completeness:

11 righe nulle -> 4.64%

10 testi nulli

ARTISTI

Table completeness:

7 celle nulle -> 0.8%

AUTORI & GENERI

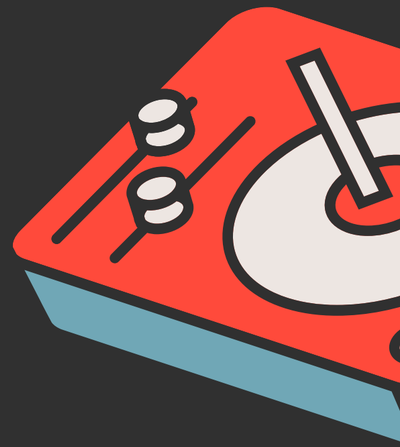
Nessun valore nullo

SVILUPPI FUTURI

Ampliare il grafo aggiungendo ulteriori edizioni precedenti, con la maggiore difficoltà di trovare i brani in Spotify.

Utilizzare altre tecniche più complesse (es. Word Embedding) di NLP per il calcolo della similarità tra testi.

Utilizzare metriche di analisi dei network (es. Centrality, Community Detection) per studiare i principali attori del Festival.



GRAZIE PER L'ATTENZIONE