

SPOTIWHY:

A document-oriented database for musical analysis

Davide Meloni, Alberto Raimondi, Silvia Santamaria

*Università degli studi di Milano-Bicocca
Progetto di Data Management*

Abstract

Spotify ricopre una posizione importante nel mercato della musica digitale. Le classifiche che rende disponibili ogni giorno per ogni Paese danno maggiore visibilità alle canzoni che vi compaiono. Questo report nasce con lo scopo di evidenziare le caratteristiche che accomunano le canzoni in classifica con particolare attenzione alla fase di raccolta ed immagazzinamento dei dati. Per la realizzazione di questo progetto si sono unite diverse fonti di dati e le informazioni raccolte hanno dato vita ad un database NoSQL MongoDB.

Keywords: Spotify, Music, Database



Febbraio, 2018

Indice

1	Introduzione	3
2	Raccolta dei dati e preparazione del dataset	3
2.1	Fonti	3
2.1.1	Spotify CSV	4
2.1.2	Spotify API	5
2.1.3	Wikipedia API	6
2.1.4	AZLyrics	6
2.2	Ulteriori dettagli	6
2.2.1	Posizione in classifica e numero di visualizzazioni . . .	6
2.2.2	Prima pubblicazione (anno) e paese di provenienza . .	7
2.2.3	Genere	8
2.3	Creazione del database	9
3	Creazione del dataset	9
3.0.1	Multiprocessing e Amazon VM	10
3.0.2	Struttura del Documento di MongoDB	12
4	Esplorazione dei dati	14
4.1	Statistiche descrittive	14
4.2	Word Cloud	16
5	Criticità e limitazioni	17
6	Conclusioni	17

1. Introduzione

Con l'avvento del digitale è cambiato radicalmente il modo in cui le persone ascoltano la musica e come gli artisti guadagnano dai loro brani. Negli ultimi anni la vendita di album è diminuita, ma grazie alla diffusione dello streaming il guadagno dell'industria musicale è addirittura aumentato.

Spotify gioca un ruolo fondamentale in questo mercato, fornendo un servizio musicale, di podcast e di video streaming. Rilasciato a Stoccolma nel 2008 dalla startup Spotify AB, Spotify fornisce accesso a più di 30 milioni di canzoni ed è disponibile nella maggior parte dell'Europa, America, Australia, Nuova Zelanda e Asia. Si stima che Spotify abbia circa 140 milioni di utenti.

Quotidianamente viene rilasciata una classifica per ogni paese con le canzoni più ascoltate. Se una canzone è presente in una classifica di Spotify, aumentano le possibilità che un altro utente la ascolti e quindi anche le possibilità per gli artisti di guadagnare di più. Inoltre le classifiche costituiscono una fotografia di quello che le persone ascoltano nella vita di tutti i giorni.

Questo report nasce con lo scopo di raccogliere dati per identificare le caratteristiche delle canzoni che ne permettono l'ascesa in classifica.

2. Raccolta dei dati e preparazione del dataset

In questo capitolo vengono riportate le metodologie adottate nella fase di raccolta dei dati. Si tratta di dati esterni, ottenuti tramite download, tecniche di scraping e in altri casi attraverso le API disponibili.

2.1. Fonti

I dati sono stati raccolti dalle seguenti fonti:

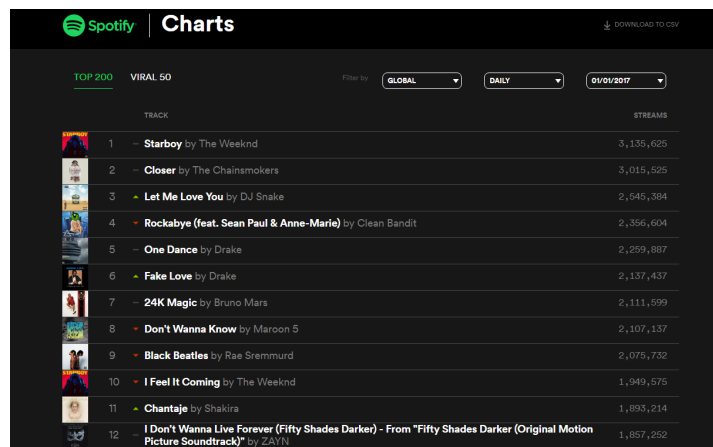
- Spotify CSV [1]
- Spotify API [2]
- Wikipedia API [3]
- Music Brainz [4]
- AZLyrics [5]

2.1.1. Spotify CSV

Spotify offre un servizio di tipo freemium: alcuni servizi sono gratis mentre ulteriori funzionalità sono offerte tramite una sottoscrizione a pagamento. Viene pubblicato quotidianamente un file in formato CSV con le 200 canzoni più ascoltate per ciascun Paese in cui opera sul sito esterno spotifycharts.com. Questi file sono stati raccolti attraverso gli script Python *global.py* e *spotify_worldwide.py*, eseguiti nella notte del 31 Dicembre 2017 per mezzo di un Raspberry pi. Si sono quindi ottenuti due csv contenenti tutte le canzoni per ogni nazione.

Nel file csv risultante sono disponibili le seguenti informazioni:

- nome della traccia
- nome dell'artista
- URL: racchiude l'ID di Spotify che permette di identificare univocamente una canzone
- posizione raggiunta in classifica (giornaliera). Per un ulteriore approfondimento si rimanda al capitolo 2.2.1
- numero di visualizzazioni (giornaliere). Per un ulteriore approfondimento si rimanda al capitolo 2.2.1



Spotify		Charts		DOWNLOAD TO CSV	
TOP 200		VIRAL 50		Filter by: GLOBAL DAILY 01/01/2017	
	TRACK		STREAMS		
1	Starboy by The Weeknd		3,135,625		
2	Closer by The Chainsmokers		3,015,525		
3	Let Me Love You by DJ Snake		2,545,384		
4	Rockabye (feat. Sean Paul & Anne-Marie) by Clean Bandit		2,356,604		
5	One Dance by Drake		2,259,887		
6	Fake Love by Drake		2,137,437		
7	24K Magic by Bruno Mars		2,111,599		
8	Don't Wanna Know by Maroon 5		2,107,137		
9	Black Beatles by Rae Sremmurd		2,075,732		
10	I Feel It Coming by The Weeknd		1,949,575		
11	Chantaje by Shakira		1,893,214		
12	I Don't Wanna Live Forever (Fifty Shades Darker) - From "Fifty Shades Darker (Original Motion Picture Soundtrack)" by ZAYN		1,857,252		

Figura 1: Classifica globale della TOP200 di Spotify dell'1 Gennaio 2017, visibile al sito [SpotifyCharts\[1\]](http://SpotifyCharts[1]).

2.1.2. Spotify API

I dati sono stati integrati anche grazie all'uso dell'API di Spotify per mezzo dell'ID di Spotify precedentemente estratto dal URL. Queste API sono state interrogate attraverso la libreria **requests** per Python e l'oggetto ottenuto in risposta è stato trattato utilizzando la libreria **JSON**. L'API mette a disposizione diverse informazioni riguardanti le canzoni e l'artista:

- artista
 - numero di follower
 - lista con i differenti generi musicali propri dell'artista
 - lista con altri artisti correlati¹
 - foto dell'artista²
- album
 - nome dell'album
 - data di pubblicazione
 - copertina dell'album²
- tracce
 - acousticness³
 - danceability³
 - energy³
 - loudness³
 - valence³
 - popolarità³
 - accordo base ("tonica") della scala musicale utilizzata
 - modo (scala maggiore o minore)
 - numero della traccia nell'album
 - durata della traccia in millisecondi

¹gli artisti correlati sono scelti da un algoritmo sviluppato da Spotify

²include un URL, larghezza e altezza dell'immagine

³questa variabile è il risultato di un algoritmo sviluppato da Spotify

2.1.3. Wikipedia API

Si è fatto uso di **Wikipedia** per il recupero di informazioni riguardanti il sesso degli artisti o la loro appartenenza ad un gruppo (si rimanda al capitolo 2.2.3 per ulteriori dettagli). Le API esposte da Wikipedia sono state le più complesse da utilizzare data la vastità delle pagine disponibili e la fantasia nei nomi d'arte scelti da cantanti e gruppi. Per il recupero del sesso dell'artista si è operata prima una richiesta della pagina attraverso il nome del cantante e, in caso di risposta negativa, lo script effettua una nuova richiesta ma aggiungendo i più comuni suffissi utilizzati per identificare gli artisti. Ottenuta la pagina, inglese relativa all'artista, il sesso viene determinato attraverso un'operazione di scraping. Per mezzo dell'API di Wikipedia si sono estratte inoltre dall'infobox le seguenti informazioni:

- anno della prima pubblicazione. Per un ulteriore approfondimento si rimanda al capitolo 2.2.2
- Paese di provenienza. Per un ulteriore approfondimento si rimanda al capitolo 2.2.2

2.1.4. AZLyrics

Dal sito **AZLyrics** si è tentato di ottenere i testi di ogni canzone. Ciascuna parola presente nel testo viene salvata in una lista. La lista contiene un valore nullo (NaN) nel caso in cui la richiesta sia fallita o il sito sia risultato sprovvisto del testo ricercato. L'interrogazione avviene attraverso lo scraping della risposta relativa alla pagina HTML con in aggiunta il titolo della canzone. Queste operazioni vengono svolte tramite l'utilizzo delle librerie **requests** e **BeautifulSoup**. Nonostante il codice fosse funzionante e testato in locale ci si è resi conto che il sito blocca l'accesso alle proprie API da servizi di cloud computing come AWS, per questo motivo l'aggiunta dei testi è rimandata a un'integrazione successiva attraverso proxy o personal computer.

2.2. Ulteriori dettagli

2.2.1. Posizione in classifica e numero di visualizzazioni

Tramite una procedura automatizzata implementata in linguaggio Python si sono scaricati i CSV contenenti le 200 canzoni più ascoltate per ogni

Paese. Si sono quindi unite le canzoni in un unico CSV senza ripetizioni. Per ciascuna canzone si è ricavata la posizione raggiunta e il numero di ascolti per ogni giorno in ogni Paese. Nel caso in cui la canzone non venisse trovata, si associa 999 alla posizione e 0 al numero di ascolti. Tutte le informazioni vengono inserite in due liste separate dove la posizione nell'array permette di identificare il giorno corrispondente.

2.2.2. Prima pubblicazione (anno) e paese di provenienza

Le variabili relative alla data di inizio carriera e al Paese di provenienza si sono ottenute tramite l'API di Wikipedia. L'API restituisce un file JSON con tutte le informazioni presenti sulla pagina Wikipedia dell'artista. Si è quindi fatto scraping sulla parte relativa all'infobox.

Background information	
Birth name	Edward Christopher Sheeran
Born	17 February 1991 (age 26) <div> Halifax, West Yorkshire, England</div>
Origin	 Framlingham, Suffolk, England
Genres	Pop ^[1] · folk pop ^{[2][3]}
Occupation(s)	Singer-songwriter · record producer · guitarist
Instruments	Vocals · guitar
Years active	2004–present
Labels	Asylum · Atlantic · Elektra
Associated acts	Taylor Swift · One Direction · James Blunt · Zeph Ellis
Website	www.edsheeran.com [ⓘ]

Figura 2: Infobox della pagina Wikipedia del cantante inglese Ed Sheeran.

Per ottenere il Paese di provenienza si è utilizzata la libreria **Geotext**[6] che raccoglie i nomi di tutti i Paesi del mondo e oltre a selezionare l'informazione corretta, uniforma i dati raccolti. Tuttavia non tutti i cantanti hanno una pagina Wikipedia in lingua inglese e quindi non è sempre stato possibile ottenere il Paese di provenienza con questa procedura. In caso di insuccesso si è interrogata la libreria MusicBrainz. Si è utilizzata questa libreria come seconda fonte in quanto le informazioni presenti non sono standardizzate e sono necessari tempi maggiori per ottenere una risposta. Anche in questo

caso per alcuni artisti non è presente l'informazione ricercata. In tal caso si è assunto che il Paese di provenienza corrispondesse al Paese in cui l'artista ha raggiunto la posizione in classifica più alta il più velocemente possibile. Infatti in questo ultimo caso si ritiene che l'artista sia poco famoso a livello globale non disponendo di una pagina Wikipedia e una sezione dedicata in MusicBrainz.

2.2.3. Genere

Dal momento in cui il genere dell'artista non è un'informazione presente nell'infobox di Wikipedia, si è trovato un escamotage per ricavarlo. Si è cercata la pagina Wikipedia inglese relativa⁴ e si sono analizzate le parole presenti nel testo e in particolare si è conteggiata la presenza di alcune parole chiave che si ritiene siano associate a un soggetto maschile, femminile o ad un gruppo di più persone. Nella fattispecie si sono conteggiati i pronomi personali e il verbo essere nella funzione di ausiliare e sono stati distinti in tre categorie:

- Locuzioni che si riferiscono ad un singolo soggetto femminile (she, is, her, singer, ...);
- Locuzioni che si riferiscono ad un singolo soggetto maschile (he, is, his, rapper, ...);
- Locuzioni che si riferiscono ad un gruppo o una band (they, are, band, group, ...).

A fronte del conteggio sono possibili i seguenti tre scenari:

1. Le locuzioni che si riferiscono ad un singolo cantante (maschile o femminile) sono superiori di oltre il doppio delle parole che si riferiscono ad un gruppo o una band. In questo caso si verifica se i pronomi personali femminili siano superiori a quelli maschili e quindi si conclude che il cantante sia un soggetto femminile. In caso contrario si conclude che il cantante sia un soggetto maschile.

⁴nel caso in cui le parole della pagina Wikipedia siano inferiori a 100 non viene fatto alcun conteggio. Questo per evitare di considerare le pagine di disambiguazione

2. Le locuzioni che si riferiscono ad una gruppo di artisti sono superiori di oltre il doppio delle parole che si riferiscono ad un singolo. In questo caso si conclude che si tratti di una band.
3. In tutti gli altri casi non è possibile dare una risposta univoca e quindi si effettua una nuova ricerca del cantante su MusicBrainz. Anche in questo caso si tiene MusicBrainz come seconda fonte in quanto più lenta a fornire i risultati.

2.3. Creazione del database

Per la creazione del database con MongoDB 3.4.1 [7] si sono unite le informazioni provenienti dalle diverse fonti e quindi si sono inseriti i documenti relativi ad ogni singola canzone.

Data la mole di dati disponibile e le finalità del progetto, si è optato per escludere le canzoni con valori mancanti in campi di un certo rilievo per le analisi, come ad esempio l'assenza del nome dell'artista.

Si contano 16539 canzoni non duplicate, relative a 5165 artisti differenti.

3. Creazione del dataset

MongoDB rientra nei database **NoSQL** di tipo Document Based Management System. La scelta è stata dettata dallo scopo del progetto, infatti, i dati che vengono trattati presentano 2 delle caratteristiche proprie dei Big Data:

- Volume: ogni documento pesa in media 0,524 MB così che la collezione di MongoDB arrivi a circa 10GB (compresi file di journaling) per 16539 canzoni.
- Varietà: si sono uniti in un unico tipo di documento dati in formato CSV a dati in formato JSON provenienti da diverse fonti e con strutture diverse.

MongoDB è dotato di una struttura basata su documenti che vengono memorizzati per mezzo di uno schema documentale molto flessibile che permette di gestire facilmente grandi moli di dati di grande varietà.

Secondo il CAP theorem è impossibile per un sistema informatico distribuito soddisfare contemporaneamente tutte le seguenti tre garanzie:

- Consistenza: se una stessa informazione è contenuta su più nodi, questa è sempre uguale per tutte le repliche.
- Disponibilità: è sempre possibile ad ogni richiesta ricevere una risposta su ciò che è riuscito o fallito.
- Tolleranza alle partizioni: corretto funzionamento delle partizioni ovvero, anche se una parte del sistema fallisce, continua ad operare.

MongoDB gode di due delle proprietà del CAP theorem: consistenza e tolleranza alle partizioni. In un ottica di utilizzo futuro del dataset per scopi di analisi è stato scelto di rinunciare alla *availability* per ottenere dati consistenti e sicuri anche se non sempre disponibili.

MongoDB è stato scelto quindi per la sua grande flessibilità nella gestione di dati molto vari come nel caso di informazioni di cui non si conosce la numerosità (ad esempio il genere musicale relativo ad un artista). La capacità di scalare efficientemente unita al linguaggio di query dichiarativo hanno reso MongoDB una scelta ottimale per le finalità di questo elaborato.

3.0.1. Multiprocessing e Amazon VM

Il tempo necessario per collezionare le informazioni relative ad una singola canzone dalle diverse API si è stimato essere pari a circa 18 secondi. Considerando che il dataset ottenuto contiene 16539 canzoni distinte il tempo totale dell'esecuzione dello script sarebbe risultato eccessivo (circa 5 giorni). Per questo motivo si sono messi in atto degli espedienti per velocizzare il processo. La tecnica del multiprocessing è sembrata essere la soluzione migliore per risolvere questo tipo di problema.

Utilizzando la libreria nativa di Python 3.6 *multiprocessing.py*[8] è stato possibile suddividere il carico di lavoro tra diversi workers pari al numero di CPU disponibili sul terminale di esecuzione. Grazie alla parallelizzazione del lavoro il tempo di esecuzione è diminuito drasticamente.

Dato il numero limitato di CPU fornite dalle macchine virtuali Azure messe a disposizione dall'Università degli Studi di Milano-Bicocca, si è optato per l'utilizzo di una istanza **EC2 Amazon** dotata di 16 CPU e 64 GB di RAM. La scelta del tipo di macchina è stata dettata dalle esigenze del caso: si è posta particolare attenzione al numero di CPU e quindi di richieste al secondo da indirizzare al sito di Spotify. Un numero limitato di CPU rende il processo estremamente lento, un numero eccessivo di CPU porta ad

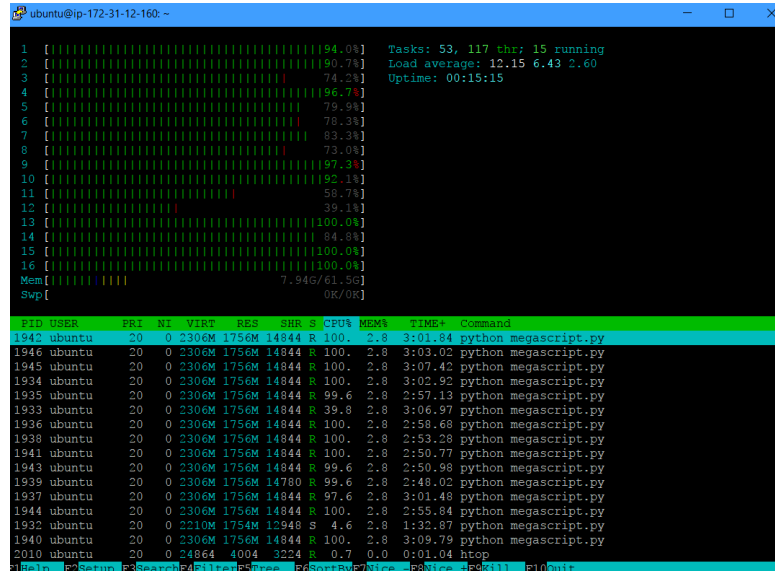


Figura 3: Le 16 CPU dell'istanza EC2 in uso durante l'esecuzione dello script.

un blocco temporaneo delle richieste. Ogni volta che un worker raccoglie tutte le informazioni relative ad una canzone, queste vengono inserite in un dizionario, poi convertito in formato BSON e aggiunto alla collezione di MongoDB denominata *songs*.

In questo modo il tempo di calcolo è risultato essere pari a 19 ore, minore di oltre 6 volte rispetto al tempo inizialmente preventivato.

Nel mettere in atto la procedura del multiprocessing sono sorte ulteriori problematiche. Infatti il tempo di esecuzione è comunque rimasto superiore ad un'ora, tempo di validità del token richiesto dall'API di Spotify. Sfortunatamente il refresh automatico del token è riservato a developers registrati. Questo problema è stato ovviato attraverso un refresh semi-manuale per mezzo di un cron-job con cadenza oraria.

Per via della difficoltà nel traceback di errori per i processi svolti in parallelo si è optato per la scrittura dei file di log sia per MongoDB - che ha questa funzionalità integrata (attivata con file di configurazione) - sia per Python attraverso il comando Unix *nohup*. I file di log sono stati analizzati per identificare la possibile presenza di dati mancanti o canzoni problematiche.

3.0.2. Struttura del Documento di MongoDB

La scelta della struttura del documento è molto importante per le performance che il database avrà poi nelle interrogazioni successive. Per questo motivo si pone l'obiettivo di trovare una struttura per il documento di MongoDB tale che la richiesta di risorse computazionali non sia troppo alta.

La struttura inizialmente ipotizzata conteneva nei campi riferiti agli stream giornalieri e alle posizioni giornaliere per ogni nazione un dizionario con la data esplicitata come chiave e il numero di visualizzazioni o la posizione in classifica come valore associato alla data corrispondente. Questa struttura avrebbe permesso query più semplici da scrivere, tuttavia avrebbe rallentato di molto i tempi di interrogazione.

La scelta finale del modello per quei campi è stata l'utilizzo di array a cui ad ogni indice corrisponde un giorno che è poi confrontato con un dizionario contenente le date effettive.

La versione finale del documento consente di risparmiare circa il 40% di spazio rispetto a quella iniziale, tuttavia abbiamo giudicato questa riduzione di dimensione in modo positivo essendo il volume nei dati utile solo se inteso come volume di informazioni. Dati pesanti perché codificati male sarebbero controproducenti per raggiungere la soglia del volume prefissata.

Data la struttura ad array di alcuni campi del documento si è deciso di utilizzare la libreria **pymongo** per svolgere le query, così da integrare il linguaggio di MongoDB con operazioni svolte per mezzo del linguaggio Python.

Viene riportata a titolo esemplificativo la struttura di un documento di MongoDB ⁵.

```
{"track_name": "God's Plan",
"artist": {"name": "Drake",
            "spotify_id": "3TVXtAsR1Inumwj472S9r4",
            "musicbrainz_id": "b49b81cc-d5b7-4bdd-aadb-385df8de69a6",
            "spotify_followers": 15962212,
            "country": "Canada",
            "gender": "male",
            "genre": ["canadian pop", "hip hop", "pop", "pop rap", "rap"],
            "related_artist"6: ["Big Sean", "J. Cole", ..., "NAV"]}
```

⁵l'esempio si riferisce alla canzone di Drake, *God's Plan*, attualmente prima in classifica Globale

⁶si riportano solo alcuni artisti correlati

```

    "first_publication": "2001",
    "image": {"height": 320,
              "url": "https://i.scdn.co/image/6bd672a0f33705eda4b543c304c21a152f393291",
              "width": 320}},
    "album": {"album_name": "Scary Hours",
              "album_publication": "2018-01-20",
              "album_image": {"height": 300,
                              "url": "https://i.scdn.co/image/2af1735e2281011bfc05353bfdee906338cbbb5b",
                              "width": 300}},
    "spotify_URL": "https://open.spotify.com/track/2XW4DbS6NddZxRPm5rMCeY",
    "features": {"acousticness": 0.0244,
                 "danceability": 0.753,
                 "energy": 0.454,
                 "loudness": -9.488,
                 "valence": 0.344},
    "popularity": 99,
    "key": 7,
    "mode": 1,
    "track_number": 1,
    "duration_ms": 198960,
    "_id": "2XW4DbS6NddZxRPm5rMCeY",
    "daily_views"7: {"United States": [0, ..., 0],
                    ...,
                    "Uruguay": [0, ..., 0]},
    "global_daily_views": [0, ..., 0],
    "daily_position"7: {"United States": [999, ..., 999],
                    ...,
                    "Uruguay": [999, ..., 999]},
    "global_daily_position": [999, ..., 999],
    "lyrics": ["Yeah", "they", "wishin", "and", ..., "br"],
    "total_streams": 0,
    "max_global_position": 0}

```

⁷si riportano la posizione e le visualizzazioni relative al 1 Gennaio e al 31 Dicembre di solo due Paesi a titolo esemplificativo. Si noti che la canzone è stata rilasciata nel 2018 e quindi non è mai comparsa in classifica nel 2017, anno oggetto dello studio

4. Esplorazione dei dati

In questa fase si è interrogato il database di MongoDB ed in particolare la collezione *songs* per ricavare informazioni relative alle variabili impiegate. Le interrogazioni sono state fatte con la libreria pymongo (si rimanda al capitolo 3.0.2 per ulteriori approfondimenti) così da poter sfruttare la potenza di MongoDB e di Python congiuntamente.

4.1. Statistiche descrittive

Si riportano le statistiche descrittive relative all'intero dataset raccolto ⁸ (Tabella 1).

Attributo	Statistiche
Volume	8.47 GB
Dimensione media documento	524 KB
Tempo di elaborazione	19 ore
N. di canzoni	16539
N. di cantanti	5165
N. Paesi	50
N. generi musicali	717
Moda della provenienza	United States

Tabella 1: Statistiche descrittive circa la collezione songs di MongoDB

La Tabella 2 contiene le statistiche descrittive relative alle canzoni che sono comparse in una classifica di uno specifico Paese. A titolo esemplificativo si riportano solo 4 dei 50 Paesi disponibili:

- Stati Uniti
- Regno Unito
- Italia
- Svezia

a cui si aggiungono nella tabella 3 le statistiche relative alle canzoni comparse almeno un giorno del 2017 nella classifica globale.

Confrontando le statistiche descrittive relative ai 4 Paesi sopra elencati si notano interessanti differenze nell'uso di Spotify da parte di persone di

⁸per il codice impiegato per le query si rimanda al file allegato spotiwhy.py

	Stati Uniti	Regno Unito	Italia	Svezia
N. di canzoni	1608	2063	1468	1682
N. di cantanti	455	817	488	723
N. di stream medi	264792	63454	22565	31399
N. di stream massimo	4068152	1357938	495971	589730
N. di stream minimo	122488	24727	5320	12090
Artista rimasto più tempo in prima posizione (no. giorni)	Post Malone (104)	Ed Sheeran (99)	J-AX (68)	Luis Fonsi (89)
Genere musicale più ricorrente	pop	pop	pop	pop
Secondo genere musicale più ricorrente	rap	dance pop	italian arena pop	dance pop

Tabella 2: Statistiche descrittive relative alle canzoni comparse nelle relative classifiche di alcuni paesi.

	Globale
N. di canzoni	1307
N. di cantanti	434
N. di stream medi	518218
N. di stream massimo	11381520
N. di stream minimo	325951
Artista rimasto più tempo in prima posizione	Post Malone (102 giorni)
Genere musicale più ricorrente	pop
Secondo genere musicale più ricorrente	rap

Tabella 3: Statistiche descrittive relative alle canzoni comparse nella classifica globale

diverse nazionalità. La prima variabile che si è considerata fa riferimento al numero di canzoni che sono comparse nel 2017 nella classifica nazionale. Si osserva che il Regno Unito ha visto susseguirsi il maggior numero di canzoni nella propria classifica, pari a 2063. Questo significa che tendenzialmente una canzone rimane nella classifica inglese per meno tempo rispetto ad esempio alla classifica italiana e si può quindi supporre che le persone inglesi amino ascoltare sempre canzoni diverse.

Al contrario negli Stati Uniti il numero di canzoni che hanno fatto parte della TOP200 di Spotify è molto inferiore a quello dell'Inghilterra, pari a 1608 canzoni di soli 455 artisti diversi. Probabilmente è più difficile entrare in competizione con gli artisti statunitensi e si ritiene che le persone negli USA ascoltino principalmente canzoni di autori già famosi.

In Svezia, sede di Spotify, si sono susseguiti numerosi artisti in classifica, mentre il numero di canzoni diverse è piuttosto limitato. Il cantante rimasto più tempo in prima posizione della classifica svedese è un artista portoricano, Luis Fonsi. Si evince che la lingua non costituisca una barriera nel campo

5. Criticità e limitazioni

Il codice di raccolta dei dati dovrebbe essere integrato con un sistema di refresh automatico del token di autenticazione delle API Spotify, cosa che richiede però procedure di autenticazione particolari riservate ai developer registrati. Il problema è stato risolto nella fase di raccolta dati attraverso un refresh manuale del token ad ogni sua scadenza, cosa che non sarebbe possibile in un'ottica di deployment in un ambiente di produzione.

Durante la creazione del database ci si è accorti che alcuni campi contenevano dati errati per via di un errore nel codice iniziale, per risolvere il problema si è utilizzata ancora una volta l'istanza EC2 con uno script per aggiornare tutti i valori in modo corretto senza ricreare il database.

Il codice elaborato è in grado di raccogliere anche le parole dei testi delle canzoni, attuando un processo di scraping dal sito AZLyrics; tuttavia non ci è stato possibile ottenere queste informazioni poiché il sito di AZLyrics, come tanti altri siti, blocca le richieste provenienti dai più popolari Cloud Services, come nel caso dell'istanza EC2 utilizzata, per evitare di incorrere in bot che scandagliano il loro database con una frequenza troppo elevata.

Per la raccolta dei dati il collo di bottiglia nell'elaborazione è stato nei tempi delle varie api di cui vengono attese tutte le risposte prima della scrittura in MongoDB. In caso di una futura rielaborazione dei dati sarebbe consigliata la divisione delle chiamate alle API in più funzioni in modo che ognuna di esse possa essere eseguita alla massima velocità consentita per poi inserire i dati nel database attraverso una procedura di update incrementale asincrono anziché la scrittura contemporanea a risposte ricevute.

Nell'ottica futura di integrazione del dataset con nuovi dati sarebbe opportuno superare la limitazione di una sola macchina sia per lo storage dei dati sia per la loro elaborazione, entrambi problemi facilmente risolvibili con l'utilizzo dello stack **Hadoop**.

6. Conclusioni

In conclusione si sono ottenute numerose informazioni riguardanti canzoni e artisti di successo in tutto il mondo. Il database così realizzato potrebbe costituire il punto di partenza ad uno studio futuro che insceni anche tecniche di machine learning per l'individuazione delle caratteristiche necessarie affinché una canzone entri nella classifica di Spotify. Le informazioni così

ottenute potrebbero essere vendute ai cantanti come regole e consigli al fine di realizzare canzoni sempre più di successo.

Alternativamente il database potrebbe essere condiviso sulla piattaforma sociale Kaggle in modo da fornire a data scientist del materiale non facilmente reperibile altrimenti.

I dati sono stati strutturati nell'ottica di realizzare delle visualizzazioni da inserire in un sito web informativo per esperti nel campo musicale e non: questo aspetto costituirà il prossimo obiettivo della ricerca.

Riferimenti bibliografici

- [1] Spotify csv, <https://spotifycharts.com/>.
- [2] Spotify api, <https://developer.spotify.com/web-api/>.
- [3] Wikipedia api, https://www.mediawiki.org/wiki/API:Main_page.
- [4] Musicbrainz, <https://musicbrainz.org/>.
- [5] Azlyrics, <https://www.azlyrics.com/>.
- [6] Geotext, <https://media.readthedocs.org/pdf/geotext/latest/geotext.pdf>.
- [7] MongoDB, <https://www.mongodb.com/it>.
- [8] Multiprocessing, <https://docs.python.org/3/library/multiprocessing.html>.
- [9] Wordcloud, https://github.com/amueller/word_cloud.