Formula Tweet

Dario Carolla[1]*, Matteo Licciardello[2]*

Abstract
La quantità di dati pubblicati dagli utenti e resi accessibili attraverso i *social network* rende l'analisi dell'attività di questi ultimi sempre più rilevante. Anche le società televisive e sportive vengono coinvolte in questo processo di raccolta e analisi dei dati volto all'aumento degli ascolti e dei profitti. Lo scopo di questo progetto è proporre un'architettura che consenta di acquisire, memorizzare ed integrare dati in tempo reale al fine di analizzare il traffico generato da un evento sportivo influente, come una gara di Formula 1, attraverso il social network *Twitter*.

Keywords
Data Management — F1 Twitter Stream Analysis — Project Report

[1,2] *Dipartimento di informatica, sistemistica e comunicazione, CdLM Data Science, Università degli studi di Milano - Bicocca*
1 - d.carolla@campus.unimib.it
2 - m.licciardello@campus.unimib.it

Contents

Introduzione

Ad oggi, Twitter [1] è il social network più diffuso che permette la raccolta di dati in tempo reale attraverso *API* proprietarie, inoltre alla base del social network vi è l'utilizzo di *hashtag* che permettono di identificare e seguire con facilità le discussioni relative ad un determinato evento. Per questi motivi Twitter è risultato il miglior candidato per prelevare informazioni dai social network. L'evento che si è voluto studiare è il Gran Premio di Germania di Formula 1 tenutosi il 28 luglio 2019. Lo scopo del progetto è analizzare il traffico di tweet per valutare quali siano gli eventi dal maggior impatto mediatico.

1. Architettura sviluppata

Per poter scaricare i dati da Twitter, oltre le sue API, è stato utilizzato Apache Kafka [2]. Si tratta di una piattaforma per il data streaming che ha lo scopo principale di ottimizzare la trasmissione e l'elaborazione di flussi di dati che vengono scambiati tramite il collegamento diretto tra il destinatario e la fonte di dati. L'utilizzo di Kafka ha dunque permesso di scaricare i dati da Twitter in tempo reale.

L'architettura di Kafka è divisibile in due componenti principali:

- Producer: applicazione che invia il messaggio;
- Consumer: applicazione che riceve il messaggio.

Prima di giungere al consumer il messaggio viene recuperato dai topic in Kafka, i quali sono una sorta di categoria utilizzata per raggruppare i messaggi. Una volta che il topic ha recuperato il messaggio un consumer può collegarsi al topic e scaricare il messaggio.

L'architettura Kafka implementata prevede un unico producer, il quale invia i dati ad un topic che, infine, vengono recuperati da un consumer.

1.1 Producer

Il producer sviluppato è stato utilizzato per scaricare i tweet riguardanti il Gran Premio di Germania tramite le API di Twitter. L'intervallo di tempo per cui il producer scarica i tweet va da due ore prima dell'inizio della gara fino a due ore dopo il termine. I tweet vengono scaricati secondo i seguenti criteri: sono stati forniti in input alle API di Twitter tutti gli hashtag principali della Formula 1 come ad esempio *#F1* e *#Formula1*, ma anche hashtag creati appositamente per il gran premio considerato come *#GermanGP*. A questo scopo, tramite uno studio della piattaforma Twitter, sono stati raccolti tutti gli hashtag riferiti ai piloti ed alle scuderie ottenendo così un totale di tre hashtag generici, quarantotto riferiti ai piloti e diciotto alle scuderie.

Volendo anche osservare il comportamento delle scuderie e dei piloti sui social network sono stati inseriti anche gli *user id* dei canali di tutte le scuderie e dei piloti. Inoltre sono stati scaricati tutti i tweet effettuati dal canale ufficiale di F1. Durante il recupero dei tweet, il producer effettua una prima pulizia dei dati eliminando dal tweet alcuni campi reputati superflui, a ciò segue la memorizzazione dei dati sul topic.

1.2 Consumer

Il consumer, come già accennato, si occupa di prelevare i dati presenti sul topic. Durante questa operazione il consumer si serve di un database MongoDB [3] in cui salvare i dati, nello specifico, ogni tweet corrisponde ad un documento del DB, il quale è costituito da sei diverse *collection* contenenti:

- Tutti i tweet prelevati;
- I tweet contenenti gli hashtag riferiti ai piloti;
- I tweet effettuati dai piloti;
- I tweet effettuati dalle scuderie;
- I tweet riferiti alle scuderie;
- I tweet effettuati dal canale ufficiale.

L'assegnazione dei tweet alla collection corretta avviene in maniera automatica: i dati provenienti da un account specifico vengono rilevati ed assegnati alla collection a cui quell'account fa riferimento, oppure, mediante l'utilizzo di un algoritmo di *pattern matching*. In questo caso il testo di ogni tweet viene scansionato parola per parola, nel caso in cui venga individuato un *hashtag* facente parte delle parole chiave di una collection allora il tweet viene salvato all'interno di quest'ultima. Nel caso in cui un tweet risulti contenere elementi chiave di due collection diverse allora verrà inserito in entrambe.

Al temine di questa operazione di assegnamento, i documenti contenuti nel DB risultano avere il formato rappresentato nella figura sottostante:

```
{
    id: "1155432698954915840"
    created_at : 2019-07-28T10:59:56.000
    text: "#F1 What a moment"
    user: Object
        id: "813735697525907457"
        name: "Andre Oliveira"
        screen_name: "MrAndreOliveira"
        location: "Portugal"
        verified: false
        followers_count: 158
        statuses_count:4990
    coordinates:null
    retweet: false

}
```

In questo caso il campo *retweet* è impostato a false, questo perché il tweet considerato non è un retweet; in caso contrario si avrebbe l'intera struttura del tweet al quale fa riferimento.

In totale sono stati raccolti 309.915 tweet i quali, all'interno del DB, occupano 203.6 MB.

2. Integrazione

Per soddisfare il requisito di varietà dei dati, i tweet raccolti in formato *JSON* sono stati integrati con informazioni esterne in formato *CSV*.

2.1 Dati utilizzati

I dati utilizzati per l'integrazione sono forniti dal sito *Ergast Developer API* [4]. All'interno del sito sono disponibili tutti i dati riguardanti la Formula 1 dal 1950 ad oggi. I dataset presenti sul sito non sono aggiornati in tempo reale, ma i nuovi dati vengono inseriti circa sei ore dopo il termine della gara. I dataset sono forniti in formato CSV e forniscono le seguenti informazioni per ogni gara disputata:

- Dati anagrafici dei piloti;
- Dati identificativi delle scuderie;

- Risultati delle scuderie e dei piloti;
- Classifiche aggiornate;
- Tempi sul giro per ogni pilota;
- Dati relativi alla gara.

2.2 Processo di integrazione

La modellazione ideata prevede la creazione di due dataset separati: il primo riguardante i piloti, mentre il secondo le scuderie. I dataset sono stati generati utilizzando i file CSV precedentemente menzionati, sui quali sono state effettuate diverse operazioni di *preprocessing* utilizzando la libreria di Python [5] denominata *Pandas* [6], attraverso la quale è possibile operare con più flessibilità sulla struttura dei dati.

Si è scelto di estrarre unicamente i dati riguardanti i piloti e le scuderie partecipanti al campionato 2019. A questi ultimi sono state aggiunte diverse informazioni riguardanti la gara presa in esame e la classifica di campionato. Per quanto riguarda le operazioni di preprocessing effettuate, il formato dei dati è stato adattato per poter essere correttamente interpretato dal DBMS MongoDB. Da quest'ultimo sono stati recuperati i tweet che vengono interpretati da Python come una lista di dizionari. Dunque anche i dataframe generati sono stati trasformati in una lista di dizionari per permetterne l'integrazione. La struttura utilizzata per i dizionari è stata ideata considerando il futuro inserimento dei dati in MongoDB. Quest'ultimo, infatti, permette anche la gestione di documenti *embedded*, ovvero documenti innestati tra loro. Tale caratteristica permette di ottenere ottime prestazioni per le operazioni di lettura e consente di aggiornare i dati innestati con una singola operazione di scrittura.

Il primo dizionario generato ha lo scopo di integrare i dati riguardanti i piloti con i tweet raccolti distinguendo tra i tweet riguardanti i piloti ed i tweet effettuati da questi ultimi. Dunque, sono stati generati venti dizionari, uno per ogni pilota. All'interno di ogni dizionario sono stati aggiunti i dati anagrafici ed i dati riguardanti la gara, come i punti conquistati e la posizione finale.

Il secondo dizionario, invece, riguarda le scuderie: in questo caso i dizionari generati sono dieci, ovvero il numero di scuderie presenti nel campionato. All'interno di ognuno di essi sono presenti tutte le informazioni che riguardano la scuderia considerata come, ad esempio, i punti totalizzati nella gara, la nazionalità ed il numero di vittorie totali. Le considerazioni effettuate sull'integrazione dei tweet sono le medesime delle scuderie. Anche in questo caso, quindi, i tweet sono stati integrati con i dati delle scuderie distinguendo i tweet effettuati dalle scuderie stesse ed i tweet riguardanti queste ultime.

2.3 Integrazione degli eventi

Particolare rilevanza è stata assegnata all'individuazione degli eventi, ovvero incidenti o avvenimenti che causano il ritiro dei piloti dalla gara.

Inizialmente, si è deciso di raggiungere l'obiettivo attraverso l'analisi del testo contenuto nei tweet pubblicati durante la gara dal canale ufficiale "F1", tuttavia, non è stato possibile ottenere un livello di accuratezza accettabile in quanto non vengono utilizzati pattern specifici, inoltre il tweet viene pubblicato con un ritardo variabile rispetto all'effettivo svolgersi dell'evento.

Al fine di aumentare l'accuratezza, si è deciso di utilizzare il dataset contenente i tempi sul giro dei piloti unendolo ai risultati finali della gara: tale dataset contiene non solo le posizioni dei piloti giunti al traguardo, ma anche i piloti ritirati, il motivo del ritiro ed il giro in cui l'evento è accaduto. Rispetto al metodo considerato precedentemente, il procedimento appena descritto consente di collocare con maggiore precisione l'avvenimento all'interno dell'arco temporale della gara.

Ai fini dell'integrazione è stata creata una lista costituita da sessantaquattro dizionari, uno per ogni giro della gara. Ad ogni dizionario è stata aggiunta la classifica aggiornata al termine del giro fornendo il tempo e la posizione per ogni pilota. Inoltre, tramite *pattern matching* sono stati rilevati i tweet effettuati dal canale ufficiale relativi ad un determinato giro e sono stati inseriti all'interno del dizionario inerente a quest'ultimo. Infine, durante la creazione del dizionario viene controllato lo stato del pilota e, nel caso in cui quest'ultimo si sia ritirato dalla gara, viene creata una chiave *"Evento rilevato"* in cui viene segnalato il tipo di evento rilevato ed il pilota che ne è soggetto.

2.4 Linea temporale della gara

Per poter analizzare il comportamento degli utenti di Twitter durante la gara è stato necessario determinare l'orario di ogni giro in modo da poter avere una linea temporale della gara. Le gare di F1 durante tutto il campionato iniziano alle ore 15:10, ma per motivi di sicurezza l'orario potrebbe variare. Questo è stato il caso della gara presa in esame dove la situazione meteorologica non ha consentito l'inizio della gara all'orario canon-

ico. Per poter recuperare l'orario di partenza è stato creato un algoritmo che, analizzando la lista di dizionari contenente i giri precedentemente creata, ricerca il primo tweet effettuato dal canale ufficiale e ne rileva l'orario. Partendo da quest'ultimo viene sottratto il tempo impiegato dal pilota in prima posizione per effettuare i giri precedenti. A quest'orario calcolato viene inoltre aggiunto un tempo arbitrario di 1.30 minuti per simulare il tempo che intercorre tra l'inizio del giro e la pubblicazione del tweet da parte del canale. Utilizzando questo algoritmo l'orario di inizio della gara è risultato essere le 15:20:01 riuscendo così a calcolare l'orario corretto di partenza.

Tramite l'utilizzo dell'orario calcolato e del tempo dei piloti è stato possibile assegnare ad ogni giro un orario di inizio e di fine.

3. Storage dei dati

Al termine di tutte le operazioni di pulizia ed integrazione dei dati, questi ultimi sono stati caricati all'interno di un DB MongoDB, il quale è costituito da quattro collection contenenti:

- Tutti i tweet raccolti: 309.915 documenti ed un peso di 191.15 MB;
- I dati dei piloti: 20 documenti ed un peso di 5.6 MB;
- I dati delle scuderie: 10 documenti ed un peso di 13.4 MB;
- I dati riguardanti la gara: 64 documenti ed un peso di 145.6 KB.

Il DB ha un peso totale di 210.4 MB.

Per quanto riguarda la collection contenente tutti i tweet, la struttura è invariata rispetto a quella già rappresentata precedentemente.

Come già accennato, la collection contenente i dati relativi alle scuderie è composta da 10 documenti, uno per ogni scuderia. La struttura è rappresentata nella seguente figura:

```
{
    constructorId: 1
    constructorName: "McLaren"
    constructorNationality:"British"
    chammpionshipPoints: 70
    chammpionshipPositions: 4
    seasonWins: 0
    Tweet scuderia: Array
    Tweet utenti: Array
}
```

I campi "Tweet utenti" e "Tweet scuderia" sono Array di documenti, all'interno dei quali sono contenuti i tweet prelevati dalla piattaforma. Il secondo in particolare è presente all'interno del documento solo nel caso in cui la scuderia abbia twittato durante il periodo analizzato.

Come per le scuderie, la collection relativa ai dati dei piloti è composta da un numero di documenti pari alla quantità di piloti. In seguito alle operazioni di integrazione possiede la struttura rappresentata nella figura successiva:

```
{
    driverId: 154
    name : "Romain Grosjean"
    number : "8"
    cod: "GRO"
    dob: 1986-04-17T00:00:00.000
    nationality: "French"
    raceId: 1020
    seasonPoints: 8
    standingPositions: 17
    seasonWins: 0
    constructorId: 210
    startPosition: 6
    finalPosition: 7
    laps: 64
    time: "+16.838"
    status: "Finished"
    Tweet utenti: Array
    Tweet pilota: Array
}
```

Anche in questo caso il campo "Tweet pilota" è presente all'interno del documento solo nel caso in cui siano presenti tweet effettuati dal pilota.

Per quanto riguarda la collection di dati relativi alla gara, ogni documento rappresenta un giro di gara. La struttura con cui si presenta è rappresentata in figura:

```
{
    Lap: 2
    Tweet official channel: null
    Classifica: Array
        0: Pilota_1
            Driver_Id: 1
```

```
            Driver: "Lewis Hamilton"
            Posizione: 1
            Millisecondi: 94720
            Time: "1.34.728"
        1: Pilota_2
            ...
    Evento rilevato: Object
        Evento: "Spun off"
        Soggetto: "Sergio Perez"
    Inizio giro: 2019-07-28T13:20:01.000
    Fine giro: 2019-07-28T13:22:00.108
}
```


Il campo denominato ”Tweet official channel” assume valore diverso dal valore nullo solamente nel caso in cui venga rilevato un tweet proveniente dal canale ufficiale riferito a quel giro. Il campo classifica contiene al suo interno 20 documenti, uno per ogni pilota. Il campo ”Evento rilevato”, invece, è presente solo nel caso in cui si sia verificato un evento durante il giro considerato.

4. Risultati e conclusioni

Per verificare l'attività di raccolta dati è stato realizzato un grafico (1) tramite il software di visualizzazione *Tableau* [7], rappresentante i tweet raccolti per ogni minuto di attività del producer.

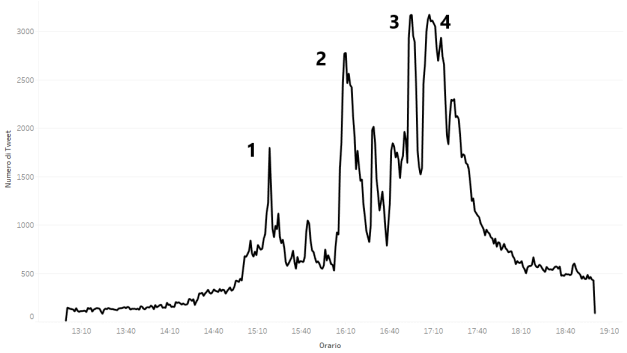

Figure 1. Tweet raccolti per minuto

Come era possibile immaginare, i picchi maggiori di tweet si sono verificati durante lo svolgimento della gara, mentre è possibile osservare che l'attività sul social network è molto maggiore al termine della gara rispetto al periodo precedente alla partenza. In totale, sono stati rilevati quattro eventi che, più di tutti gli altri, hanno spinto gli spettatori a twittare. In particolare, i picchi rilevati sono i seguenti:

1. Ore 15:18, tweet appena precedenti la partenza della gara;
2. Ore 16:09 - giro 28, incidente di Leclerc mentre era in seconda posizione;
3. Ore 16:54 - giro 57, Bottas colpisce il muro e si ritira;
4. Ore 17:07 - giro 64, termina la gara con la vittoria di Verstappen.

In aggiunta agli eventi che hanno determinato i picchi di tweet, si sono rilevati altri eventi che però hanno riscosso meno successo tra gli spettatori, si tratta di:

- Ore 15:23 - giro 2, Perez esce di pista e si ritira;
- Ore 15:45 - giro 14, Ricciardo rompe il motore;
- Ore 16:03 - giro 26, una perdita di potenza costringe Norris al ritiro;
- Ore 16:23 - giro 40, Hulkenberg si ritira in seguito ad un incidente;
- Ore 17:02 - giro 62, Gasly si ritira a causa di un contatto.

Per verificare l'integrazione dei dati svolta, invece, è possibile interrogare il DB creato effettuando delle *query*. Tramite queste ultime, è stato possibile notare che il tweet del canale ufficiale più retwittato riguarda l'incidente di Valtteri Bottas al giro 57 con un totale di 688 retweet. Il pilota più menzionato su Twitter, invece, è stato Sebastian Vettel, che compare in 10416 tweet: il pilota della Ferrari, infatti, pur non arrivando primo è riuscito ad arrivare secondo partendo dalla ventesima posizione.

In conclusione, risulta evidente come l'interazione con i social network durante gli eventi sportivi sia molto consistente. Questo, però, dipende fortemente dagli eventi che si verificano durante la manifestazione sportiva: in questa particolare occasione, infatti, l'imprevedibilità data dalla condizione meteorologica durante la gara ha causato molti eventi rilevanti. In particolare, si è osservato come l'alto numero di tweet sia fortemente influenzato dagli incidenti più che dai sorpassi avvenuti durante la gara.

References

[1] Twitter. https://twitter.com/.

[2] Kafka stream. https://kafka.apache.org.

[3] Mongodb. https://www.mongodb.com.

[4] Ergast. http://ergast.com/mrd/.

[5] Python. https://www.python.org.

[6] Pandas. https://pandas.pydata.org/.

[7] Tableau. https://www.tableau.com.