Università degli studi di
Milano-Bicocca

Data Management e Visualization

Progetto finale

Toronto Raptors vs Golden State Warriors: Analisi dei tweet pubblicati durante l'ultimo game della finale di NBA 2019

Authors:
Beatrice Civelli - 754920 - b.civelli@campus.unimib.it
Luca Contini - 836675 - l.contini@campus.unimib.it
Edoardo Gervasoni - 790544 - e.gervasoni4@campus.unimib.it

July 8, 2019

Contents

Abstract

Il Social Listening, anche conosciuto come "monitoraggio dei social media", è il processo grazie al quale è possibile l'individuazione e la valutazione delle conversazioni in rete attinenti particolari parole chiave o frasi che identificano determinati brand, eventi, individui e qualsiasi altro argomento d'interesse. Possiamo pensare al social listening come a un processo continuo e costante che ha la finalità di analizzare i comportamenti delle persone e consente di elaborare strategie orientate all'ascolto delle informazioni postate su internet. In una realtà in cui chiunque è in rete, che sia un individuo o un'azienda, le informazioni postate sul Web producono enormi quantità di dati che possono essere raccolte in tempo reale mediante tecniche e strumenti di data streaming. Con tali dati è possibile rispondere ad una serie di domande importanti e scoprire fatti interessanti che permettono di capire il contesto dell'ascolto ed il relativo comportamento degli utenti online. In questo report presenteremo il nostro progetto di social listening effettuato su Twitter durante l'ultimo game degli NBA Playoff 2019 e di come le informazioni raccolte sono state elaborate per permetterci di scoprire fenomeni interessanti avvenuti in contemporanea con la partita.

1 Introduzione

Il 14 Giugno 2019 si è giocato il sesto e ultimo game della finale di NBA presso l'Oracle Arena (Oakland, California) che si è concluso con la vittoria dei Toronto Raptors contro i Golden State Warriors con un punteggio di 114 a 110. La National Basketball Association, comunemente nota come NBA, è la principale lega professionistica di pallacanestro degli Stati Uniti d'America e del Canada. Da anni le partite dell'NBA, ricche di competizione e di azioni spettacolari da parte dei giocatori, sono seguite da un gran numero di persone, soprattutto in America, e sono conosciute in tutto il mondo. L'obiettivo del nostro progetto è stato quello di raccogliere i dati relativi al traffico di tweet avvenuto durante la partita, effettuando quindi un social listening di

Twitter.

Twitter è una delle principali piattaforme social al mondo che offre un servizio web di condivisione delle notizie e di microblogging. Su Twitter gli utenti postano e interagiscono con messaggi chiamati Tweet e in ogni istante milioni di persone da tutto il mondo condividono la loro opinione riguardo agli argomenti più disparati, andando a formare uno dei più grandi database da cui poter estrapolare qualsiasi tipo di informazione. Con i dati che siamo riusciti a raccogliere ci è stato possibile individuare e valutare il flusso di informazioni avvenuto durante l'evento ma anche di scoprire fatti interessanti che abbiamo visualizzato mediante l'utilizzo di infografiche. Nello svolgimento del progetto abbiamo cercato di affrontare due delle V dei Big Data: la *Velocity*, ovvero la capacità di raccogliere dati in tempo reale, e la *Variety*, cioè la capacità di integrare, analizzare e ricavare informazioni da dati provenienti da fonti diverse.

2 Velocity

2.1 Streaming dei dati

Lo streaming è una procedura con la quale avviene il trasferimento continuo e progressivo di flussi di informazione digitale consentendo di ottenere dati in tempo reale. La parola deriva dal significato del termine inglese *to stream* , ovvero fluire, scorrere. Per il progetto abbiamo scritto il codice di streaming con Python utilizzando in particolare Tweepy[1], una libreria che permette di fare richieste alle Web API di Twitter in maniera molto semplice. Le API (Application Programming Interfaces) rappresentano il mezzo con cui i programmi informatici "parlano" tra di loro per richiedere e fornire informazioni. Permettono sia di accedere a determinate funzioni o dati di un programma o di un servizio web, sia di manipolarli e utilizzarli per diversi scopi.
Twitter, come altre piattaforme social, utilizza l'autenticazione OAuth. Si tratta di uno dei più importanti protocolli del web 2.0 che permette di

[1] https://www.tweepy.org/

accedere a specifiche risorse protette da autenticazione (username e password) senza necessità di dover condividere pubblicamente tali credenziali. L'autenticazione OAuth, in altri termini, permette a una terza parte di gestire sezioni riservate di una risorsa, di un sito o di un'applicazione, il tutto senza costringere l'utente a mettere in circolazione la propria password per accedervi.

Come primo passo abbiamo creato un'applicazione presso il seguente indirizzo https://apps.twitter.com/, con la quale abbiamo ottenuto quattro credenziali di accesso che ci hanno permesso di reperire le informazioni e i dati di nostro interesse:

- *consumer key*: chiave che identifica il client che accede alle risorse.
- *consumer secret*: password del client che viene utilizzata per l'autenticazione con il server di Twitter.
- *access token*: token di accesso che viene rilasciato al client dopo l'autenticazione. Questo token permette di definire i privilegi del client, ovvero a quali dati può e non può accedere.
- *access token secret*: ogni volta che il client desidera accedere alla risorsa, il token secret funziona come una password e viene rilasciato con il token di accesso.

Una volta inserite le nostre credenziali nello script di Python, abbiamo avviato la funzione Tweepy.StreamListener per raccogliere i tweet in tempo reale. In particolare abbiamo scelto di selezionare solo quelli contenenti cinque hashtags ufficiali per la partita: **#NBA, #NbaFinals, #Nbaplayoffs, #TorontoRaptors** e **#GoldenStateWarriors**. Sono state registrate sia le informazioni riguardanti ogni tweet (testo e orario di pubblicazione in UTC, Coordinated Universal Time) sia quelle riguardanti gli utenti (nome, nome dell'account twitter, numero di followers e numero di amici). Lo streaming è stato avviato il 14 giugno 2019 alle 00:45 (orario Italiano) e si è concluso alle 06:31 con la raccolta di 429.500 tweet complessivi.

2.2 MongoDB

MongoDB è uno dei più noti DBMS non relazionali o NoSQL. Si tratta di una soluzione orientata ai documenti che sfrutta il formato JSON per la memorizzazione e la rappresentazione dei dati. A differenza di un database relazionale, i dati vengono salvati all'interno di una *collection* composta da diversi documenti.
Per poter lavorare con MongoDB durante la fase di collezione dei dati, abbiamo utilizzato la libreria *Pymongo*[2] implementabile in Python. Grazie all'utilizzo di questa, i dati sono stati salvati in tempo reale in un file JSON, all'interno di una collezione creata appositamente per lo streaming, racchiusa a sua volta in un database chiamato TwitterDB. Finita la fase di streaming, con il comando mongoexport abbiamo esportato i dati dal database di MongoDB in un formato CSV per poi procedere alla fase di elaborazione e analisi delle informazioni.

3 Variety

3.1 Dataset principale

I tweet raccolti sono stati analizzati grazie a Jupyter Notebook, un ambiente computazionale interattivo per scrivere ed eseguire codici Python. Il dataset, caricato come Pandas DataFrame, presenta una struttura di 429.500 righe e 8 colonne. Ogni riga rappresenta un tweet, mentre le colonne rappresentano le seguenti variabili:

- *_id*: un identificatore univoco di ciascuna riga
- *created_at*: l'orario di pubblicazione del tweet in UTC
- *text*: testo del tweet
- *user.name*: il nome dell'utente

[2] https://api.mongodb.com/python/current/

- *user.screen_name*: il nome dell'account Twitter dell'utente
- *user.location*: la località
- *user.followers_count*: numero dei follower dell'utente
- *user.friends_count*: numero degli amici dell'utente

Gli attributi di maggiore interesse per lo sviluppo del progetto sono 'created_at' e 'text'; la variabile location, per quanto possa essere interessante individuare la posizione geografica degli utenti, presenta numerosi valori NaN e molte location fittizie, rendendola quindi inutilizzabile. User.follower_count e user.friends_count, invece, non presentano informazioni rilevanti per lo scopo del progetto e pertanto sono state ignorate.

3.2 Integrazione di altri Dataset

Con l'obiettivo di approfondire l'analisi dei tweet raccolti, si è deciso di integrare il dataset principale con altri dati ricavati dal sito www.basketball-reference.com, dove sono presenti numerose informazioni sull'andamento della partita del 14 Giugno, come la timeline degli eventi del gioco, statistiche sulle performance dei giocatori di entrambe le squadre e diversi grafici. I dati che abbiamo considerato per lo scopo del progetto sono stati il resoconto delle azioni del gioco (ciascuna col corrispondente minuto della partita) presente all'indirizzo https://www.basketball-reference.com/boxscores/pbp/201906130GSW.html, e diversi dataset sulle statistiche dei giocatori, presenti alla pagina https://www.basketball-reference.com/boxscores/201906130GSW.html. Questi dati sono stati esportati come file .csv e caricati come Pandas DataFrame nello script Python.

Il file della timeline degli eventi della partita inizialmente caricato presenta 6 colonne, relative alle seguenti variabili:

- *Time*: il minuto dell'azione al dato quarto della partita. Questa è infatti suddivisa in 4 parti da 12 minuti ciascuno, e per ogni quarto il conteggio dei minuti va da 12 a 0
- *Toronto*: eventi relativi ai Toronto Raptors
- *TorontoScore*: valore dei canestri realizzati dai Raptors (1, 2 o 3 punti)

- *Score*: punteggio complessivo alla determinata azione
- *GoldenState*: eventi relativi ai Golden State Warriors
- *GoldenStateScore*: valore dei canestri realizzati dai Warriors

Gli eventi riportati possono essere un canestro, un fallo o un'altra azione rilevante del gioco.

Uno dei problemi principali nello svolgimento del progetto è stato quello di integrare il dataset della timeline della partita con quello dei tweet che abbiamo raccolto, data la differenza tra le scale temporali. Infatti, nonostante il primo sia molto dettagliato nel riportare ogni tipo di informazione su ciò che è accaduto in gioco per entrambe le squadre, l'attributo *Time* riporta solo i minuti della partita e non l'orario assoluto. Per rendere quindi possibile l'integrazione delle informazioni tra le due fonti, si è scelto di selezionare solo quelle azioni della partita corrispondenti ad un evento di cui si potesse sapere l'orario con sicurezza. In particolare abbiamo considerato i tweet delle pagine ufficiali di NBA e NBAItalia riferiti a precise azioni del gioco, e al loro orario di pubblicazione. In tal modo siamo riusciti a collocare alcuni eventi rilevanti della partita in scala UTC, con un errore stimato di circa 1-2 minuti dato un possibile ritardo della la pubblicazione dei tweet.

Per quanto riguarda gli altri dataset utilizzati, questi sono 4 e riportano le statistiche generali (Basic Box Score Stats) ed avanzate (Advanced Box Score Stats) dei giocatori per le due squadre. I parametri di particolare interesse sono: *MP*, i minuti giocati durante la partita, *True Shooting Percentage* (TS%), una stima che misura l'efficienza di un giocatore nei tiri; *Effective Field Goal Percentage* (eFG%), una statistica che descrive l'efficienza nella realizzazione dei canestri; *+/-*, una statistica che misura l'impatto di un giocatore sul gioco, rappresentata dalla differenza tra il punteggio totale della propria squadra rispetto a quello dell'avversario quando il giocatore è in gioco.

4 Preprocessing

La fase di preprocessing comprende la pulitura e la strutturazione dei dataset per consentire una migliore analisi dei dati. Considerando il dataset principale, il primo problema affrontato è consistito nella rimozione delle righe contenenti valori NaN per facilitare e alleggerire l'analisi dei tweet. Quindi si è proceduto ad aggregare per minuti i tweet raccolti e a rappresentarne l'andamento:

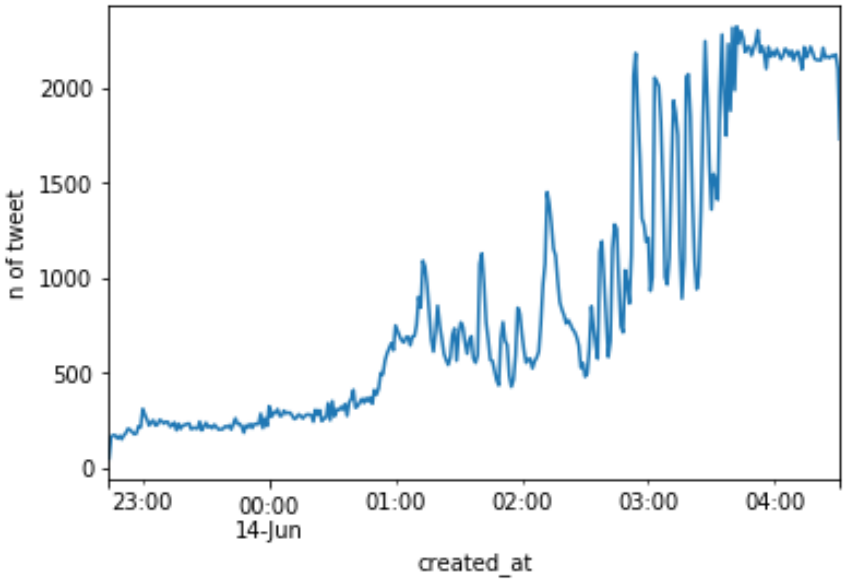

Figure 1: Tweet totali al minuto

Com'è possibile notare dal grafico, fino all'ora 1:00 UTC il numero di tweet al minuto resta basso, dai 250 ai 500; dall'inizio della partita (1:11) si assiste ad un loro aumento, caratterizzato da numerose fluttuazioni. Intorno alle 3:45 UTC si può invece notare un andamento particolare: il numero di tweet satura e non presenta più oscillazioni evidenti.
Abbiamo ipotizzato che questo sia dovuto ad un limite dell'API di Twitter per cui non è possibile raccogliere oltre un certo numero di tweet al minuto. Ai fini dell'analisi e dello scopo del progetto è stato pertanto deciso di utilizzare solo i dati fino alle 03:42 del 14 Giugno 2019 (UTC), quando la partita è ormai finita. Inoltre è stato anche deciso di escludere i tweet postati prima

delle 00:45 perché ritenuti poco interessanti per il nostro fine.

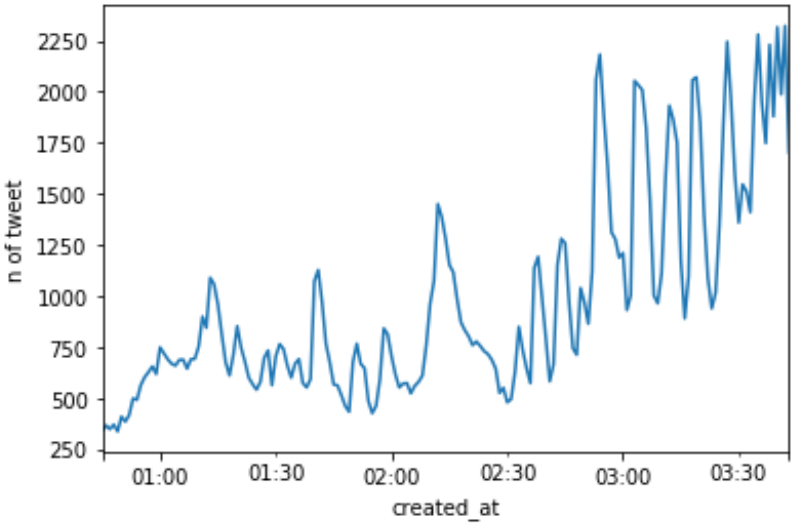

Figure 2: Tweet uscenti dalla fase di preprocessing

Un'ulteriore manipolazione dei dati, descritta successivamente, è stata effettuata per la realizzazione della terza infografica.

Per quanto riguarda i dataset delle statistiche, abbiamo escluso le osservazioni riguardanti i giocatori che hanno giocato per meno di 25 minuti di partita. Per il resto non hanno subito modifiche dato che sono stati ricavati già strutturati.

5 Visualizzazione delle informazioni raccolte

5.1 I infografica

La prima infografica presenta i dati che abbiamo raccolto da un punto di vista generale, mostrando i tweet del nostro dataset rispetto al tempo UTC in cui sono stati postati. Nel grafico sono state evidenziate le varie suddivisioni della partita tra i quarti di gioco (in blu) e le pause, l'intervallo e

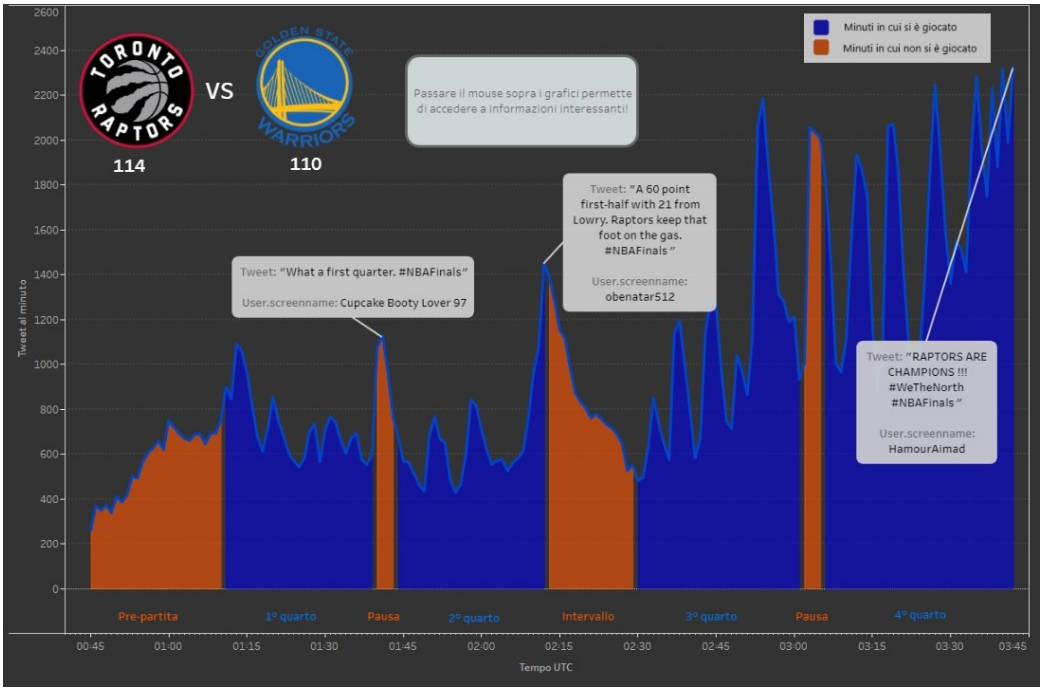

Figure 3: Prima infografica. Versione interattiva su https://tinyurl.com/y559nhhd

il pre-partita (in rosso). Andando al link nella descrizione dell'immagine si trova la versione interattiva dell'infografica, e scorrendo con il mouse sulla figura è possibile accedere all'ora, al numero di tweet e agli eventi della finale che sono stati selezionati e ricollocati in scala temporale UTC, come spiegato nella sezione 2.2. Sono stati inoltre riportati tre tweet pubblicati durante la partita considerati particolarmente significativi.

Dall'infografica si nota che, come già sottolineato precedentemente, i tweet tendono in generale ad aumentare col procedere della partita. Più in particolare si può osservare come, in corrispondenza delle pause e quindi dell'interruzione del gioco, vi sia un aumento dei tweet al minuto. Ciò è probabilmente dovuto al fatto che durante le pause il pubblico non segue la partita e puo' dedicare la sua attenzione ai social network. Un andamento simile si può notare all'inizio dell'intervallo tra le due metà della finale; il fatto che nel corso di questo si registri una diminuzione dei tweet, si può spiegare con la presenza di uno spettacolo musicale d'intermezzo.

5.2 II infografica

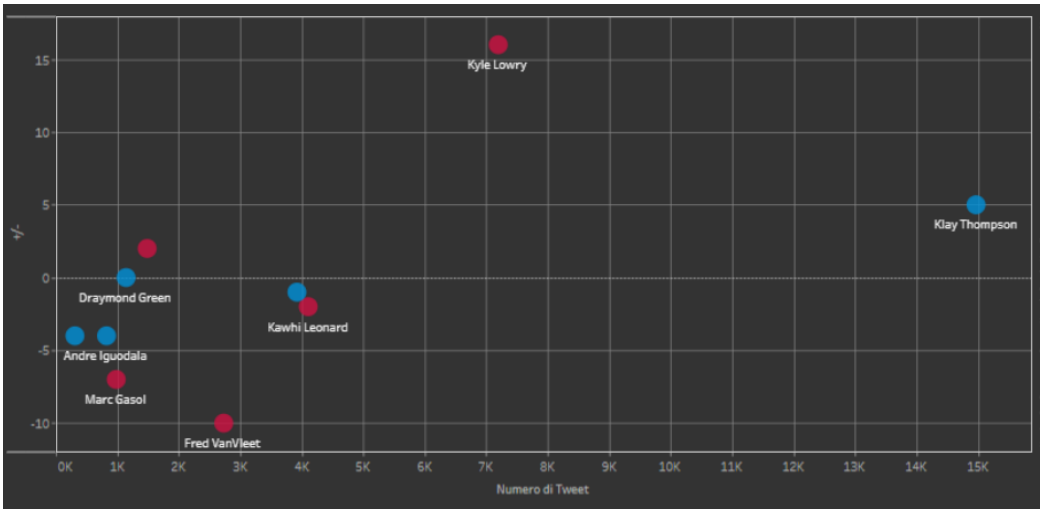

Figure 4: Primo grafico della seconda infografica. Versione interattiva su https://tinyurl.com/y4ry7jla

Nella seconda infografica ci siamo concentrati sulle relazioni tra i tweet e vari giocatori. Tra questi ultimi in particolare abbiamo considerato solo quelli che hanno giocato per più di 25 minuti di partita. Quello che abbiamo voluto mostrare nel primo grafico, è la presenza di una correlazione tra le performance in campo dei giocatori e il numero di tweet in cui il loro nome è comparso. I dati sulla bravura di ciascun uomo sono stati ricavati dai dataset precedentemente descritti, in particolare dagli *Advanced Box Score Stats* per ciascuna delle due squadre. Le statistiche di nostro interesse sono state la True Shooting Percentage, l'Effective Field Goal Percentage e +/-, già descritte sopra. Per quanto riguarda invece i tweet, per ogni giocatore è stato calcolato il numero di post in cui compare il proprio cognome. Eccezioni sono Klay Thompson e Kahwi Leonard per i quali è stato considerato il nome, dato si è visto che venivano citati con questi nella maggior parte dei casi; per Draymond Green invece si è scelto di considerare i tweet in cui comparivano sia nome che cognome data la presenza di un omonimo nell'altra squadra.
Attraverso uno slider a destra del grafico, l'utente può vedere come varia il rapporto tra barvura in gioco di ciascun giocatore e numero di tweet: si potrà notare che i due giocatori migliori rispetto agli indicatori considerati, Klay

Thompson per i Golden State Warriors e Kyle Lowry per i Toronto Raptors, sono anche i più twittati.

Nella seconda parte dell'infografica abbiamo quindi deciso di mostrare come varia nel tempo il flusso dei tweet relativi a questi due campioni. Come si

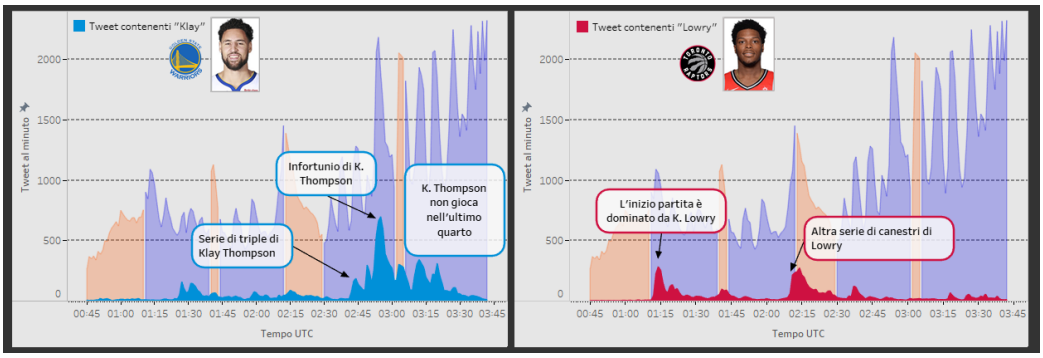

Figure 5: Thompson vs Lowry

può notare dall'immagine, i tweet al minuto riguardanti Klay Thompson rimangono su valori abbastanza bassi fino al terzo quarto. Qui il giocatore inizia a realizzare diversi canestri da 3 punti, a cui corrisponde un aumento del numero di tweet, che raggiunge un picco netto in corrispondenza del suo infortunio. Il flusso di post riguardanti Thompson continua a mantenere un valore rilevante anche durante la successiva pausa e nel corso dell'ultimo quarto, nonostate non stia giocando. Una spiegazione per questo si può trovare nello scalpore che l'infortunio del campione ha provocato sul pubblico, anche in relazione all'incidente in campo subito da un'altra superstar dei Warriors nella partita precedente: Kevin Durant.

Per quanto riguarda invece Kyle Lowry, si può notare subito un picco consistente di tweet intorno all'1:15 UTC: fin dall'inizio della partita Lowry infatti è protagonista di varie azioni vincenti tra cui due triple. Un secondo picco rilevante si ha alla fine del secondo quarto, anche questo corrispondente ad una sua serie di punti.
Notiamo infine che per entrambi i giocatori si assiste ad un aumento dei tweet relativi durante le pause: in particolare per Thompson questo si vede nella prima e nell'ultima pausa; mentre per Lowry si vede nella prima e nell'inizio dell'intervallo.

5.3 III infografica

Figure 6: Wordcloud della terza infografica. Versione interattiva su https://tinyurl.com/y5rbkx6k

Nella terza infografica abbiamo infine voluto mostrare le parole che sono comparse con più frequenza nei tweet raccolti. Questo è stato fatto con la realizzazione di un Wordcloud, per mostrare visivamente i termini con maggiori occorrenze, e con un lollipop chart in cui si è mostrato il numero delle volte con cui i 10 termini più frequenti comparivano. Per poter realizzare questi ultimi grafici, è stata necessaria un'ulteriore fase di manipolazione dei

dati. Grazie alla funzione *.clean()* della libreria python *preprocessor*[3] è stato possibile rimuovere dal testo dei tweet tutti i caratteri "speciali" quali hashtags, URL, links, etc. Successivamente si è scelto di filtrare tutte le parole composte da una o due lettere e i più comuni articoli, pronomi e preposizioni dato che sono stati considerati irrilevanti al fine della nostra indagine.

Il Wordcloud è stato realizzato grazie alla libreria di python *WordCloud*[4]. Come si nota da entrambi i grafici la parola nettamente più twittata è NBA, mentre per quanto riguarda i nomi dei giocatori, i più frquenti sono Klay Thompson (che ha sia nome che cognome tra i termini più ricorrenti) seguito da Lowry. È infine interessante notare come i Raptors abbiano sopravanzato i Warriors non solo in partita ma anche in ricorrenze tra i tweet che sono stati raccolti.

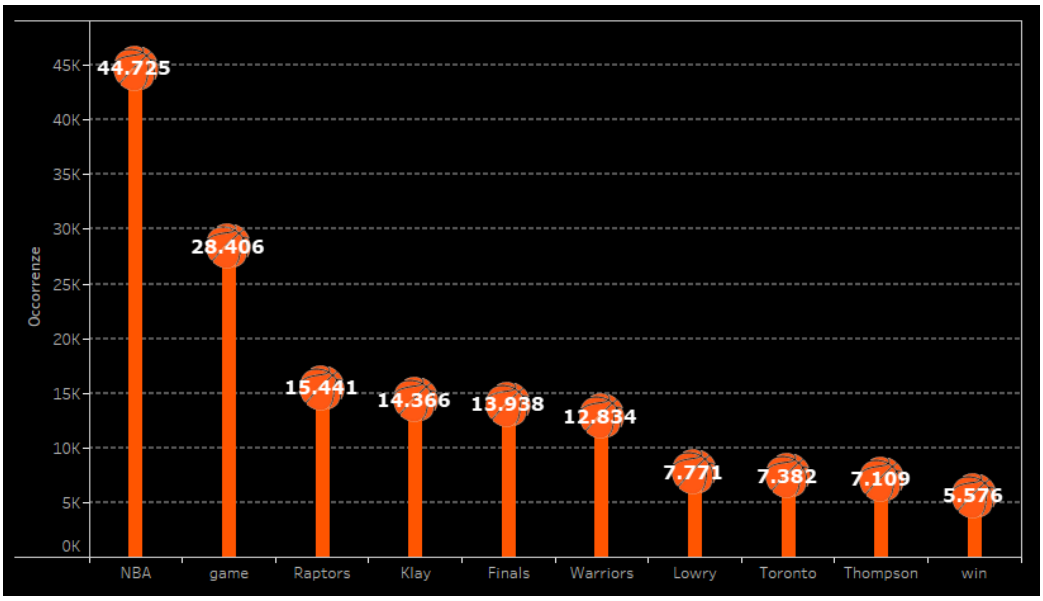

Figure 7: Lollipop chart delle occorrenze delle parole più frequenti.

[3] https://pypi.org/project/tweet-preprocessor/

[4] https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html

6 Conclusioni

Il lavoro che abbiamo fatto è consistito nel raccogliere e nel descrivere i tweet (contenenti determinati hashtags) che sono stati pubblicati durante il sesto game della finale di NBA 2019. Sfruttando Tweepy è stato possibile recuperare i post in tempo reale, affrontando quindi il problema della *Velocity* dei Big Data. Per avere invece una contestualizzazione e una descrizione più completa dei dati così ricavati, abbiamo sfruttato informazioni provenienti da fonti diverse, dedicandoci al problema della *Variety*. In questo modo siamo riusciti a mettere in relazione i tweet con gli episodi più rilevanti della partita, andando a scoprire corrispondenze interessanti, come ad esempio il fatto che durante le pause vi sia un evidente aumento dell'utilizzo di Twitter; abbiamo inoltre individuato una correlazione tra le performance e le azioni dei singoli giocatori durante il game e il numero di tweet contenenti il loro nome. In particolare è stato interessante osservare l'evidente picco di post relativi a Klay Thompson in corrispondenza del suo infortunio. Infine abbiamo esplorato le parole contenute nei tweet e le loro occorrenze, da cui abbiamo visto che Klay, Thompson e Lowry sono alcune delle parole più scritte, e che "Raptors" compare più di "Warriors".

References

[1] *2018-19 NBA Rulebook*, 2019 https://official.nba.com/rulebook/.

[2] *Basketball reference*, 2019, https://www.basketball-reference.com/.

[3] *Twitter Developer*, 2019, https://developer.twitter.com/.

[4] Rezzani A., *Big Data Analytics - Il manuale del data scientist*, Maggioli editore, 2017.