

UNO SGUARDO SULL'ITALIA DI TOKYO 2020

Corso di Laurea Magistrale in Data Science
Data Management - A.A. 2021/2022

MATTEO CAMPIRONI - 801850

LAURA CAVENATI - 864000

SERENA DI MAGGIO - 821063

ABSTRACT

A causa della pandemia di COVID-19 le Olimpiadi di Tokyo 2020 si sono svolte senza pubblico. L'unico modo quindi per seguirle dall'Italia era in diretta dalla televisione o vederle via streaming web. Un ulteriore ostacolo per i telespettatori che volevano seguire i Giochi Olimpici dal nostro paese è stata la differenza di fuso orario di 8 ore, che ha costretto gli appassionati a guardare la maggior parte degli eventi durante la notte.

In questo elaborato si spiegano i passaggi che hanno portato alla creazione di un dataset a partire dai tweet raccolti durante tutto il periodo dei giochi olimpici, arricchito con le informazioni riguardanti gli atleti, le discipline e gli eventi associati ad ogni post. Durante la fase di raccolta dati è stato utilizzato Apache Kafka e, tramite NiFi, i tweet sono stati immagazzinati inizialmente in MongoDB. Dopodiché i dati sono stati esportati in un nuovo database, Neo4j, in modo tale da poter successivamente dare la possibilità di effettuare qualsiasi tipo di analisi desiderata.

INDICE

1	Introduzione	1
2	Raccolta e storage dati	1
3	Data Quality	4
4	Data Enrichment	6
5	GraphDB	7
6	Analisi esplorativa dei dati	8
7	Criticità, conclusioni e Sviluppi Futuri	10

1 INTRODUZIONE

Le Olimpiadi moderne sono il complesso di competizioni sportive internazionali istituite nel 1896 per iniziativa del barone Pierre de Coubertin, che intendeva far rivivere lo spirito dei più famosi giochi sacri dell'antichità, celebrati in onore di Zeus a Olimpia. Sono organizzate con cadenza quadriennale dal Comitato internazionale olimpico, in città diverse e con regole precise. Oggi le Olimpiadi rappresentano l'evento sportivo più importante e più presente sui media del mondo. L'interesse che suscitano, non solo in ambito sportivo, è tale che nel 2002, per motivi logistici, il CIO ha posto un limite al numero dei partecipanti: 10.500 atleti per un totale di 28 discipline sportive e 301 specialità. Alle Olimpiadi estive si affiancano, dal 1924, anche le Olimpiadi invernali e le paralimpiadi, dedicate ad atleti disabili, organizzate a partire dal 1960. I Giochi della XXXII Olimpiade denominati dal Comitato Olimpico Internazionale come Tokyo 2020 erano inizialmente programmati dal 24 luglio al 9 agosto 2020. A causa della pandemia di COVID-19 sono stati posticipati e si sono svolti a Tokyo dal 23 luglio all'8 agosto 2021.

Il nostro elaborato ha come obiettivo la creazione di un dataset a partire dai tweet pubblicati durante tutto questo periodo e riguardanti principalmente la squadra olimpica italiana. Tale dataset, tramite l'utilizzo di tecniche di text mining, potrebbe essere impiegato per capire cosa ha influenzato il flusso di conversazioni online relative alle Olimpiadi.

La raccolta dei tweet è avvenuta tramite un'architettura producer-consumer, Apache Kafka, che consente l'interazione con le API Twitter attraverso la libreria Python Tweepy. Per la gestione dei dati in streaming è stata utilizzata l'interfaccia NiFi, che ha consentito di filtrare i campi di interesse di ogni tweet e di immagazzinarli in MongoDB. Successivamente, per poter arricchire il dataset ottenuto, è stato effettuato web scraping dal sito del CONI, reperendo informazioni relative ad eventi, discipline, atleti in gara e medaglie vinte dalla squadra italiana. Tutti i dati sono stati sottoposti ad una valutazione della qualità, per permettere così un buon processo di data enrichment. Infine, si è scelto di immagazzinare il dataset finale arricchito nel database a grafo Neo4j. In questo modo, è possibile reperire dei dati accurati e facilmente accessibili, che rappresentano una visione dettagliata e informativa dei tweet pubblicati durante i Giochi olimpici.

2 RACCOLTA E STORAGE DATI

Twitter API

Per la fase di raccolta dei dati ci siamo serviti di Apache Kafka, un'architettura producer-consumer. Il producer, permette di raccogliere i tweets in streaming e mandarli ad un topic creato appositamente dall'utente. Il consumer invece, effettuando la sottoscrizione al topic creato, riesce a recuperare in qualsiasi momento i dati e immagazzinarli. Il nostro producer, Tweepy, permette l'interazione con le API Twitter e immette nella coda i tweet rac-

colti in tempo reale. Per acquisire i dati in streaming dalla piattaforma si è implementato un workflow di Nifi contenente:

- un nodo Kafka Tweet Consumer;
- una serie di nodi per il filtraggio dei retweet;
- due nodi per la scrittura su MongoDB in locale e su Atlas.

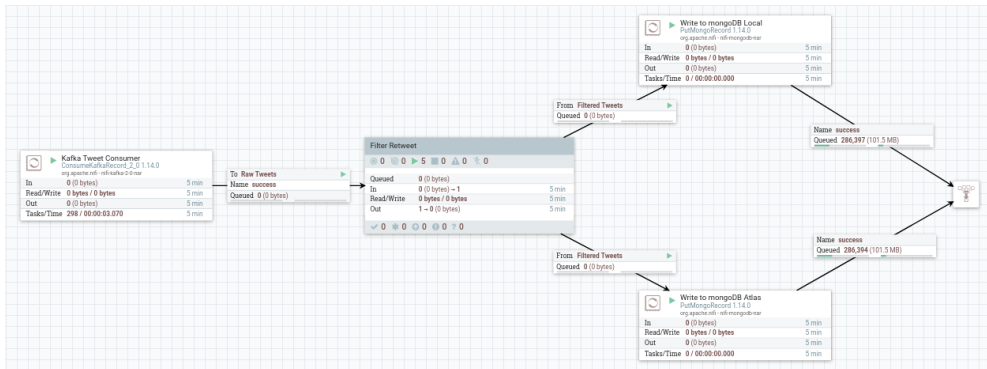


Figura 1: Workflow di NiFi

La ricerca dei tweet di interesse avviene per hashtag, ossia si cercano tutti i tweet in italiano che all’interno del loro testo presentino un hashtag inerente all’evento, in particolare: Tokyo2020, ItaliaTeam, Olimpiadi, GiochiOlimpici. Ogni tweet prima di essere immagazzinato viene filtrato tramite NiFi, andando ad eliminare i retweet in quanto considerati non utili ai nostri scopi. Infatti il nostro interesse non era studiare il social in termini di volume, ma analizzare i messaggi distinti e le discussioni degli utenti. Alcuni campi superflui vengono scartati già in fase di cattura, in quanto ci si è resi conto durante la fase di ingestion che diversi attributi erano sempre nulli e grazie alla flessibilità di MongoDB si è deciso di cambiare lo schema a raccolta avviata. Il producer è stato fatto partire alle 02:00 italiane del 23/07/2021 e fermato alle 23:00 italiane del 08/08/2021, permettendo di raccogliere in tutto 394212 tweet, dati relativi all’intera durata delle Olimpiadi, per un totale di circa 150MB. I principali vantaggi che NiFi ha fornito sono stati la gestione facilitata della queue e la possibilità di filtraggio in tempo reale dei tweet. Per quanto riguarda lo storage dei dati si è pensato di immagazzinarli in MongoDB, andando a creare una collezione diversa per ogni giornata, in modo da essere più comodi a recuperare i tweet in caso di eventuali problemi che però non ci sono stati. I dati sono stati salvati su Atlas in modo che tutti i membri del gruppo potessero averne accesso e in locale per una maggiore sicurezza.

Web scraping

Una seconda fase di acquisizione dati viene effettuata mediante lo scraping dal sito del CONI. Il portale mette a disposizione i nomi di tutti gli atleti italiani partecipanti ai Giochi olimpici e i rispettivi sport praticati. Inoltre sono presenti le informazioni riguardanti tutti gli eventi disputati dagli italiani,

andando ad elencare per ogni gara l’impianto sportivo in cui si è svolta, gli atleti che ne hanno preso parte e l’ora di inizio e fine. Infine sono elencate tutte le gare nelle quali si è vinta una medaglia e il nome dei partecipanti.

CalendarioAzzurri.csv

Per ottenere questo dataset viene utilizzata la libreria BeautifulSoup di Python, che permette di estrarre la tabella dalla pagina HTML e di importarla come DataFrame su Pandas.

Azzurri in gara 30 Luglio						STAMPA
ORA JAP	ORA ITA	SPORT	EVENTO	IMPIANTO	AZZURRI IN GARA	
07:30 - 15:30	00:30 - 08:30	Golf	Secondo giro U	Kasumigaseki Country Club	GUIDO MIGLIOZZI RENATO PARATORE	
08:30 - 11:00	01:30 - 04:00	Sport Equestri	Completo / Dressage individuale e a squadre I giorno sessione I	Equestrian Park	SUSANNA BORDONE VITTORIA PANIZZON ARIANNA SCHIVO	
09:00 - 11:50	02:00 - 04:50	Beach Volley	Fase a gironi U/D	Shiokaze Park	ADRIAN IGNACIO CARAMBULA RAURICH ENRICO ROSSI	
09:15 - 09:33	02:15 - 02:33	Canottaggio	Singolo / U Finale B	Sea Forest Waterway	GENNARO DI MAURO	
09:15 - 12:30	02:15 - 05:30	Atletica	Alto U - Qualificazioni	Olympic Stadium	STEFANO SOTTILE GIANMARCO TAMBERI	
09:30 - 11:15	02:30 - 04:15	Tiro con l'arco	Ottavi Individuale D	Yumenoshima Park Archery Field	LUCILLA BOARI	
09:30 - 12:30	02:30 - 05:30	Atletica	3000m siepi U - Qualificazioni	Olympic Stadium	AHMED ABDELWAHED OSAMA ZOGLAMI ALA ZOGLAMI	
09:45 - 12:30	02:45 - 05:30	Atletica	Disco U - Qualificazioni	Olympic Stadium	GIOVANNI FALOCI	

Figura 2: Esempio di una pagina del sito del CONI contenente le gare

Uno dei problemi a cui si va incontro è quello dei nomi multipli nel campo *Azzurri in gara*. Si sarebbe potuto utilizzare un JSON in modo da salvare gli atleti in una lista, ma la decisione presa è stata quella di inserire un atleta per riga e salvare in CSV, in quanto il dataset non è molto grande e viene usato esclusivamente per la fase di enrichment.

ListaAzzurri.csv

Sempre tramite l’utilizzo di BeautifulSoup viene creato un dataset in formato CSV contenente il nome completo di tutti gli atleti e la rispettiva disciplina.



GIOVANNI
ABAGNALE

Canottaggio

qualificato il 29/08/2019

PASS INDIVIDUALE

CLUB OLIMPICO



VINCENZO
ABBAGNALE

Canottaggio

qualificato il 02/07/2021

PASS INDIVIDUALE

CLUB OLIMPICO



AHMED
ABDELWAHED

Atletica

qualificato il 19/05/2021

PASS INDIVIDUALE

CLUB OLIMPICO




DOMENICO
ACERENZA

Nuoto

qualificato il 02/07/2021

PASS INDIVIDUALE

CLUB OLIMPICO



VLADIMIR
ACETI

Atletica

qualificato il 02/07/2021

PASS INDIVIDUALE

CLUB OLIMPICO



ALEXANDRA
AGIURGUCULESE

Ginnastica

qualificata il 05/07/2021

PASS INDIVIDUALE

CLUB OLIMPICO

Figura 3: Esempio di una pagina del sito del CONI contenente gli atleti e le rispettive discipline

Medaglie.csv

Infine, sempre utilizzando Python, viene effettuato lo scraping della pagina web che contiene gli eventi in cui l’Italia ha vinto una medaglia e i rispettivi atleti che vi hanno partecipato.

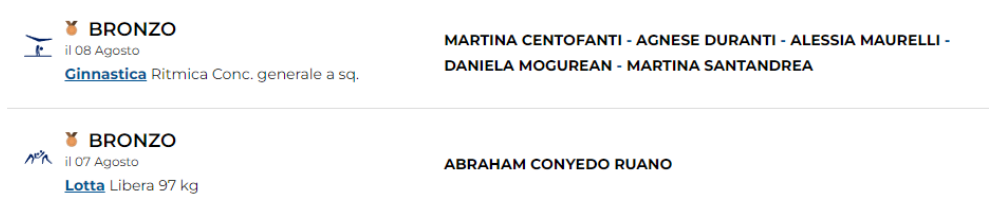


Figura 4: Esempio di una pagina del sito del CONI contenente gli eventi in cui è stata vinta una medaglia

Ricapitolando, alla fine del processo di raccolta dei dati si hanno:

- i tweet salvati su MongoDB;
- il dataset `ListaAzzurri.csv` con tutti gli atleti italiani;
- il dataset `CalendarioAzzurri.csv` con tutte le gare e l’atleta che vi ha partecipato;
- il dataset `Medaglie.csv` contenente tutti gli eventi in cui è stata vinta una medaglia e gli atleti che hanno gareggiato.

3 DATA QUALITY

Una volta ottenuti ed immagazzinati i dati, inizia un lavoro mirato ad ottenere dei Dataframe utilizzabili per rispondere alle domande di ricerca. Sfruttando PyMongo si ha facilmente accesso alle collezioni MongoDB, create durante la raccolta in streaming dei dati. I dati raccolti presentano problemi legati principalmente a due dimensioni della qualità: completezza e consistenza.

Consistenza

Questa dimensione di qualità può assumere due significati differenti:

- consistenza dei dati con i vincoli di integrità della base di dati;
- consistenza delle differenti rappresentazioni presenti nelle basi di dati di uno stesso oggetto della realtà.

Il dataset `CalendarioAzzurri.csv` presenta problemi di qualità legati al primo significato di consistenza, ovvero legato ai vincoli di integrità su tabella, definiti su due attributi, inizio e fine di un evento. Sono presenti 2 righe del dataset in cui l’orario di inizio e di fine dell’evento sono invertiti e 11 record in cui la data di inizio dell’evento è posticipata di un giorno rispetto alla realtà. Quest’ultima inconsistenza è causata da come il sito del CONI stesso gestisce alcuni dati. Se ad esempio un evento si è svolto tra le 23.30 del 25/07/2021 e le 2:00 del 26/07/2021 il sito lo riporta come evento totalmente avvenuto il 26/07/2021. Tuttavia, questo non corrisponde alla realtà e i dati risultano incoerenti tra loro. Entrambi questi problemi, sia legati agli orari invertiti sia legati al giorno anticipato, sono stati risolti modificando manualmente i valori inconsistenti. La collezione di tweet ottenuta in streaming

presenta invece un problema di consistenza rappresentativa, fondamentale per la futura integrabilità dei dati. Ci si è accorti, infatti, che il campo timestamp, in cui vengono riportati orario e data di un tweet, presenta l'orario in formato UTC, il quale è indietro di due ore rispetto all'orario italiano. Per questo motivo, questo campo è stato trasformato in formato UTC+2, ossia l'ora italiana, come è riportata nel dataset ottenuto dal sito del CONI. Un altro miglioramento della qualità indispensabile per l'integrabilità dei dati è stato quello di riscrivere i nomi degli sport nei dataset ottenuti tramite scraping nello stesso modo, ossia in lettere minuscole, fatta eccezione la prima lettera della prima parola. Per analoghe motivazioni, si è aggiunto in una riga di `ListaAzzurri.csv` il secondo nome all'atleta BENEDICTA CHIGBOLU, diventando quindi MARIA BENEDICTA CHIGBOLU, come è riportata nelle restanti righe a lei riferite. In questo modo, si crea uniformità tra i nomi di tutti gli atleti presenti in entrambi i dataset.

Completezza

La completezza di un insieme di dati è la copertura con cui il fenomeno che è stato preso in considerazione, è rappresentato nell'insieme dei dati. Durante la raccolta dei tweet, alcuni dei campi considerati sono: geo, coordinates e user_location, cioè geolocalizzazione, coordinate e nome del luogo in cui si trova l'utente mentre pubblica il tweet. Nell'intera collezione però sono stati contati 394134 valori nulli sia per il primo che per il secondo campo e 151901 valori mancanti o nulli per il terzo, corrispondenti rispettivamente al 99% e 39% del totale dei tweet raccolti. Purtroppo si tratta di informazioni non reperibili e, per questo motivo, non è possibile sfruttare questi campi per effettuare enrichment o per ulteriori analisi e utilizzi. Anche il dataset `CalendarioAzzurri.csv` ottenuto attraverso scraping è incompleto. Infatti, si valuta la completezza dell'attributo "Azzurri in gara", che riporta i nomi degli atleti italiani che partecipano ad un determinato evento. Questo attributo presenta 53 valori mancanti, quasi l'11% del totale nel dataset originario. Per migliorare la qualità dei dati si procede con una strategia data-driven e, in particolare, si utilizza la tecnica source trustworthiness, che consiste nel selezionare una fonte di dati che si ritiene di buona qualità per sostituire i valori mancanti. In questo caso, la fonte è il sito ESPN [1]. La complessità del sito web ha reso complicato automatizzare il processo, per cui si sono cercati gli eventi mancanti uno per volta e si sono individuati gli atleti coinvolti. Non tutti i campi mancanti sono stati riempiti, infatti 28 eventi facevano riferimento a gare dove non ha partecipato nessun atleta italiano. Il motivo della loro presenza è dovuto al fatto che la tabella venisse caricata la sera prima degli eventi e alcuni di questi prevedevano qualificazioni e fasi finali tutti nella stessa giornata e di conseguenza era impossibile sapere chi vi avrebbe partecipato. Tali record sono stati rimossi.

Infine, oltre a questi problemi legati a completezza e consistenza, ne è stato individuato un altro legato all'accuratezza sintattica: `CalendarioAzzurri.csv` presenta nell'attributo evento una cella con la stringa "Batterie StEffetta 4x200m stile libero D", corretta manualmente in "Batterie StAffetta 4x200m stile libero D".

4 DATA ENRICHMENT

Ottenuti tutti i dati preliminari, come descritto nelle precedenti sezioni, è stato necessario definire uno schema per l'arricchimento dei dati ottenuti mediante API con i dati ottenuti tramite web scraping. In particolare si è cercato di assegnare ai tweet gli atleti, gli sport e l'evento a cui si riferivano. In questo modo i tweet con contenuto relativo ad un certo atleta piuttosto che ad uno sport o ad un evento presenteranno questa ulteriore informazione. La decisione di come assegnare ad un tweet queste informazioni non è stata semplice. Non è facile lavorare con il linguaggio naturale e per questo motivo sono state selezionate le seguenti accortezze:

- Per arricchire il tweet con gli sport a cui sono riferiti, abbiamo utilizzato le discipline olimpiche estratte dal dataset `CalendarioAzzurri`. È stato scansionato il testo dei tweet per verificare se contenesse i nomi degli sport. A supporto di questo processo è stato creato un dizionario che contiene come chiavi i vari sport e per valori alcuni sinonimi e parole identificative dei vari sport, in modo da cogliere le varie sfumature del linguaggio italiano.
- Per assegnare il tweet ad un certo atleta invece si è cercato se il testo contenesse il cognome (solo se non c'erano più atleti con lo stesso cognome) o almeno un nome e il cognome (alcuni atleti avevano più di un nome).
- Per assegnare gli eventi invece si è ragionato su come si potesse fare un'integrazione temporale. Il fatto che un tweet sia stato postato durante un determinato evento non è sufficiente per affermare che stia parlando proprio di quello. Per questo motivo viene verificato anche se il testo contenga il nome di uno degli atleti in gara. La combinazione di queste condizioni ci permette di dedurre che un tweet sia riferito ad un certo evento.

Utilizzando queste regole si è riusciti ad assegnare un atleta a circa il 37% dei tweet totali, una disciplina al 35% e un evento all'11%, sempre tenendo conto che in generale non tutti i tweet parlano di prestazioni olimpiche, ma molti erano legati al fuso orario o ai problemi di trasmissione della Rai.

Record Linkage

Una procedura di record linkage è una tecnica algoritmica il cui scopo è identificare quali coppie di record di due basi di dati corrispondono ad una stessa unità. In questo caso particolare ci si è resi conto che il dataset `Medaglie.csv` presentava i nomi degli eventi scritti in modo più approssimativo rispetto a quanto riportato nel dataset `CalendarioAzzurri.csv`. Per questo motivo è stato effettuato un lavoro di record linkage. Il primo passo è stata la selezione delle variabili chiave, che abbiamo individuato nel nome dell'evento e negli atleti che gareggiavano. Questi ultimi sono stati normalizzati in modo che fossero espressi nello stesso formato in entrambi i dataset, così da rendere più semplice individuare le differenze. Date le dimensioni ridotte dei

CSV non è stata effettuata nessuna operazione di blocking. Utilizzando la libreria di Python fuzzymatcher sono stati individuati correttamente tutti gli eventi e sostituite le versioni approssimative del testo con quelle complete.

Sentiment Analysis

Come ultima attività di arricchimento si è deciso di effettuare una sentiment analysis dei tweet. In questo modo viene aggiunta un'informazione riguardante il riconoscimento delle emozioni, che può essere utile per dei task di opinion mining. A questo scopo è stato utilizzato il modello FEEL-IT [2], addestrato appositamente per questo scopo su tweet italiani. Il risultato è 0 se il tweet è negativo, mentre 1 se considerato positivo.

5 GRAPHDB

Per lo storage del dataset finale si è deciso di immagazzinare i dati in un database a grafo. La scelta è ricaduta su Neo4j ed è dovuta alla struttura dei dati, che ben si prestano ad essere schematizzati tramite nodi, proprietà e relazioni che li interconnettono. In questo modo abbiamo voluto dare maggiore importanza alle relazioni tra i nodi. L'impossibilità di un'eventuale scalabilità orizzontale non è stata vista come un problema in quanto si è pensato che difficilmente il database potrebbe essere ingrandito a tal punto da mettere in crisi Neo4j, dal momento che i Giochi olimpici sono terminati e non è previsto un ulteriore streaming di tweet. Utilizzare un DB document-based nel nostro caso avrebbe sicuramente portato ad una ripetizione dei dati, situazione da evitare per non insorgere in errori soprattutto per eventuali aggiornamenti. Ad esempio, se si fossero divisi i documenti in base agli atleti, come ci si sarebbe dovuti comportare con i tweet senza atleta? Dove si sarebbero inserite le informazioni relative ai singoli eventi? Sarebbe stato necessario ripeterle per ogni tweet. Contrariamente, se si fosse impostato il DB in base agli eventi, non si sarebbero potuti considerare i tweet a cui non è associato un evento. Viene quindi effettuato l'import dei dati in Neo4j sfruttando il linguaggio Cypher, l'ottimizzazione dell'import via CSV e l'utilizzo della libreria APOC per i file JSON.

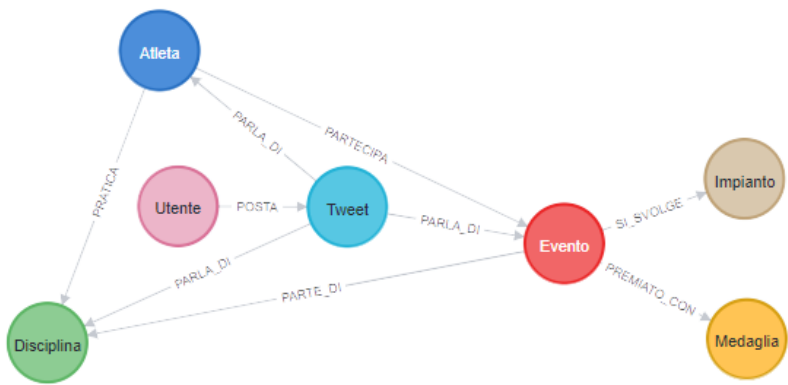


Figura 5: Modello implementato su Neo4j

6 ANALISI ESPLORATIVA DEI DATI

Una prima analisi esplorativa è stata svolta subito dopo la raccolta dei dati. Essendo il dataset composto prevalentemente da variabili di tipo qualitativo non è stato possibile ottenere statistiche di base sulla distribuzione dei dati come media o varianza. Si è scelto quindi di investigare su quali fossero le varie problematiche, come per esempio il numero di valori nulli dei vari attributi oppure esplorare il numero di tweet arricchiti con successo. Tutti questi numeri sono stati riportati nelle sezioni precedenti e hanno permesso di migliorare la qualità dei dati.

Ottenuto il dataset finale è stato possibile svolgere alcune analisi esplorative attraverso delle visualizzazioni. Tale processo ha consentito di approfondire e rendere maggiormente esplicativi i dati raccolti.

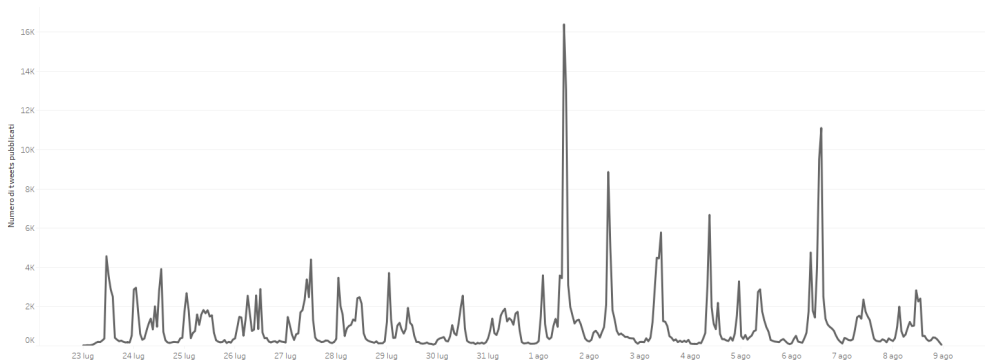


Figura 6: Numero di tweet pubblicati

Come mostrato dal grafico in Figura 6, il numero di tweet pubblicati su Twitter presenta una stagionalità giornaliera. Il 1 agosto verso mezzogiorno, momento della vittoria di Jacobs nei 100m e di Gianmarco Tamberi nel salto in alto, è stato registrato il maggior numero di tweet pubblicati.

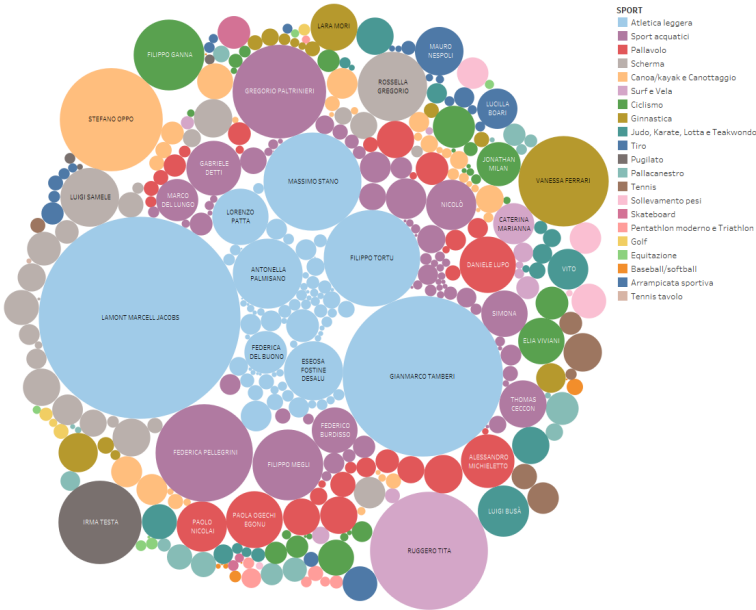


Figura 7: Atleti più citati nei tweet

Come mostra il grafico in Figura 7, Lamont Marcell Jacobs è l'atleta più citato, seguito da Gianmarco Tamperi e Ruggero Tita, tutti vincitori di medaglie d'oro a Tokyo 2020.



Figura 8: Parole più frequenti nei tweet

Anche la Figura 8 mostra come gli atleti Lamont Marcell Jacobs e Gianmarco Tamberi siano notevolmente citati nei tweet. Questo grafico è una wordcloud a forma di fiamma olimpica, che riporta le parole più frequenti all'interno dei tweet, escludendo i tag considerati per raccogliarli e stopwords. Vengono evidenziate parole come FINALE, ATHLETIC, VOLLEYBALL, NUOTO, SCHERMA, MONDO, BRONZO, ORO, la cui frequenza testimonia come siano stati maggiormente commentati le finali o sport come nuoto, pallavolo, atletica e scherma, molto seguiti solitamente o in cui la squadra italiana ha vinto medaglie.

Altre visualizzazioni, preparate per il progetto di data visualization, sono reperibili al seguente link [3].

7 CRITICITÀ, CONCLUSIONI E SVILUPPI FUTURI

Un possibile sviluppo futuro per questo progetto potrebbe essere quello di riuscire a superare il limite di tweets raccolti imposto dalle API di Twitter, al fine di riuscire ad effettuare analisi ancora più dettagliate e approfondite, soprattutto nei momenti in cui si riscontrano vittorie importanti. In futuro potrebbe essere utile arricchire il dataset con ulteriori informazioni relative all'argomento di quei tweet che non si riferiscono né ad atleti né a sport. Da una prima analisi esplorativa dei dati ci si è accorti che molti tweet riportavano un sentiment negativo, in quanto gli utenti si lamentavano del fatto che alcune gare non erano trasmesse in diretta mentre altre, a causa del fuso orario, bisognava seguirle a notte inoltrata. Un'analisi dettagliata del testo dei tweet potrebbe aiutare i canali di trasmissione a programmare la trasmissione televisiva e radiofonica delle prossime Olimpiadi, stabilendo quali eventi trasmettere, a quali sport e atleti dare più spazio ed eventualmente quali gare replicare.

RIFERIMENTI BIBLIOGRAFICI

- [1] Espn - tokyo2020. [Online]. Available: <https://www.espn.com/olympics/summer/2020/schedule>
- [2] F. Bianchi, D. Nozza, and D. Hovy, "'FEEL-IT: Emotion and Sentiment Classification for the Italian Language'," in *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2021.
- [3] Link progetto data visualization. [Online]. Available: <https://mcampironi.github.io/>