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

Andrea Maver 828725, Simone Radaelli 845065, Oscar Zanotti 873763

9 settembre 2022

1 Introduzione

Lo sviluppo del seguente progetto si basa sulla creazione di un database riguardante i brani e gli artisti che hanno partecipato alle ultime dieci edizioni del Festival di Sanremo, dal 2013 al 2022. Il Festival di Sanremo è un importante evento musicale nato nel 1951, che da oltre settant'anni cattura l'attenzione di tutta la nazione e monopolizza l'interesse dei media durante il suo svolgimento. Sul famoso palco del teatro Ariston si sono esibiti artisti che hanno scritto pagine della storia della musica italiana, perciò si è deciso di prendere in considerazione questa manifestazione.

Per creare il database sono state utilizzate diverse fonti di dati, che verranno spiegate adeguatamente nelle sezioni successive, sia tramite l'uso di API che di tecniche di scraping. La scelta di utilizzare un modello a grafo risiede nel fatto che questo, meglio di ogni altro, permette di modellare le relazioni che intercorrono tra le varie istanze; essendo l'ambito di lavoro scelto riguardo ad autori, interpreti e brani, si è pensato che questo fosse il più adatto. L'utilizzo di grafi "musicali" è una pratica ben conosciuta e utilizzata, infatti molti sistemi di raccomandazione si basano proprio su modelli di questo tipo, che permettono di riconoscere brani con caratteristiche simili a quelli ascoltati dall'utente per potergliene suggerire di nuovi. Quindi, un uso del seguente grafo potrebbe essere proprio quello di raccomandazione, o una ricerca di relazioni e piccole comunità all'interno del gruppo di cantanti e autori che hanno partecipato al Festival.

Il report si costituisce di una sezione in cui viene spiegato il processo logico e tutte le fasi che hanno portato all'ottenimento dei dati e alla creazione del modello; successivamente si andrà più nello specifico per quanto riguarda lo scaricamento dei dati e tutte le sorgenti utilizzate, oltre alle fasi legate all'unione coerente delle diverse fonti; dopo una parte di analisi esplorativa dei dati si passa alla modellazione effettiva del grafo, nella quale verranno spiegate le sue relazioni e la struttura; nella parte conclusiva si valuteranno aspetti positivi e criticità riscontrate, oltre che eventuali sviluppi futuri.

2 Processo di sviluppo

Dopo aver deciso di voler sviluppare un progetto in ambito musicale, la prima fase dello svolgimento riguarda il capire da dove sia possibile ottenere i dati. La prima fonte che è stata presa in considerazione è Spotify, che mette a disposizione delle comode API. Spotify è una piattaforma di streaming musicale utilizzata in tutto il mondo e leader del mercato; vanta diversi milioni di abbonati al proprio servizio a pagamento solo in Italia, che diventano oltre un centinaio in tutto il mondo. A causa della sua grandissima diffusione è stata ritenuta una fonte attendibile per l'ottenimento di dati riguardanti brani e artisti. Inoltre, come detto, la piattaforma mette a disposizione delle API che permettono di scaricare facilmente diversi parametri legati alle canzoni. Dopo una prima fase preliminare di conoscenza delle informazioni disponibili, si è iniziato a pensare a che tipo di dati poter integrare a quelli di Spotify; dato che questa fornisce per lo più informazioni legate ai brani, si è pensato di svilup-

pare la parte dei cantanti e più in generale quella di tutte le persone coinvolte nel Festival. Per questo è stato scelto di utilizzare Wikipedia, l'enciclopedia digitale più famosa al mondo, facendo scraping di diverse tabelle e testi all'interno delle pagine riguardanti la competizione; inoltre, si è utilizzata la ricerca in Google per ottenere ulteriori informazioni riguardo ai cantanti e ai gruppi in gara.

A questo punto, con questi dati in mente, si è pensato a come poterli modellare. Dopo una fase di valutazione delle diverse alternative è stato scelto di utilizzare il modello a grafo, che permette di mostrare sia le diverse caratteristiche dei brani sotto forma di proprietà, ma anche le relazioni che assumono i cantanti uno con l'altro: in particolare, diversi cantanti hanno partecipato a più di un'edizione, a volte magari formando un duetto con un altro artista. Inoltre, si è pensato agli autori dei brani in gara, che spesso non sono conosciuti dal grande pubblico; infatti, mentre l'interprete è spesso un singolo individuo o una band, gli autori sono più di uno e si è pensato che molti di questi si sarebbero potuti trovare nei crediti di diversi brani. Questo permetterebbe, almeno a livello teorico, di infittire notevolmente il grafo, andando ad aggiungere molte relazioni e molti nodi, creando cricche e percorsi di attraversamento interessanti.

Dopo aver ottenuto tutti i dati dalle diverse fonti e averli integrati uno con l'altro in un modo fruibile per le fasi successive, si è modellato il grafo e si sono eseguite delle analisi di qualità.

3 Ottenimento dei dati

In questa sezione verranno esposte tutte le tecniche utilizzate per ottenere i dati necessari alla creazione del grafo; non potendo trovare online dei dataset fruibili, è stato necessario scaricare la totalità delle informazioni, utilizzando API e scraping.

3.1 API

Le API (Application Programming Interface) sono delle interfacce che permettono a delle applicazioni di comunicare comodamente l'una con l'altra. Questo permette di accedere ai database di diversi siti in modo sicuro, avendo la certezza della bontà della fonte. Sono una tecnica di ottenimento dati molto robusta ma che, per grandi moli, potrebbe a volte richiedere degli accordi particolari con i proprietari del dato. Per questo progetti sono state utilizzate due piattaforme, cioè Spotify per ottenere informazioni sui brani e Genius per quanto riguarda i testi.

3.1.1 Spotify

Per la prima fase della raccolta dei dati ci si è focalizzati sull'utilizzo delle API di Spotify, per capire quali e quanti tipi di dati si sarebbe riusciti a ottenere. Per usufruire comodamente di tutti i comandi forniti, si è deciso di utilizzare un pacchetto R chiamato "spotifyr", che facilita l'uso rispetto alle classiche interrogazioni dal browser. Il pacchetto contiene tutte le funzionalità messe a disposizione dalla piattaforma, oltre che una ricca documentazione, e permette di ottenere dati in un formato tabellare, al posto del JSON utilizzato dalle API classiche.

Il primo importante step è quello di creare le playlist contenenti le canzoni in gara; nonostante esistano sulla piattaforma alcune compilation create da utenti sconosciuti si è deciso di crearle appositamente, in modo tale da essere certi della correttezza dei brani inseriti. Andando indietro negli anni, si nota che alcune canzoni non sono presenti all'interno del catalogo di Spotify; purtroppo per queste non esiste alcuna soluzione di integrazione, in quanto i parametri numerici che verranno descritti successivamente sono calcolati attraverso algoritmi specifici, e non sono reperibili altrove. Queste sono in particolare: "È colpa mia" e "Quando non parlo" (2013), "Da lontano", "Sing in the rain" e "Un abbraccio unico" (2014), "Vita d'inferno" e "Voce" (2015). Inoltre, "Domenica" (2022) porta degli strani problemi nella fase di download, per i quali tutti i valori numerici, che sarebbero serviti a descriverla, sono risultati uguali a zero.

A questo punto può iniziare lo scaricamento tramite le API e R: tramite alcune fun-

zioni e delle manipolazioni dei risultati ottenuti, si ottengono tre dataset, uno per descrivere i brani, uno per gli artisti e uno per i generi musicali. Quello dei brani contiene la lista di tutti i titoli, un codice identificativo unico e diversi parametri numerici che descrivono le canzoni. Quello degli artisti invece contiene un codice univoco, il nome dell'artista o del gruppo, e l'edizione in cui ha partecipato. È importante far notare che alcuni brani sono interpretati da più artisti, per cui per evitare duplicazioni viene anche creata una tabella pivot che collega i codici di canzone e artista. Tutti i parametri verranno descritti successivamente, nella parte di esplorazione dei dati, dopo l'integrazione con le altre fonti. Il dataset dei generi musicali contiene per ogni riga una lista di generi che vengono associati all'artista da parte di Spotify e l'ID identificativo del cantante. Si nota che il campo genere non viene assegnato alla singola canzone, ma piuttosto all'interprete; questo permetterà di creare nodi con cluster di artisti accomunati dallo stesso genere musicale, utile in un potenziale sistema di raccomandazione.

3.1.2 Genius

Per arricchire ulteriormente le informazioni riguardanti i brani, si è deciso di scaricarne i testi, con l'obiettivo di poter creare un ambiente che dia spazio a possibili sviluppi riguardanti la loro analisi, come per esempio valutazioni di similarità. Per tale scopo viene utilizzata l'API resa disponibile in modo gratuito dal sito Genius, specializzato nella distribuzione ed interpretazione dei testi musicali.

Poiché l'uso di queste API richiede di indicare sia il titolo che l'interprete del brano, non si è potuto usare solamente la tabella riguardante le tracce, in quanto non contiene informazioni relative agli artisti. Utilizzando la tabella pivot artisti, nella quale sono presenti sia gli id delle canzoni che quelli dei cantanti, sono stati effettuati due join: il primo con la tabella artisti, grazie al quale è possibile ottenere il nome dell'artista; il secondo con la tabella tracce che permette di ottenere il titolo della canzone. Rimuovendo i duplicati, e mantenendo così solamente un'osservazione per canzone, si è ottenuto un dataset di 227 righe per 4 colonne: 'id track', 'id artist', 'nome', 'titolo'. Il numero di righe corrisponde al numero delle canzoni per cui è presente il valore di 'id track'.

In quanto l'API di Genius permette lo scaricamento di 73 lyrics alla volta, è stato necessario suddividere il dataset appena ottenuto in quattro sezioni, che verranno poi unite nuovamente a processo concluso. Siccome all'interno di Spotify possono essere presenti più versioni della stessa canzone, che differiscono leggermente nel titolo con costrutti come *featuring* e *edit*, prima di effettuare le ricerche all'interno del database di Genius è stato necessario effettuare un'azione di pulizia da questi elementi di disturbo. Una volta scaricati i testi sono stati concatenati i quattro dataset contenenti il testo, riottenendo così il formato originale. Infine, effettuando un merge sull'id delle canzoni con la tabella tracce, questa viene arricchita dall'aggiunta della colonna 'testo'.

3.2 Scraping

Lo scraping è una tecnica di ottenimento dati basata sull'analisi delle pagine HTML che compongono il web. Andando a lavorare su di esse è possibile isolare parti di informazioni utili; anche se potenzialmente permette di ottenere qualsiasi informazione presente in una pagina, è una tecnica molto poco robusta a causa del fatto che le strutture sono sempre diverse e per ognuna è necessario applicare delle manipolazioni ad hoc. Questa fase viene eseguita per aggiungere informazioni sia ai brani che agli artisti. In questa fase, le tabelle ottenute dalle manipolazioni in R sono state utilizzate come base di partenza, a cui sono state integrate le nuove informazioni.

3.2.1 Query Google

La prima idea di integrazione è stata quella di aggiungere l'anno di inizio della carriera degli artisti partecipanti al Festival; questo può dare informazioni sulla longevità del loro percorso e conseguentemente, anche sulla loro età. La lunghezza della carriera potrebbe rivelarsi utile per analisi che prendono

in considerazione proprio l'età dei cantanti, che si può legare a un periodo storico in cui l'industria musicale era caratterizzata da particolari trend che sono magari andati a sfumare nel tempo.

Per aggiungere questo tipo di informazione si è pensato inizialmente di utilizzare delle altre API, cioè quelle di Wikidata, che permettono di cercare i cantanti tramite il loro codice identificativo di Spotify. All'interno del grafo di conoscenza poi si sarebbe cercata la proprietà "work period", che indica l'anno di inizio dell'attività. Tuttavia, dopo dei test preliminari si è notato che questo metodo permetteva di assegnare l'anno solamente a circa metà di tutti gli interpreti in gara; questo perché alcuni di loro nella loro pagina di Wikidata non contengono l'ID Spotify, altri perché non hanno la proprietà "work period". Si è deciso per questo motivo di utilizzare un'altra tecnica, cioè lo scraping.

Per farlo si è usufruito di alcune librerie Python, cioè Requests e BeautifulSoup: queste permettono di porre una query a un motore di ricerca per poi ottenere la pagina in formato HTML, nella quale è possibile navigare per isolare solo le informazioni utili al caso. Dopo aver creato una lista di nomi di cantanti e gruppi, si sono eseguite delle query in questo formato: "periodo di attività musicale + *nome artista*". Tramite delle espressioni regolari si è poi estratta solamente la parte contenente la risposta alla domanda e dopo dell'ulteriore pulizia grazie ad altre espressioni regolari si è riusciti ad assegnare correttamente l'anno di inizio carriera ai vari artisti e gruppi. Si nota che per sette righe non avviene l'assegnazione, quindi presenteranno valori nulli; queste problematiche verranno discusse nella parte di qualità.

3.2.2 Wikipedia

Membri band

Un'ulteriore idea di integrazione riguarda la creazione di un dataset contenente i membri delle band che hanno partecipato al Festival. Questo può risultare interessante per vedere se sono presenti artisti che prima facevano parte di un gruppo e successivamente hanno partecipato alla competizione da solisti; una potenziale analisi potrebbe riguardare il confronto delle tracce prima e dopo questo cambiamento.

Come primo passaggio è necessario dividere la tabella riguardante gli artisti, separando i cantanti solisti dai gruppi; nascono così due tabelle con la stessa struttura. In particolare, 'solo cantanti' è composta da 199 righe, mentre 'solo gruppi' da 34, che se unite permettono di ottenere nuovamente la tabella originale. Questa divisione viene effettuata sfruttando delle conoscenze di dominio, anche grazie al fatto che le band non sono numerose, e sarà importante anche per le successive fasi di modellazione del grafo, permettendo di creare dei nodi di tipo diverso.

Per ottenere le informazioni si eseguono delle interrogazioni alla pagina Wikipedia basate sul nome del gruppo tramite la libreria 'wikipedia' di Python. Nonostante sia un'enciclopedia libera compilata dagli utenti, è talmente di uso comune che è ragionevole pensare che le informazioni contenute siano corrette. All'interno della pagina HTML risultante poi vengono cercati e isolati dei costrutti come 'Formazione' o 'composto da', che contengono blocchi di testo riguardanti i componenti dei gruppi. In questo modo si può associare a ogni band il proprio paragrafo, che però necessita ancora di molte manipolazioni per poter essere portato a un formato utilizzabile. È importante notare che a causa della diversa struttura delle pagine, per alcune band è necessario operare in modo personalizzato, perché l'approccio generale non avrebbe portato i risultati sperati. Per ogni gruppo viene effettuato uno split e un'esplosione delle righe in modo tale da ottenere su ogni riga una singola persona. Dopo un'ulteriore pulizia da costrutti come date e alias si ottiene un dataset di 104 righe e 3 colonne, cioè gruppo, nome e strumento; quest'ultima fornisce informazioni sul ruolo della persona all'interno del gruppo, che sia la voce, chitarra o altro.

Brani

Nelle fasi successive verrà arricchita la tabella riguardante i brani. Come prima infor-

mazione aggiuntiva si è pensato di aggiungere la posizione di arrivo nella classifica finale della canzone, per poter eventualmente analizzare se esistano dei pattern o delle caratteristiche tipiche delle migliori classificate o al contrario di quelle con meno successo. Per aggiungere il dato si è deciso di utilizzare lo scraping sulla pagina ufficiale del Festival di Sanremo su Wikipedia per ogni annata. Utilizzando le stesse librerie delle query Google si è ottenuta una tabella contenente la posizione di arrivo, il titolo, l'interprete e gli autori del brano. Questo è stato effettuato per tutte e dieci le edizioni, stando attenti ad alcune differenze di struttura presenti nelle pagine.

A questo punto, osservando i dati appena ottenuti, si è pensato di lavorare sull'attributo autori. Questo presentava sulla stessa riga una lista di nomi in formato "iniziale nome. cognome" (M. Mengoni) e da una veloce esplorazione si è notato che spesso anche l'interprete della canzone risultava nella lista degli autori, ma con formato diverso, cioè "nome cognome" (Marco Mengoni). Per questi motivi si è deciso di cercare un'altra fonte online da cui ottenere la lista di tutti gli autori che hanno contribuito alle ultime dieci edizioni del Festival, nello stesso formato della colonna autori prodotta da Spotify. In questo modo sarebbe risultato immediato per il sistema riconoscere che due istanze con lo stesso nome si riferiscono alla stessa persona.

3.2.3 Recensiamomusica.com

Dopo una prima ricerca si è deciso di focalizzare l'attenzione su un blog di musica chiamato recensiamomusica.com che contiene diversi articoli riguardanti il Festival, tutti di un formato simile, con specificati anche gli autori delle canzoni descritti da nome e cognome per intero. Facendo un veloce controllo con la lista presente su Wikipedia si è potuto constatare la bontà della fonte, per procedere così allo scraping. A differenza di quella precedente, queste pagine HTML non presentano delle tabelle per rendere più comodo l'ottenimento dei dati, ma solo blocchi testuali. Con l'utilizzo però di una serie di espressioni regolari si è riusciti a isolare solo le parti necessarie, cioè la lista degli autori separati da virgola e il titolo del brano, necessario per poter poi effettuare il match con la tabella precedente. Il processo è stato ripetuto per tutte e dieci le edizioni, con dei leggeri cambiamenti necessari tra una e l'altra, in quanto il codice non era particolarmente pulito; questo argomento verrà approfondito nella parte sulla qualità dei dati. Alla fine dello scraping ripetuto, unendo i risultati per le singole edizioni si ottiene una grande tabella, che può essere unita a quella creata nel punto precedente, cioè quella di Spotify, con aggiunta della posizione in classifica.

3.3 Integrazione e arricchimento

Questa fase è necessaria per unire in un unico dataset tutti i dati ottenuti nelle precedenti fasi di download, in particolare l'unione della tabella riguardante i brani, con l'aggiunta della posizione in classifica, e quella riguardante gli autori. Il periodo di attività è già presente in modo corretto all'interno della tabella adatta, infatti il processo di unione è stato sviluppato parallelamente a quello di scaricamento.

Per effettuare un merge corretto e il più completo possibile si è proceduto a step: prima di tutto si è deciso di applicare una left join, considerando come tabella di sinistra quella proveniente da Wikipedia, e Spotify. Questo perché la tabella proveniente dall'enciclopedia online contiene tutte le canzoni in gara, anche quelle che invece mancano sulla piattaforma di streaming. In questi casi il match non può evidentemente accadere, e i brani in questione presenteranno dei valori nulli per i parametri numerici che dovrebbero descriverli.

L'unione viene effettuata tramite il titolo e l'edizione, per evitare che canzoni con lo stesso titolo creassero problemi. Dopo un primo tentativo di integrazione la tabella risultante presenta trentatré righe che non sono state accoppiate; osservando i titoli di queste, provenienti da Spotify, si nota che molti presentano delle espressioni come "- festival di sanremo 2014" e "- sanremo 2019", perciò si decide di eliminare questi costrutti tramite delle espressioni regolari.

Una nuova merge mostra che una decina di casi sono stati risolti. A questo punto i problemi più evidenti sono in titoli che contengono espressioni come "feat." o "with"; eliminandoli e provando ancora gli elementi non accoppiati sono diventati solo sedici.

Tramite ulteriori trasformazioni e unioni si ottiene un risultato decisamente buono, in cui i brani che presentano valori nulli sono solo undici: è importante ricordare che nove di questi, non essendo disponibili in streaming sulla piattaforma, non possono essere accoppiati in alcun modo. Per gli altri casi il merge non è potuto avvenire a causa di differenze nelle particolari nelle stringhe dei titoli. La tabella risultante è composta da titolo, id, edizione, posizione e i vari parametri numerici. A questo punto si può procedere con la manipolazione degli autori.

Per creare una tabella fruibile per gli usi futuri è necessario isolarli, insieme agli ID dei brani, per poi dividere la lista quando si presenta una virgola e duplicare il codice per il numero di persone che compaiono nel gruppo. Questo è facilmente realizzabile sfruttando le funzioni split e explode. Si nota però che sono presenti alcune eccezioni, cioè casi in cui, nell'insieme degli autori di un brano, risultano due fratelli o sorelle: in questo caso non sono separati da virgola ma hanno un formato del tipo "nome e nome cognome" (ad esempio Max e Francesco Gazzè). Queste occorrenze non separate in precedenza vanno trattate in altro modo, cioè creando degli split in presenza di spazi bianchi, ricomponendo nome e cognome e inserendo una virgola tra i due. Effettuando nuovamente le operazioni di split per virgola e explode si ottiene così una tabella pivot finale per gli autori, in cui a ogni brano vengono associate tutte le persone che hanno partecipato alla sua creazione.

Dopo queste fasi di ottenimento di dati si ottengono quattro tabelle che verranno descritte dettagliatamente nella prossima sezione. Per la successiva modellazione del grafo verranno applicate alcune ulteriori manipolazioni, per rendere più agile la trasformazione dal formato tabellare a quello a grafo. Queste modifiche verranno tutte esposte all'inizio della fase di modellazione.

4 Analisi esplorativa

In questa sezione vengono descritte le tabelle ottenute tramite le tecniche di API e scraping e integrate tra loro, prima della modellazione nel grafo. Sono dati in formato tabellare, di cui verranno esposti i vari attributi e mostrate alcune analisi esplorative.

4.1 Tracce

Il dataset tracce si compone di 237 righe e 16 colonne, cioè:

- posizione: posizione di arrivo del brano nella classifica finale; assume valori numerici o NF, che indica che la canzone non si è qualificata alla serata finale
- titolo: titolo del brano
- edizione: anno in cui la canzone è stata portata in gara
- esplicita: attributo binario che indica se la canzone contenga testi espliciti
- id_track: codice identificativo univoco del brano
- ballabilità: valore da 0 a 1 che indica quanto una canzone sia adatta a essere ballata, basandosi su tempo e ritmo
- energia: valore da 0 a 1 che indica la percentuale di intensità e attività. Solitamente, tracce energiche sembrano veloci, forti e rumorose
- chiave: chiave della canzone, espressa tramite una convenzione numerica
- rumorosità: rumorosità generale della traccia espressa in dB; assume valori compresi tra -60 e 0
- tonalità: tonalità della canzone, può essere maggiore o minore
- parlato: valore da 0 a 1 che individua la presenza di parole all'interno della traccia. Valori alti indicano tracce unicamente parlate, come podcast, mentre quelli più bassi si riferiscono a canzoni; valori medi sono spesso assunti da canzoni rap, caratterizzate dall'avere un gran numero di parole

- positività: valore da 0 a 1 che indica la positività trasmessa dalla canzone. Vicino a 1 è felice, gioiosa e euforica; al contrario, vicino a 0, risulta triste, depressa o arrabbiata
- tempo: misura della frequenza espressa in battiti al minuto, tipica per la descrizione di tracce musicali
- durata: durata della canzone espressa in minuti
- ritmo: valore che indica quanti battiti sono contenuti in ciascuna misura
- testo: stringa contenente il testo delle canzoni

Delle 237 righe che compongono il dataset, 11 di queste presentano valori nulli; un approfondimento a questo riguardo verrà effettuato nella sezione di data quality.

Per quanto riguarda le variabili numeriche utilizzate per descrivere i brani, si può notare che hanno tutte delle distribuzioni normali, in quanto media e mediana sono sempre indicativamente dello stesso valore. I valori di media, deviazione standard e quantili sono osservabili per intero nella tabella 3 in Appendice. Facendo una sorta di identikit della canzone media di Sanremo nelle ultime dieci edizioni, si potrebbe dire che questa è mediamente ballabile e piuttosto energica, senza un numero particolarmente elevato di parole e indicativamente più triste che felice, senza parole scurrili e tendenzialmente in tonalità maggiore. Osservando una matrice di correlazione di Pearson (figura 1) non si notano particolari correlazioni tra le variabili, a parte una positiva tra rumorosità ed energia, che raggiunge un valore di 0.68, e tra positività, ballabilità ed energia, con valori intorno a 0.5.

Infine, osservando la distribuzione dei brani per edizione (figura 2), si nota che ci sono stati diversi cambiamenti nel numero di tracce partecipanti; questo perché negli anni il format del Festival è cambiato: nel 2013 e 2014 ogni artista portava in gara due canzoni, delle quali una si qualificava alla serata finale e l'altra no. Questo porta ad avere un alto numero di tracce ma basso numero

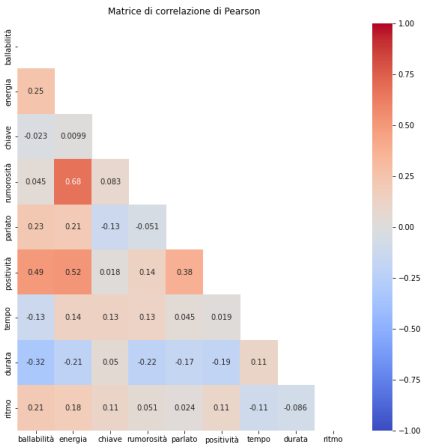

Figura 1: Correlazione di Pearson

di interpreti; quando nel 2015 la struttura è stata cambiata, il numero di canzoni partecipanti è calato di quasi un terzo, avendo poi una risalita con il passare degli anni.

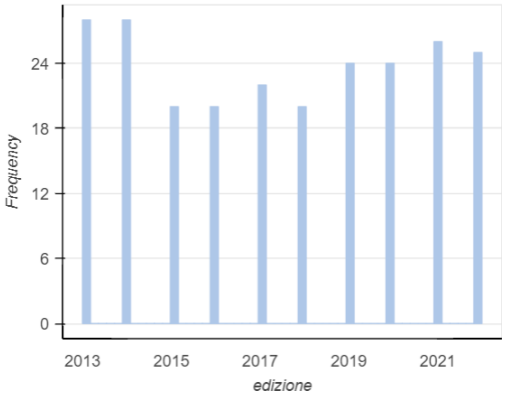

Figura 2: Numero di brani per edizione

4.2 Artisti

Il dataset riguardante gli artisti si compone di 233 righe e 4 colonne, cioè:

- id_artist: codice identificativo per ogni artista
- nome: nome dell'artista
- edizione: edizione in cui ha partecipato
- periodo_attivita: anno di inizio carriera

Nelle ultime dieci edizioni alcuni artisti hanno partecipato più volte: Noemi e Annalisa sono state le più popolari, con cinque partecipazioni a testa, seguite da Arisa a quattro; naturalmente la gran parte dei cantanti ha partecipato ad una sola edizione (figura 3). In particolare si possono osservare le distribuzioni delle partecipazioni in tabella 1.

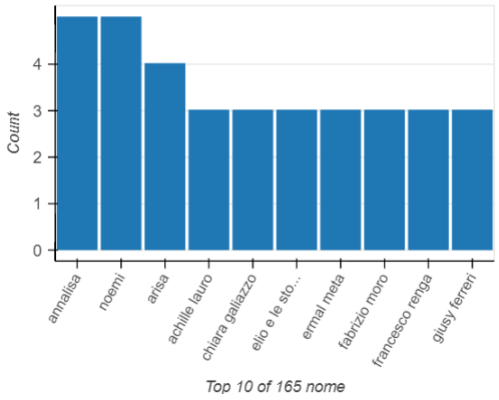

Figura 3: Numero di partecipazioni per artista

Partecipazioni	% artisti
1	71.51
2	18.78
3	7.87
4	0.60
5	1.21

Tabella 1: Distribuzioni percentuali delle partecipazioni

Come anticipato nell'analisi dei brani, i cambiamenti di format del Festival hanno generato anche distribuzioni diverse per quanto riguarda gli artisti. In particolare le prime edizioni presentavano meno interpreti, con un aumento dal 2015 in poi; in figura 4 si possono osservare queste informazioni.

Considerando invece il periodo di attività dei partecipanti, si nota come la distribuzione non sia propriamente simmetrica, in quanto media e mediana differiscono leggermente. Il valore mediano per l'anno di inizio carriera è il 2002; la tabella 4 con media e quantili è consultabile nell'Appendice. Osservando edizione per edizione, si scopre che il 2018 è stato l'anno in cui hanno partecipato artisti con le carriere più lunghe,

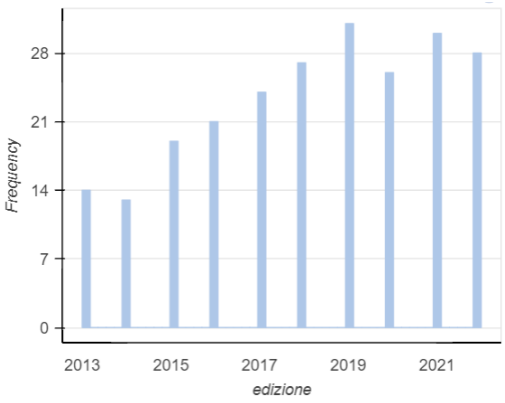

Figura 4: Numero di artisti per edizione

quindi, ragionevolmente, anche più anziani di età. Negli anni successivi si è riscontrato invece un generale "ringiovanimento", messo in atto forse per attirare anche il pubblico più giovane che sembrava aver perso interesse nella competizione. Tramite i violin plot in figura 5 è possibile osservare le distribuzioni degli anni di carriera per gli artisti nelle varie edizioni; la linea bianca indica l'andamento del valore medio con il passare del tempo

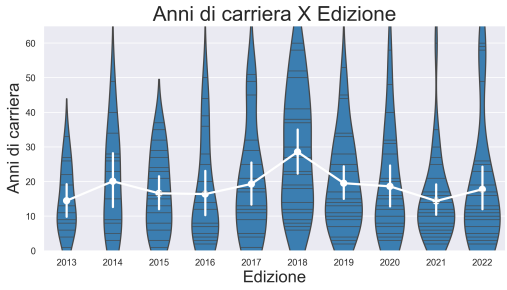

Figura 5: Numero di artisti per edizione

4.3 Autori

Il dataset riguardante gli autori si compone molto semplicemente di due colonne, ed è simile al pivot che lega tracce e cantanti. Sono presenti l'ID identificativo della traccia e il nome degli autori che l'anno composta. Data che ogni riga contiene solo un nome, nel caso in cui a una canzone abbiano collaborato più persone è necessario ripetere più volte il codice identificativo.

Analizzando gli autori analogamente agli artisti, si nota che la maggior parte di questi 378 ha scritto o contribuito alla scrittura di

una sola canzone, mentre il numero diminuisce al crescere di tracce composte. Quasi il 90% degli autori ha prodotto uno o due brani. Le percentuali precise sono osservabili in tabella 2.

Canzoni scritte	% autori
1	65.43
2	22.42
3	3.95
4	4.22
5	1.84
6	0
7	0.52
8	0.26
9	0.52
10	0.79

Tabella 2: Distribuzioni percentuali delle produzioni

In particolare, gli autori più popolari sono stati Roberto Casalino, Luca Chiaravalli e Dario Faini che hanno composto ben dieci canzoni, seguiti da Pacifico e Davide Petrella con nove a testa. La distribuzione è osservabile in figura 6. Invece il brano a cui hanno collaborato più persone è 'Combat pop' de Lo Stato Sociale, a cui hanno partecipato sette persone.

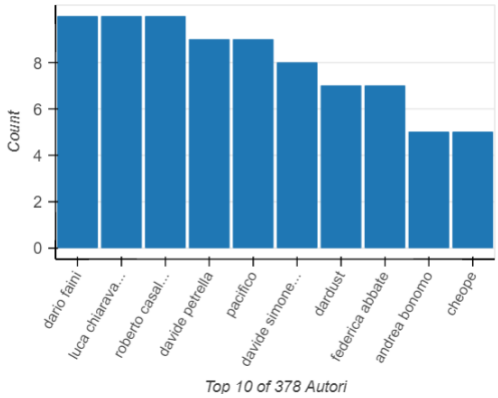

Figura 6: Numero di autori con più canzoni prodotte

4.4 Generi

Questo dataset ha una struttura molto semplice, in quanto è composta da 2 colonne e 288 righe. Ogni riga presenta il codice identificativo Spotify di un artista e il genere a cui la piattaforma lo assegna. È possibile che per ogni cantante siano considerati più generi, caso in cui, necessariamente, l'ID sarà ripetuto tutte le volte necessarie. Osservando le denominazioni dei generi si nota che molti di essi sembrano essere particolarmente simili, ma si è deciso di non applicare ulteriori operazioni, come raggruppamenti o aggregazioni, in quanto non avendo una controparte di verifica non se ne sarebbe potuta valutare la correttezza. Dei 60 generi diversi trovati, quello più comune è italian adult pop, a cui vengono assegnati 71 artisti; lo seguono italian pop e classic italian pop con 39 e 34 a testa (figura 7).

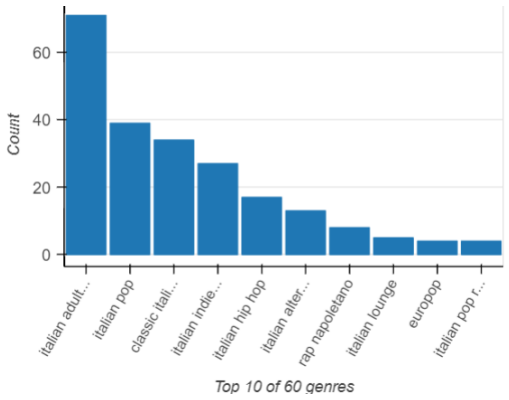

Figura 7: Numero di artisti per genere

L'artista a cui sono stati assegnati più generi, 6, è in realtà una band, cioè i The Bloody Beetroots, seguiti da Neffa e Dargen D'amico con 5.

5 Modellazione del grafo

In questa sezione verranno esposti i vari procedimenti che hanno portato alla modellazione del grafo, dettagliando tutte le operazioni preliminari e le scelte prese per la creazione del database.

5.1 Manipolazione tabelle

Pensando a come voler realizzare il grafo, si è realizzato che sarebbe stato necessario manipolare ulteriormente le tabelle ottenute nelle sezioni precedenti, per rendere più agevole l'inserimento nel database.

5.1.1 Solo Autori, Artisti Autori, Solo Artisti

Sono state effettuate delle trasformazioni per creare tre tabelle: solo autori, artisti autori e solo artisti. Queste tre permettono di modellare nodi con etichette diverse all'interno del grafo, in modo tale da ottenere un risultato più preciso ed esauriente.

Per prima cosa è stato effettuato un right merge sulle colonne nome e Autori tra le tabelle 'solo cantanti' e 'pivot autori'; è stato scelto di eseguire un right merge in maniera tale da mantenere i valori nulli per alcune colonne, che sono state successivamente utilizzate per separare le diverse categorie. Nello specifico, se la riga presenta valori nulli per la colonna 'id artist' significa che la persona in questione è solo un autore, in caso contrario viene considerata come artista autore; in questa maniera è stato possibile eseguire una prima distinzione. Per semplificare il processo i valori nulli sono stati trasformati in 0. Selezionando le righe questo valore per la colonna 'id artist' viene creata la tabella 'solo autori'. Questa è composta da 442 righe e 2 colonne (id track, Autori); essendo una tabella pivot è possibile trovare sia autori che canzoni duplicate, in quanto una canzone può essere scritta da più autori e un autore può aver scritto più canzoni. Eseguendo delle group by sulle colonne è possibile vedere che la tabella contiene 272 individui che hanno contribuito a scrivere 188 tracce. In questo dataframe sono contenute solo persone che hanno esclusivamente partecipato alla scrittura e composizione dei brani, ma senza mai salire sul palco a interpretarli.

In maniera equivalente, selezionando le righe che con valori diversi da '0' per la colonna 'id artist', viene creata la tabella 'artisti autori'. Da questa viene eliminata la colonna riguardante l'edizione per essere in grado di effettuare una rimozione dei duplicati, provenienti dall'operazione di merge precedentemente effettuata, in quanto alcuni artisti hanno partecipato a più edizioni; questa rimozione è stata necessaria in quanto altrimenti si sarebbero creati problemi di ridondanza per la creazione della relazione SCRIVE all'interno del grafo. La tabella finale è composta da 190 righe per 4 colonne (id artist, nome, periodo attivita, id track). Anche in questo caso sono state effettuate delle group by sulle colonne 'id artist' e 'id track' individuando così 106 artisti autori che hanno contribuito a scrivere 158 canzoni; le persone in questione hanno sia partecipato alla scrittura che interpretato le canzoni durante la manifestazione.

Considerando la tabella generata dal merge inziale, il numero totale di autori univoci risultava essere di 378, dei quali 272 solo autori mentre 106 sia artisti che autori.

Infine, effettuando una nuova fusione su 'id artist' tra la tabella 'artisti autori', raggruppata per 'id artist' in maniera da mantenere per ogni artista autore una sola riga, e la tabella 'solo cantanti', è stato possibile individuare i rimanenti cantanti che non hanno contribuito alla fase di scrittura della canzone. La tabella finale 'solo artisti' è composta da 33 righe per 3 colonne (id artist, nome, periodo attivita). Ogni id artist è univoco e non ripetuto, in quanto non essendo presenti informazioni riguardanti i brani non è possibile che esistano valori duplicati.

Effettuando queste trasformazioni e con la creazione di queste tabelle è stato possibile assegnare ai nodi del grafo etichette diverse o combinazioni di esse.

5.1.2 Generi

Partendo dalla tabella generi, ottenuta tramite API Spotify, sono state effettuate delle trasformazioni per renderla utilizzabile per l'inserimento della relazione APPARTIENE nel grafo. La tabella iniziale è composta da 231 righe, tante quante gli artisti in gara, e 2 colonne: 'id artist', codice univoco per identificare i partecipanti e 'genres', una lista contenente i vari generi associati ad ogni cantante.

Il primo step è stata la rimozione dei duplicati, in quanto un'artista appariva tante volte quante le sue partecipazioni. In seguito, sono state rimosse le righe contenti valori nulli per la colonna del genere. Dopo questi primi passaggi il numero di righe è sceso a 135, corrispondente agli artisti per i quali è associato almeno un genere, considerati una volta sola. In quanto Neo4j non supporta le liste, la colonna è stata esplosa, duplican-

do 'id artist' per il numero di generi associato, creando così una tabella pivot 'artisti generi' composta da 288 righe, utilizzata per creare la relazione APPARTIENE. Eseguendo un raggruppamento sulla colonna 'genres' si nota la presenza di 60 generi univoci.

Isolando solamente i generi univoci è stata, inoltre, creata una tabella 'nodi generi' composta appunto da 60 righe.

5.1.3 Collaborazioni

Per poter infittire il grafo è stato deciso di creare la relazione COLLABORA che unisce gli autori che hanno collaborato per la scrittura di una canzone, potendo vedere così le varie collaborazioni che si sono create nel corso degli anni.

Per fare questo si è partiti dal dataset riguardante gli autori, nel quale per ogni canzone sono presenti le persone che hanno contributo a scriverla. Per prima cosa viene creata la colonna 'collaboratore' come copia della colonna già presente 'Autori'; successivamente viene creata una nuova tabella raggruppando per id track e aggregando le liste di autori e collaboratori. Queste due colonne identiche vengono poi esplose, in maniera tale da andare a rimuovere quelle righe che presentano lo stesso valore per le due colonne, in quanto una persona non può collaborare con sé stessa. Nella fase successiva sono state eliminate le ridondanze che si sono create, situazioni nelle quali, in righe diverse, una persona A collabora con una persona B e viceversa; di tutte queste coppie è stata mantenuta solo un'occorrenza. La tabella finale collaborazioni ha una struttura pivot ed è composta da 797 righe per 3 colonne (id track, Autori, collaboratore). Per arricchire quella che sarà la relazione COLLABORA, viene eseguito un merge su 'id track' con la tabella canzoni, selezionando per questa solo le colonne 'titolo' ed 'edizione', permettendo così di aggiungere queste due proprietà alla relazione. Questa aggiunta fornisce informazioni nei casi in cui due individui abbiano collaborato più volte, in quanto permette di osservare per quali brani e in che anni è avvenuta la collaborazione.

5.1.4 Featuring

In maniera analoga a quanto descritto per la relazione COLLABORA, è stato deciso di creare la relazione FEATURING. Questa mette in relazione gli artisti che hanno duettato durante la competizione. Per far ciò sono stati eseguiti gli stessi passaggi descritti in precedenza utilizzando la tabella 'pivot artisti', escludendo quelle canzoni che sono state cantante da un artista solista o da band. In questo modo è stata creata la tabella 'featuring' composta da 35 righe.

5.1.5 Similarità tra canzoni

Per arricchire ulteriormente il database è stato deciso di creare una relazione SIMILE che unisce i vari nodi canzone. Si è deciso di andare a calcolare la cosine similarity tra i testi delle canzoni, che potrebbe risultare utile nella ricerca di brani simili non solo a livello musicale, ma anche a livello di testo, cercando anche se esitano corrispondenze o pattern all'interno di generi o edizioni specifiche. Si potrebbe inoltre analizzare se diversi autori abbiano stili di scrittura che caratterizzano tutti i brani a cui hanno partecipato, e osservarne le differenze.

Partendo dalla tabella dei brani che comprende i testi delle canzoni, sono state rimosse le righe con valori nulli per in questo campo. Per poter calcolare la similarità dei testi è stato necessario effettuare delle operazioni di pre-processing delle stringhe contenenti le lyrics, nello specifico:

- conversione del testo in minuscolo;
- il testo scaricato aveva un formato del tipo: titolo, lyrics, testo; quindi, è stato eseguito uno split sulla stringa utilizzando la parola lyrics ed è stata mantenuta solamente la parte della stringa contenente il testo vero e proprio;
- trasformazione del carattere '/n', che corrisponde a testo a capo, con uno spazio;
- rimozione del termine 'embed', della punteggiatura, dei numeri, delle stopwords (ottenute tramite la libreria Nl-

tk), delle single digits, di spazi bianchi multipli e degli accenti;

- stemming per avere valori più alti di cosine similarity, a causa di una matrice doc-term meno sparsa rispetto ad altre tecniche come lemmatization

Successivamente è stato creato un grande corpus contenente tutti i testi delle canzoni. Tramite la funzione TfidfVectorizer presente nella libreria sklearn, è possibile rappresentare il testo in uno spazio vettoriale andando a creare una matrice tf x idf grazie alla quale sarà possibile calcolare la similarità, in questo caso la cosine, presente anch'essa in sklearn. Viene creata una funzione 'get_recommendations' tramite la quale per ogni canzone è possibile estrarre la top 5 delle canzoni più simili. Osservando i valori si è notato che non erano particolarmente alti, con un valore massimo intorno a 0.45; per questo motivo si è deciso di non considerare un numero troppo elevato di relazioni, in quanto alcune sarebbero potute risultare poco significative. Inoltre è stato scelto di non filtrare i risultati ponendo un valore soglia, facendo sì che tutte le canzoni avessero dei collegamenti. Infine, come per la creazione delle tabelle 'collaborazioni' e 'featuring', sono state eliminate le ridondanze che si erano create. Da questo processo viene così creato un dataset 'df similarity' composto da 887 righe per 3 colonne: 'song', 'recommendation' e 'similarity', in cui le prime due sono i codici identificativi delle canzoni legate dalla relazione e la terza il valore effettivo di similarità.

5.1.6 Membri delle band

Per modellare i membri delle band all'interno del grafo è necessario prima manipolare leggermente il dataset che ne contiene le informazioni; considerando che alcuni di questi sono già presenti all'interno di altre tabelle, come artisti o come autori, per evitare di creare ridondanze nella fase di creazione dei nodi è necessario filtrarli. Per fare ciò si crea una grande lista contenente tutti i nomi di interpreti e autori, senza ripetizioni; questa viene confrontata con la lista dei nomi dei membri. Il dataset viene diviso in due, osservando se un nome fosse già presente all'interno degli altri dataset o meno: per quelli già trovati sarà necessario semplicemente creare una relazione di appartenenza alla band; per quelli nuovi sarà invece richiesto anche di crearne i nodi.

5.2 Grafo Neo4j

Per creare il modello a grafo viene utilizzato Neo4j, una piattaforma nativa per la creazione, gestione e analisi di questo tipo di database. Per creare il modello vengono utilizzati dei comandi, tramite i quali si leggono i file .csv appena creati e si associano a nodi, relazioni e proprietà. Tutti questi comandi sono consultabili all'interno dell'Appendice. A questo punto si può descrivere la logica alla base del modello a grafo.

Si è deciso di creare nodi con diverse etichette, anche per distinguere entità dello stesso tipo ma con caratteristiche differenti:

- Canzone: ogni nodo indica un brano e possiede come proprietà tutte quelle elencate nella fase di esplorazione. Per le tracce non presenti nel catalogo Spotify, quindi senza questi valori, non c'è alcun problema, in quanto il grafo permette di avere nodi dello stesso tipo ma con famiglie di attributi diversi. Queste particolari tracce saranno caratterizzate solo da codice identificativo, titolo e edizione

- Persona: questo nodo indica tutte le istanze di persone, sia che siano interpreti, autori o membri di band

- Artista: indica tutti gli interpreti che hanno partecipato al Festival; questi sono caratterizzati da essere nodi con almeno due etichette, cioè Persona e Artista

- Autore: nodo per rappresentare gli autori dei brani; come per il precedente, ogni Autore è una Persona, e può essere anche Artista, nel caso in cui abbia interpretato la canzone ma l'abbia anche composta

- Band: indica i gruppi in gara nelle diverse edizioni
- Edizione: nodo riguardante l'anno. È stato deciso di modellare l'edizione come nodo e non solo come proprietà; ha relazioni unicamente con le canzoni in gara, ma permette di ottenere dei percorsi di percorrenza interessanti e distinguere facilmente i diversi anni
- Genere: genere a cui appartengono i vari artisti

Per quanto riguarda i membri delle band viene scelto di non creare un nodo apposito MEMBRO, per evitare di complicare troppo la struttura del grafo. Questa informazione sarà invece modellata come relazione, esposta successivamente. I membri non presenti tra i crediti dei brani, quindi non presenti nella tabella autori, saranno caratterizzati dall'essere semplicemente nodi di tipo Persona legati al nodo della propria Band. Non gli viene assegnata l'etichetta di Artista perché questa viene riservata agli artisti provenienti dalle API Spotify, in quanto possiedono anche il codice identificativo, condizione necessaria alla creazione di un indice per gli interpreti, che permette una navigazione più rapida del grafo.

A questo punto i nodi vanno collegati con delle relazioni, che possono assumere anch'esse delle proprietà; le relazioni scelte e create sono le seguenti.

- PARTECIPA: collega ogni Canzone con l'anno in cui è stata portata in gara; creare questo tipo di relazione invece che modellare l'edizione solamente come proprietà risolve problemi che potrebbero nascere in casi in cui lo stesso artista abbia partecipato a più edizioni e permette di creare dei cluster di tracce divisi per anno. La relazione non possiede alcuna proprietà specifica. ((Brividi)-[:PARTECIPA]→(2022))
- SCRIVE: unisce ogni nodo di tipo Autore alle tracce che ha composto; naturalmente ogni autore può essere collegato a diverse Tracce, sia della stessa edizione che edizioni diverse. La relazione non possiede alcuna proprietà specifica. ((Mahmood)-[:SCRIVE]→(Brividi))
- CANTA: relazione tra il brano e l'interprete che l'ha cantato al Festival, sia esso sia un Artista o una Band; come per la precedente, ogni nodo riferito agli interpreti può avere più relazioni CANTA, nel caso in cui questo abbia partecipato a più edizioni. La relazione non possiede alcuna proprietà specifica. ((Mahmood)-[:CANTA]→(Brividi))
- APPARTIENE: relazione che unisce un Artista o Band ai generi musicali a cui appartiene; anche qua, a ogni cantante possono essere assegnati diversi generi. Non possiede alcuna proprietà specifica. ((Mahmood)-[:APPARTIENE]→(italian pop))
- COLLABORA: unisce tra loro gli Autori che hanno collaborato alla scrittura di un brano. Ogni Autore, che si ricorda può essere anche un artista, è potenzialmente collegato a diversi colleghi, con cui ha collaborato nel tempo. Per rendere più chiara ed esaustiva la relazione, ad essa sono state anche assegnate alcune proprietà, cioè id track, titolo e edizione. In questo modo, nel caso in cui due persone avessero collaborato più di una volta, si potrebbero distinguere i collegamenti andando a scoprire per quali canzoni e in che anno. Naturalmente, la presenza di più collegamenti tra le stesse due persone non è un errore, ma un'indicazione del fatto che hanno collaborato diverse volte. ((Mahmood)-[:COLLABORA]→(Charlie Charles))
- FEATURING: unisce due nodi Artista che si sono esibiti insieme durante le serate del Festival. Come per la precedente è caratterizzata da proprietà di titolo e ID della traccia in questione. ((Mahmood)-[:FEATURING]→(Blanco))
- MEMBRO: relazione che lega i Membri delle Band al proprio gruppo di riferimento; è possibile che contenga la proprietà strumento. ((Stefano

Belisari)-[:MEMBRO]→(Elio e le storie tese))

- SIMILE: relazione che lega due nodi Canzone in base alla similarità dei loro testi. Contiene una sola proprietà numerica, cioè similarity, che si riferisce al valore di cosine similarity tra i lyrics delle due tracce. ((Brividi)-[:SIMILE]→(Così sbagliato))

È necessario notare che per evitare ridondanze è stato scelto di eliminare tutte le relazioni COLLABORA tra membri della stessa Band, perché si è ritenuto che fosse scontato che tutti questi partecipassero alla creazione dei brani del gruppo, anche se alcuni non fanno parte dei crediti ufficiali della traccia.

Nell'immagine 8 è possibile osservare la struttura finale del grafo, semplificata eliminando i nodi Artista e Autore. Questa scelta è stata presa con il solo scopo di rendere più efficace la visualizzazione, in quanto ogni artista e ogni autore è anche una Persona, perciò le relazioni vengono mantenute. Si può notare come le Persone siano caratterizzate da connessioni come featuring e collaborazione che girano su loro stesse. Queste inoltre si legano anche ai vari generi e, naturalmente, alle canzoni in gara, a loro volta a contatto con l'edizione di appartenenza. In totale il grafo è composto da 806 nodi e 3376 relazioni.

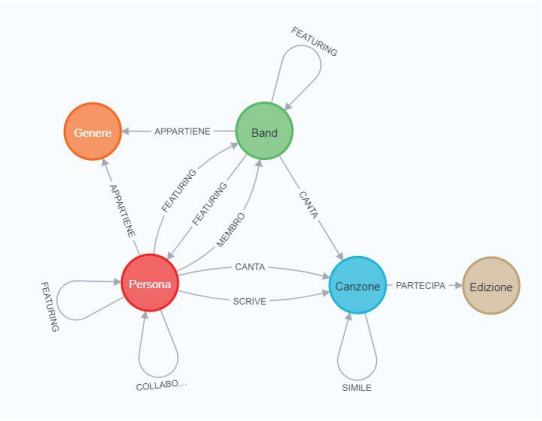

Figura 8: Struttura del grafo

Per concludere, il dataset è salvato in formato *.graphml*, in modo tale da poter essere condiviso senza il bisogno di eseguire tutti i comandi presenti in Appendice.

6 Data quality

Per eseguire un'analisi della qualità dei dati sono state scelte due misure, cioè coerenza e completezza, che sono però state analizzate in momenti diversi dello sviluppo. La prima è stata necessaria appena conclusa la fase di ottenimento dati, per rendere coerenti tutte le informazioni riferite alla stessa persona. La seconda è stata utilizzata per osservare quanto le tecniche di download dei dati abbiano prodotto dei risultati esaustivi.

6.1 Coerenza

Come detto, la coerenza dei dati provenienti dalle differenti sorgenti è stata verificata appena conclusa la fase di download; questo perché avendo utilizzato diverse fonti eterogenee, si è notato che a volte la stessa persona era citata in modi differenti. Ai fini della creazione di un grafo corretto, è fondamentale che non siano presenti nodi diversi che si riferiscono allo stesso individuo.

Le sorgenti analizzate da cui sarebbero potuti nascere dei problemi sono quelle di Spotify, Wikipedia e recensiamomusica.com, perché è proprio da queste che vengono scaricate le informazioni riguardo le persone. Più nello specifico, è necessario studiare come ci si riferisca a interpreti, autori e membri delle band. Per correggere eventuali errori sono stati utilizzati metodi di analisi automatica, in cui viene eseguito un confronto tra stringhe (Record Linkage), e metodi manuali, per i quali è stato necessario cercare manualmente casi particolari non identificati nella prima fase.

Il primo passaggio dell'analisi prende in considerazione cantanti e autori, per cercare corrispondenze ed errori. Dopo aver eliminato i duplicati contenuti nelle tabelle a loro riferite, ma mantenendo comunque l'informazione sull'edizione di partecipazione, utilizzata successivamente per creare un raggruppamento, si crea un'unica lista di nomi a cui verrà applicata la tecnica. Il record linkage consiste in un metodo di ricerca di stringhe simili, che perciò si riferiscono alla stessa entità; applicandola alla lista e scegliendo diversi valori di soglia si ottengono diversi risultati. Con una soglia a 1 vengono mostrati

solo le corrispondenze perfette, identiche carattere per carattere; in questo caso ne sono state trovate 76. Applicando invece un valore più basso, come 0.85, si ottengono delle coppie di stringhe simili ma non perfettamente uguali, che spesso differiscono per un solo carattere. In questo modo si può scoprire quali persone siano citate in modo diverso, così da poter applicare delle correzioni. In particolare tra queste si notano degli errori riferiti al diverso carattere apostrofo e ad alcuni errori di battitura per quanto riguarda recensiamomusica.com. Dopo averli corretti si procede all'analisi manuale dei dati.

Questo è un processo dispendioso in termini di tempo e non applicabile per grandi moli di dati, ma permette di avere un controllo più efficace e mirato su casi limite non riconoscibili dagli algoritmi. In particolare, è stato necessario perché si sono notate delle differenze per quanto riguarda i nomi d'arte, quelli di battesimo e il modo in cui venivano utilizzati nelle diverse fonti. Infatti, recensiamomusica.com a volte si riferiva alla stessa persona utilizzando sia l'alias che il nome reale (Elio - Stefano Belisari); in questi casi si è scelto di uniformare al nome di battesimo. Altre volte si sono notate incongruenze in quanto Spotify si riferisce a una persona con il proprio nome da artista, cosa non necessariamente vera per le altre fonti (Annalisa - Annalisa Scarrone). In questi casi si è deciso di mantenere come corretto il nome derivante dalla piattaforma streaming, in quanto si è ritenuto che fosse una fonte affidabile a cui uniformare le altre. Come detto, questi casi limite non sono individuabili facilmente dagli algoritmi e richiedono delle conoscenze di dominio; la combinazione dei due porta sicuramente ai risultati migliori.

Lo stesso procedimento è stato poi eseguito anche per quanto riguarda i membri dei gruppi, ripuliti di alias già nella fase di download, ma sui quali è stato comunque necessario effettuare un controllo. Molti di questi, infatti, sono presenti anche come autori delle canzoni che la loro band interpreta. I risultati rispecchiano quelli precedenti, con diverse entità accoppiate correttamente, e altre che presentano errori di battitura; anche in questo caso, dopo aver corretto gli errori trovati, viene effettuato un controllo manuale.

Un ulteriore step di controllo della qualità dei dati viene effettuato sulla tabella tracce; analizzando infatti i valori assunti dall'attributo Posizione, ottenuto tramite scraping di tabelle Wikipedia, si notano delle incongruenze per alcuni valori. In particolare, alcuni brani non classificati alla serata finale invece che NF presentano il costrutto "NF (E4)", è presente un errore per la sesta posizione e per gli esclusi. Cercando le istanze con questi errori è possibile correggerli per ottenere delle rappresentazioni uniformi e corrette.

6.2 Completezza

L'analisi di questa misura è stata effettuata come ultimo step dopo l'ottenimento e la pulizia dei dati. La completezza si riferisce alla quantità di informazioni contenute nel database rispetto alla realtà che rappresentano; per analizzarla sono state considerate tutte le tabelle riferite alle entità e se ne sono studiati i valori mancanti, per righe e colonne. Per fare ciò è stata utilizzata la libreria Python 'dataprep': questa permette di ottenere un report accurato e compilato automaticamente utilizzando una sola semplice istruzione. All'interno del report sono contenute diverse esplorazioni, come numero di valori nulli per ogni colonna, correlazioni e istogrammi. Questi grafici sono anche stati utilizzati nella fase di analisi descrittiva delle varie tabelle.

Brani

Per quanto riguarda i brani, nel dataset sono presenti 237 righe e 16 colonne. A livello di tabella, si sono riscontrate 140 celle con valori nulli, corrispondenti al 3.7% del totale. Per quanto riguarda gli attributi invece, quelli numerici derivanti da Spotify, come ballabilità ed energia, presentano 11 righe con valori nulli, cioè il 4.64%. Questi brani non sono presenti all'interno del catalogo della piattaforma, perciò non c'è alcun modo per integrarli. L'altra colonna che presenta valori nulli è quella relativa al testo, all'interno della quale se ne trovano 10; questo è dovuto al fatto che le API di Genius neces-

sitano sia di titolo che artista per ottenere il lyrics. Utilizzando la tabella pivot che lega brani e interpreti è possibile ottenere queste informazioni, ma, come già osservato precedentemente, per le canzoni non in catalogo non è possibile ottenere il nome del cantante.

Artisti

Delle 233 righe presenti nel dataset solo sette hanno valori nulli (0.8% del totale) e si riferiscono solamente a tre artisti diversi per il campo periodo attività, in particolare Noemi (4 volte), Le vibrazioni (2 volte) e Frenetik&Orang3 (1 volta); le ripetizioni sono dovute al fatto che alcuni di essi hanno partecipato a più edizioni. Questo problema deriva dal fatto che l'informazione è stata ottenuta tramite scraping, tecnica potente ma non particolarmente robusta, che può quindi portare a risultati inconsistenti. In alcuni casi le espressioni regolari utilizzate per isolare la sezione di pagina utile non sono risultate corrette, altre volte la ricerca Google non porta a una risposta immediata, per cui per ottenere il dato sarebbe stato necessario consultare altri siti; in alcuni casi limite si è notato che l'informazione non è assolutamente presente online.

Autori

Analizzando la tabella autori non si riscontrano valori nulli, infatti questa è un pivot che lega a ogni canzone tutte le persone che hanno contribuito a comporla. Non essendoci altri attributi che si riferiscono agli autori, non sono presenti valori nulli in nessuna delle 632 righe che compongono il dataset.

Generi

Anche per quanto riguarda i generi valgono le stesse considerazioni fatte per gli autori; essendo un pivot non sono presenti valori nulli.

7 Conclusioni e sviluppi futuri

Dopo aver sviluppato il progetto si possono trarre delle conclusioni, riflettendo sulle difficoltà incontrate e ciò che invece è stato implementato più facilmente. Per quanto riguarda il download dei dati, si è riscontrato che le API permettono di sviluppare un work-flow molto ordinato, in quanto forniscono dati da fonti certe e in formati facili da utilizzare. Un punto a loro sfavore è il fatto che i dati ottenibili sono limitati a quelli che il provider decide di fornire. Al contrario lo scraping permette di scaricare sostanzialmente qualsiasi informazione online, ma il processo è molto lungo. È necessario prestare molta attenzione alle fasi di pulizia e controllo, che anche in questo caso sono state le più dispendiose e lente. Un'altra accortezza da avere riguarda il fatto che le pagine online da cui si sono ottenuti i dati potrebbero essere modificate e questo richiederebbe un controllo periodico per verificarne la correttezza. Detto questo, il risultato finale si può ritenere buono, avendo inserito all'interno del database oltre ottocento nodi legati da molte connessioni. Sono nate piccole comunità di individui legati da diversi tipi di collaborazione, che si sono poi riflettuti anche per quanto riguarda i brani. Avere delle relazioni di similarità sia tra i cantanti, che tra i brani potrebbe rivelarsi una base interessante per degli sviluppi futuri. Da notare anche che non si sono riscontrate criticità dal punto di vista del calcolo computazionale, in quanto tutti i task sono risultati leggeri.

In particolare, il principale sviluppo potrebbe riguardare l'ampliamento del grafo tramite l'aggiunta di altre edizioni, sempre però considerando il fatto che più si va indietro e più è probabile che Spotify non abbia i brani in catalogo. Un ulteriore miglioramento potrebbe riguardare l'uso di tecniche più avanzate per il calcolo della similarità tra i testi, come ad esempio degli embedding.

Riferimenti bibliografici

[1] developer.spotify.com/documentation /web-api/reference/#/operations/get-several-audio-features

[2] github.com/charlie86/spotifyr

[3] www.rcharlie.com/spotifyr/

[4] www.crummy.com/software/BeautifulSoup
/bs4/doc/index.html?highlight=select
_one#

[5] pypi.org/project/requests/

[6] pypi.org/project/wikipedia/

[7] docs.genius.com/

[8] neo4j.com/labs/apoc/4.3/overview/
apoc.export/apoc.export.graphml.all/

[9] https://neo4j.com/labs/apoc/4.1/ installation/#:~:text=APOC%20Full
%20can%20be%20installed,see
%20the%20%22Installed%22%20message

Appendix

	ballabilità	energia	rumorosità	parlato	positività	tempo	durata
count	226	226	226	226	226	226	226
mean	0.556	0.676	-5.844	0.064	0.431	120.703	3.525
std	0.112	0.153	1.651	0.054	0.210	26.040	0.375
min	0.255	0.257	-12.075	0.026	0.069	61.000	2.550
25%	0.478	0.583	-6.747	0.034	0.272	101.250	3.280
50%	0.557	0.687	-5.459	0.043	0.393	122.000	3.525
75%	0.640	0.795	-4.713	0.065	0.601	136.750	3.737
max	0.831	0.976	-2.379	0.361	0.963	206.000	5.040

Tabella 3: Describe tabella brani

	periodo_attivita
count	226.000
mean	1998.646
std	15.187
min	1956.000
25%	1990.000
50%	2002.00
75%	2011.000
max	2020.000

Tabella 4: Describe tabella artisti

Comandi di creazione del grafo

- LOAD CSV WITH HEADERS FROM 'file:///tracce_con_testo.csv' AS row FIELDTERMINATOR ',' CREATE (c:Canzone) SET c=row
- LOAD CSV WITH HEADERS FROM 'file:///edizione.csv' AS row FIELDTERMINATOR ',' CREATE(e:Edizione) SET e=row
- MATCH (c: Canzone) MATCH (e: Edizione) WHERE c.edizione = e.anno MERGE (c)-[:PARTECIPA]->(e)
- LOAD CSV WITH HEADERS FROM 'file:///solo_autori.csv' AS row MERGE (c:Canzone id_track:row.id_track) merge (p:Persona :Autore nome:row.Autori) create (p)-[s:SCRIVE]->(c)
- LOAD CSV WITH HEADERS FROM 'file:///artisti_autori.csv' AS row MERGE (c:Canzone id_track:row.id_track) merge (p:Persona :Artista :Autore id_artist:row.id_artist, nome:row.Autori, periodo_attivita:row.periodo_attivita) create (p)-[s:SCRIVE]->(c)
- LOAD CSV WITH HEADERS FROM 'file:///solo_artisti.csv' AS row MERGE (p:Persona :Artista id_artist:row.id_artist, nome:row.nome, periodo_attivita:row.periodo_attivita)
- LOAD CSV WITH HEADERS FROM 'file:///solo_gruppi.csv' AS row CREATE (:Band nome:row.nome, id_artist: row.id_artist, periodo_attivita:row.periodo_attivita)

- LOAD CSV WITH HEADERS FROM 'file:///pivot_artisti_final.csv' AS row match (c:Canzone id_track:row.id_track) match (p:Persona :Artista id_artist:row.id_artist) create(p)-[:CANTA]->(c)

- LOAD CSV WITH HEADERS FROM 'file:///pivot_artisti_final.csv' AS row match (c:Canzone id_track:row.id_track) match (b:Band id_artist:row.id_artist) create(b)-[:CANTA]->(c)

- LOAD CSV WITH HEADERS FROM 'file:///generi_nodi.csv' AS row CREATE (g:Genere) SET g = row

- LOAD CSV WITH HEADERS FROM 'file:///generi_artisti.csv' AS row MATCH (p:Persona id_artist: row.id_artist) MATCH (g:Genere genere: row.genres) CREATE (p)-[:APPARTIENE]->(g)

- LOAD CSV WITH HEADERS FROM 'file:///generi_artisti.csv' AS row MATCH (b:Band id_artist: row.id_artist) MATCH (g:Genere genere: row.genres) CREATE (b)-[:APPARTIENE]->(g)

- LOAD CSV WITH HEADERS FROM 'file:///collaborazioni.csv' AS row MATCH (p:Persona nome: row.Autori) MATCH (g:Persona nome: row.collaboratore) CREATE (g)-[:COLLABORA anno: row.edizione, canzone: row.id_track, titolo: row.titolo]->(p)

- LOAD CSV WITH HEADERS FROM 'file:///featuring_cantanti.csv' AS row MATCH (p:Persona id_artist: row.id_artist) MATCH (g:Persona id_artist: row.id_artist2) CREATE (g)-[:FEATURING anno: row.edizione, canzone: row.id_track, titolo: row.titolo]->(p)

- LOAD CSV WITH HEADERS FROM 'file:///featuring_cantanti.csv' AS row MATCH (p:Persona id_artist: row.id_artist) MATCH (g:Band id_artist: row.id_artist2) CREATE (g)-[:FEATURING anno: row.edizione, canzone: row.id_track, titolo: row.titolo]->(p)

- LOAD CSV WITH HEADERS FROM 'file:///featuring_cantanti.csv' AS row MATCH (p:Band id_artist: row.id_artist) MATCH (g:Persona id_artist: row.id_artist2) CREATE (g)-[:FEATURING anno: row.edizione, canzone: row.id_track, titolo: row.titolo]->(p)

- LOAD CSV WITH HEADERS FROM 'file:///featuring_cantanti.csv' AS row MATCH (p:Band id_artist: row.id_artist) MATCH (g:Band id_artist: row.id_artist2) CREATE (g)-[:FEATURING anno: row.edizione, canzone: row.id_track, titolo: row.titolo]->(p)

- LOAD CSV WITH HEADERS FROM 'file:///df_similarity.csv' AS row MATCH (p:Canzone id_track: row.song) MATCH (g:Canzone id_track: row.recommendation) CREATE (g)-[:SIMILE cosine_similarity: row.similarity]->(p)

- LOAD CSV WITH HEADERS FROM 'file:///membri_ripetuti.csv' AS row MATCH (b:Band nome: row.gruppo) MATCH (p:Persona nome: row.nome) CREATE (p)-[s:MEMBRO strumento: row.strumento]->(b)

- LOAD CSV WITH HEADERS FROM 'file:///membri_nuovi.csv' AS row CREATE (p:Persona nome: row.nome)

- LOAD CSV WITH HEADERS FROM 'file:///membri_nuovi.csv' AS row MATCH (b:Band nome: row.gruppo) MATCH (p:Persona nome: row.nome) CREATE (p)-[s:MEMBRO strumento: row.strumento]->(b)
- MATCH p=(a:Persona)-[r:MEMBRO]->(:Band)<-[:MEMBRO]-(b:Persona) where id(a)>id(b) and (a)-[:COLLABORA]-(b) with a,b match (a)-[x:COLLABORA]-(b) delete x
- call db.schema.visualization
- call apoc.export.graphml.all("sanremo.graphml",)