Università degli Studi di Milano - Bicocca

Corso di Laura Magistrale di Data Science, a.a. 2020/2021

Le reazioni emotive degli utenti di Twitter alle notizie covid-19 in Italia

Guida operativa, Giugno 2021

Daniele Quattrocchi – matricola 825418

Vittorio Menardo – matricola 812341

Raffaele Moretti – matricola 794537

DISCLAIMER

L'intero codice di programmazione Python, versione 3.8, è stato scritto e sviluppato creando inizialmente un ambiente virtuale in Jupyter Notebook con il programma *Anaconda*.

INTRODUZIONE

Le domande che sono state poste per lo sviluppo di questo progetto sono le seguenti: *come ha reagito la piattaforma italiana di Twitter alle notizie in merito al COVID-19? In particolare, quali sono state le emozioni principalmente espresse sulla piattaforma dagli utenti?* Per rispondere a queste domande, abbiamo deciso innanzitutto di definire un periodo temporale di analisi, compreso tra il 24 febbraio 2020, data che rappresenta la pubblicazione ufficiale del primo bollettino Covid, e il 2 maggio 2021. In seguito, sono stati cercati i dati che consentivano di rispondere al point del progetto. Per questo motivo, il gruppo ha lavorato distintamente su due temi:

- Come collegarsi e scaricare i dati storici dal server di Twitter.
- Scaricare i dati relativi al bollettino COVID-19.

Di seguito verrà prima fornita la descrizione del primo topic.

COVID, TWITTER, EMOZIONI.

Per risolvere la prima questione sopra indicata, abbiamo deciso di utilizzare la libreria Python *snscrape*, che per essere installata richiede un'attenta configurazione [1].

Successivamente, è stata posta un ulteriore domanda*: quali emozioni si possono rilevare sulla piattaforma? E come si possono rilevare?* È noto che per esprimere la stessa emozione le persone possono utilizzare parole diverse. Quindi tali parole possono essere intese come sinonimi. Per questa ragione, abbiamo deciso di procedere con una categorizzazione delle emozioni, ossia parole che esprimono la medesima sensazione.

Per poter eseguire questa classificazione, ci siamo affidati alla teoria sulla categorizzazione delle emozioni primarie dello psicologo Daniel Goleman [2].

In questo modo, è stato realizzato lo script contenuto nella cartella 'selezione emozioni'. In tale script, abbiamo riportato le categorie delle emozioni definite da Goleman, e per ciascuna parola nella categoria abbiamo calcolato il numero di tweet e di risposte[3] in lingua italiana che contengono la parola in questione e la parola covid, nel periodo di analisi sopra descritto. Abbiamo fatto questo per avere una diretta correlazione tra il fenomeno emozionale e il COVID-19. Successivamente, una volta effettuati i conteggi per ogni parola di ciascuna categoria, abbiamo deciso di procedere prendendo per ciascuna categoria la parola al quale corrisponde il numero maggiore di tweets. Tali parole sono:

- Amore, per la categoria amore
- Dolore, per la categoria tristezza
- Gioia, per la categoria gioia
- Odio, per la categoria collera

[1] Per ulteriori informazioni visita il tutorial per la configurazione e installazione della libreria snscrape al link: https://betterprogramming.pub/how-to-scrape-tweets-with-snscrape-90124ed006af

[2] Per sapere di più su tale teoria visita il link: https://www.giuliamartino.it/le-otto-emozioni-primarie/

[3] Si è deciso di considerare le riposte oltre ai tweet per avere maggiori testi da sottoporre ad analisi. Riteniamo che in questo modo, le analisi risultato maggiormente complete, nonostante possano essere soggette ad una percentuale di errore maggiore.

- Paura per la categoria paura
- Schifo, per la categoria disgusto
- Shock, per la categoria sorpresa
- Vergogna, per la categoria Vergogna

A questo punto, è stato realizzato la seconda parte del codice relativo a questo primo topic, contenuto nella cartella 'estrazione tweets emozioni'. La cartella contiene otto script, ciascuno corrispondente all'analisi di una parola chiave considerata. L'analisi di ciascuna parola chiave è identica, per tale ragione si procederà a descrivere l'analisi della parola Paura, come esempio di riferimento per le altre emozioni.

Il primo step consiste nell'estrarre i tweet in lingua italiana contenente le parole paura e covid, nel periodo di tempo considerato. Successivamente, i dati, che sono estratti nel formato JSON, sono stati convertiti nel formato CSV. Successivamente, il dataset è stato pulito in quanto la ricerca dei tweet contenenti le parole covid e paura presenta un problema da non sottovalutare. La ricerca permette si l'estrazione dei tweet contenenti queste parole, tuttavia queste parole possono essere:

- Nel testo del tweet
- Nello username
- Nel display name

Per questa ragione, è probabile avere diverse combinazioni di quanto espresso sopra, portando all'estrazione di tweet di utenti che sulla piattaforma hanno il nome PAURA e nel testo parlano tutt'altro rispetto alla paura da covid. Nonostante ciò, tweet di questo tipo sono stati piuttosto rari nell'analisi di tutte le parole chiave. Per risolvere questo problema, abbiamo deciso di pulire il dataset tenendo solamente i tweet che contenevano le parole covid e paura solamente nel testo del tweet.

Successivamente, si è proceduto ad analizzare un campione dei tweet corrispondente all'1% dei tweet totali estratti. Dunque, per la parola paura sono stati analizzati circa 308 tweet. Il fine di questa procedura è individuare se nei tweet è presente ironia o negazione dell'emozione del tipo:

- *"Non ho paura"*, per la parola paura
- *"Mai una gioia"*, per la parola gioia
- *"L'amore prima del covid + emoji che ride"*, per la parola amore

e così via.

Per accettare la validità di tutto l'insieme di tweets, abbiamo deciso di studiare se il numero di tweets analizzati nel campione risultati come invalidi superavano il numero dei tweets classificati come validi.

Le parole Amore e Gioia sono risultate estremamente soggette all'ironia degli utenti, rendendo lo studio di queste emozioni non valido. Pertanto, non si è proceduto ulteriormente nell' analisi di queste due emozioni, tenendo solamente sei emozioni in totale sottoposte ad analisi.

Successivamente, abbiamo proceduto a raggruppare i tweet per giorni, generando un conteggio per ogni giorno, relativo all'emozione analizzata (conteggio giornaliero non cumulato). Tuttavia, si è presentato un problema ossia una mancanza temporale tra i giorni raggruppati. In altri termini, per

le parole chiave analizzate, vi sono stati giorni in cui non ci sono stati tweet in merito alle parole chiave considerate. Dunque, abbiamo risolto questo problema introducendo i giorni mancanti nel dataframe e assegnano a questi giorni il conteggio 0.

Successivamente, i dati sono stati convertiti nella struttura JSON tramite un'opportuna funzione, che di fatto mantiene il tipo di dato come testo. Questo è stato necessario per poter caricare su mongoDB locale i dati testo che sono stati poi opportunamente riconosciuti/convertiti in formato JSON. Nella cartella allegata, i dati in questione sono nella sottocartella emozioni_db, che prende il nome del database creato in mongodb locale.

STATISTICHE COVID

Per il secondo topic, sono stati scaricati i dati del bollettino ufficiale[4]. Di tale dataset sono state tenute solamente le colonne di principale interesse per lo studio e a scopo esplorativo, ossia:

- Nuovi positivi (giornalieri, non cumulati)
- Deceduti (cumulati)

Successivamente, sono stati estratti dal dataframe i record compresi nel periodo di analisi sopra descritto, per avere concomitanza con il periodo di analisi con i tweets relativi alle emozioni durante la pandemia.

A questo punto, i dati sui deceduti sono stati calcolati giornalmente, in quanto tale decisione è necessaria perché i conteggi dei tweets sono giornalieri. Questa decisione è dovuta al fatto che per poter analizzare lo stato d'animo della piattaforma è necessario considerare dati giornalieri non cumulati.

Successivamente, si è proceduto ad eseguire i seguenti step:

1. ***Importazione dataset casi nel mondo***

Lettura dell'ultimo dataset di Our World In Data, tramite github[5], in modo da avere il dataset sempre aggiornato. Da questo dataset andremo a prendere la colonna casi_mondo. Tale colonna presenta conteggi giornalieri in funzione del giorno di riferimento (i.e. non cumulato)

2. ***Importazione dataset vaccini***

Lettura dell'ultimo dataset di Our World In Data, tramite github[6], in modo da avere il dataset sempre aggiornato. Da questo dataset andremo a prendere la colonna total_vaccinations. Dopodiché procediamo per calcolare i dati giornalieri, come per quanto riguarda i dati sui deceduti sopra descritti.

[4] I dati sono disponibili su Github al seguente link: https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv
[5] https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv
[6] https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations/country_data/Italy.csv

3. Dataset finale

In questa fase andremo a creare il dataset finale. Per fare questo facciamo prima di tutto una merge tra il dataset della protezione civile e il dataset vaccini; merge tra il dataset appena creato e il dataset con i casi nel mondo. Entrambe le merge vengono fatte attraverso la data.

4. Conversione dati dal formato CSV al formato JSON

Conversione del dataset nella struttura JSON, mantenendo il tipo di dato testuale: ogni colonna viene portata in JSON come intero, ad eccezione della colonna data

5. Importazione dei dati in MongoDB locale

Importazione di MongoClient e collegamento a MondoDB locale.

Nella cartella allegata, i dati finali ottenuti in questione sono nella sottocartella statistiche_covid_db, che prende il nome del database creato in mongodb locale.

TWEETS COVID

Per concludere l'analisi, a scopo esplorativo si è deciso di estrarre dalla piattaforma Twitter i tweets in lingua italiana contenente la parola covid, senza specificare altre parole di ricerca, tra il 24 febbraio 2020 e il 2 maggio 2021. Il codice è riportato nella cartella 'estrazione tweets covid'. I dati estratti sono stati innanzitutto puliti, tenendo solamente i tweets che contenevano la parola covid almeno nel testo del tweet. Successivamente tali dati sono stati raggruppati per giorno generando un conteggio giornaliero e sono stati poi convertiti nella struttura JSON (mantenendo comunque il tipo di dato testuale). In questo modo, i dati sono poi stati caricati in mongoDB locale. Nella cartella allegata, i dati in questione sono nella sottocartella covid_tweets_db, che prende il nome del database creato in mongodb locale.

INTEGRAZIONE PARTE 1

Dopo aver esportato i dati in mongoDB locale, si è proceduto ad importare i conteggi relativi alle singole emozioni in uno script notebook, contenuto nella cartella 'integrazione e dati finali' e nominato ' integrazione_emozioni_script_definitivo'. Importati i dati, questi sono stati convertiti dal formato JSON al formato CSV, per poi essere mergiati sulla colonna 'data'. In questo modo si è ottenuta una tabella contenente per ciascun giorno il numero di conteggi di tweets contenenti l'emozione indicata nella colonna. Successivamente tali dati sono stati poi convertiti nella struttura

JSON ed esportati in mongoDB locale. Nella cartella allegata, i dati in questione sono nella sottocartella dati_integrati_db, che prende il nome del database creato in mongodb locale.

Successivamente si è creato uno script analogo a questo indicato, dove i conteggi per ciascuna colonna sono stati resi cumulati (i.e. il conteggio del giorno 2 contiene anche il conteggio del giorno 1). Inoltre, si è proceduto ad un unpivot della tabella cumulata in modo da avere le emozioni come valori al quale sulla stessa riga vengono associati i conteggi giornalieri di tweet. Infine, la procedura di conversione del tipo di dato ed esportazione in mongoDB locale è analoga a quella descritta sopra. Questi dati creati sono fondamentali per la produzione del grafico race chart, che verrà spiegato opportunamente nel report di data vizualization. Nella cartella allegata, i dati in questione sono nella sottocartella data_per_viz, che prende il nome del database creato in mongodb locale.

INTEGRAZIONE PARTE 2

Analogamente alla procedura indicata sopra, i dati contenenti i conteggi delle emozioni giornalieri non cumulati sono stati importati in uno script notebook contenuto nella stessa cartella ma nominato 'integrazione_emozioni_script_parte_2'. Inoltre, in questo script sono stati importati da mongodb locale anche i dati relativi alle statistiche covid e i conteggi giornalieri dei tweets contenenti la parola covid (e non altre parole di ricerca). I dati sono stati convertiti in formato CSV, mergiati sulla colonna data. Tali dati sono stati poi calcolati in media mobile a 7 giorni e successivamente riconvertiti nella struttura JSON per poi essere esportati in mongodb locale. Nella cartella allegata, i dati in questione sono nella sottocartella dati_integrati_db, che prende il nome del database creato in mongodb locale.